Divergence-based Framework for Common Spatial Patterns Algorithms

Wojciech Samek, *Member, IEEE,* Motoaki Kawanabe and Klaus-Robert Müller, *Member, IEEE,*

***Abstract*—Controlling a device with a Brain-Computer Interface (BCI) requires extraction of relevant and robust features from high-dimensional electroencephalographic recordings. Spatial filtering is a crucial step in this feature extraction process. This work reviews algorithms for spatial filter computation and introduces a general framework for this task based on divergence maximization. We show that the popular Common Spatial Patterns (CSP) algorithm can be formulated as a divergence maximization problem and computed within our framework. Our approach easily permits enforcing different invariances and utilizing information from other subjects, thus it unifies many of the recently proposed CSP variants in a principled manner. Furthermore it allows to design novel spatial filtering algorithms by incorporating regularization schemes into the optimization process or applying other divergences. We evaluate the proposed approach using three regularization schemes, investigate the advantages of beta divergence and show that subject-independent feature spaces can be extracted by jointly optimizing the divergence problems of multiple users. We discuss the relations to several CSP variants and investigate the advantages and limitations of our approach with simulations. Finally we provide experimental results on a data set containing recordings from 80 subjects and interpret the obtained patterns from a neurophysiological perspective.**

I. Introduction

Brain-Computer Interface (BCI) systems [1] [2] provide a novel communication channel for healthy and disabled people to interact with the environment. The core idea of a BCI is to decode the mental state of a subject from its brain activity and to use this information for controlling a computer application or a robotic device such as a wheelchair. There are several ways to voluntarily induce different mental states, one common approach is motor imagery. In this paradigm, participants are asked to imagine the movements of their hands, feet or mouths. This alters the rhythmic activity over different locations in the sensorimotor cortex and can be measured in the Electroencephalography (EEG). However, reliable decoding of mental state is a very challenging task as the recorded EEG signal contains contributions from both task-related and task-unrelated processes. In order to enhance the task-related neural activity, i.e. increase its signal-to-noise ratio, it is common to perform spatial filtering. A very popular method for this is Common Spatial Patterns (CSP) (e.g. [3] [4] [5] [6] [7]). Spatial filters computed with CSP are well suited to discriminate between different mental states induced by motor imagery as they focus on the synchronization and desynchronization effects occurring over different locations of the sensorimotor cortex after performing motor imagery. Although impressive improvements in BCI efficiency have been achieved with CSP (see e.g. BCI Competitions[1] [8] [9] [10] [11]) the current BCI systems are far from being perfect in terms of reliability and generalizability. This suboptimal performance can be mainly attributed to a low signal-to-noise ratio [12] [4] [13], the presence of artifacts in the data [14] [15] [16] and the non-stationary nature of the EEG signal [17] [18] [19].

Several extensions of vanilla CSP have been proposed to increase the robustness and discriminativity of the extracted features by applying regularization, incorporating data from other sessions/subjects or using robust estimators (see Section II for an overview). All these different algorithms were not designed as part of a general robust approach to spatial filtering, but rather each method was proposed for a specific application scenario with its own optimization strategy. This diversity of algorithms not only poses a practical implementation problem, but the lack of flexibility of these methods may result in suboptimal solutions (see Section V). In addition, they do not optimize the same objective, consequently they can neither be easily combined nor compared.

The main goal of this paper is to propose an unifying optimization framework for spatial filter computation based on divergence maximization. For that we first provide a novel view on CSP, namely we prove that the CSP spatial filters

We thank Adam Wilkins for the support in writing. 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 work was supported by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein. *Asterisks indicate corresponding author.*

W. Samek is with Berlin Institute of Technology (TU Berlin), 10587 Berlin, Germany. (e-mail: wojciech.samek@tu-berlin.de)

M. Kawanabe is with Advanced Telecommunication Research Institute International, 619-0288 Kyoto, Japan. (e-mail: kawanabe@atr.jp)

K.-R. Müller is with the Berlin Institute of Technology (TU Berlin), 10587 Berlin, Germany, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Republic of Korea (e-mail: klaus-robert.mueller@tu-berlin.de)

W. Samek, M. Kawanabe and K-R. Müller, Divergence-based Framework for Common Spatial Patterns Algorithms, *IEEE Reviews in Biomedical Engineering*, 2013. http://dx.doi.org/10.1109/RBME.2013.2290621

[1] http://www.bbci.de/competition/

span a subspace with maximum symmetric Kullback-Leibler divergence between the average class distributions. This relation permits reformulating CSP as divergence maximization problem. Furthermore we propose to add different regularization terms (measured as divergences) to the optimization problem in order to increase the robustness and generalizability of the extracted features. With this generic regularization approach our framework unifies many of the state-of-the-art CSP variants. The fact that all quantities used in the optimization process are measured as divergences enables us to easily combine and compare different regularization schemes. Since divergences have a clear mathematical foundation and can be interpreted from an information geometric perspective [20] we can easily obtain meaningful CSP-like spatial filtering algorithms with novel properties by using other divergences. In this paper we will investigate the usage of beta divergence (generalization of Kullback-Leibler divergence) within our framework.

This paper is organized as follows. In the next section we review the Common Spatial Patterns algorithm and its state-of-the-art variants. In Section III we introduce the divergence-based framework for spatial filter computation, we prove its equivalence to CSP and discuss two optimization algorithms. Section IV extends the divergence-based framework by introducing different regularization terms and deriving a beta divergence-based version of the algorithms. It also discusses the relations to some of the published CSP variants. In Section V we investigates the advantages and limitations of our approach using simulations. The experimental evaluation on a data set containing EEG recordings from 80 subjects is presented in Section VI. This work concludes with a discussion in Section VII. An implementation of our framework is available at `http://www.divergence-methods.org`.

II. Spatial Filtering Algorithms

A. Why Spatial Filtering ?

In Electroencephalography (EEG) we record the electrical activity on the scalp (e.g. [21] [22] [23]). The recorded signal at an electrode does not only reflect neural voltage fluctuations underneath that electrode, but it also captures the activity of distance current sources through volume conduction effects. Thus the EEG signal $\mathbf{x}(t) \in \mathbb{R}^D$ generated by the brain sources $\mathbf{s}(t) \in \mathbb{R}^D$ is usually represented as a (noisy) linear mixture [24]

$$\mathbf{x}(t) \quad = \quad \mathbf{A}\mathbf{s}(t) \; + \; \mathbf{n}(t), \tag{1}$$

where the matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ maps the activity of each source to the electrode space. Note that many blind source separation algorithms (e.g. ICA) assume that the number of brain sources and electrodes coincides. On the other hand inverse solution algorithms (such as LAURA, ELECTRA, LORETA, etc.) often rely on more realistic assumptions, namely that the number of sources is much larger than the number of sensors [25]. These algorithms compute spatial filters trying to address the volume conduction effect. Contributions not captured by $\mathbf{A}$ are considered as normally distributed noise $\mathbf{n}(t)$.

The imagination of movement execution attenuates the sensorimotor rhythms (SMRs) [22] in the corresponding cortical areas. For instance, left hand motor imagery mainly affects the SMRs over the right motor cortex. In order to distinguish motor imagery tasks of different body parts it is necessary to recognize the sources of SMR modulation. This is usually achieved by *spatial filtering*

$$\hat{\mathbf{s}}(t) \quad = \quad \mathbf{W}^\top \mathbf{x}(t), \tag{2}$$

where $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_d] \in \mathbb{R}^{D\times d}$ projects the EEG signal to a d-dimensional subspace. A spatial filter $\mathbf{w}_i$ weights each electrode to extract information about the true source of interest $\mathbf{s}(t)$. Since changes in the SMR are visible in the band-power of the signal (= variance of band-pass filtered signals), one can enhance the SMR modulation by projecting the data to a subspace with maximum band-power differences between the motor imagery classes. This criterion reflects the underlying physiology of event-related desynchronization (ERD) / event-related synchronization (ERS) [22].

B. Common Spatial Patterns

The Common Spatial Patterns (CSP) method (e.g. [3] [4] [5] [6] [7]) is probably the most popular algorithm for computing spatial filters in motor imagery experiments. It is well suited to discriminate different mental states induced by motor imagery as it maximizes the band-power ratio between two motor imagery classes. The spatial filters can be computed by solving the generalized eigenvalue problem

$$\mathbf{\Sigma}_1 \mathbf{w}_i \quad = \quad \lambda_i \mathbf{\Sigma}_2 \mathbf{w}_i \tag{3}$$

with $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ being the $D \times D$-dimensional average covariance matrices estimated from two different motor imagery classes. Note that the generalized eigenvalues λ_i measure the variance ratio between class 1 and class 2. A large λ_i indicates high variance of class 1 and low variance of class 2, a small λ_i indicates the opposite. Since the goal is to extract a subspace with large band-power differences between both conditions, irrespectively whether the variance of class 1 or class 2 is high, we sort the extracted spatial filters $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_D]$ according to their ability to capture these differences in decreasing order $\alpha_1 = \max\{\lambda_1, \frac{1}{\lambda_1}\} > \ldots > \alpha_D = \max\{\lambda_D, \frac{1}{\lambda_D}\}$.

C. Limits of standard CSP

The CSP method computes spatial filters in a naive data-driven manner. This makes the algorithm vulnerable and may produce suboptimal results, i.e. do not extract the true motor imagery related activity, in certain situations.

One major source of errors results from the difficulty to properly estimate the class covariance matrices. Since poorly estimated covariance matrices do not well represent the underlying neural processes this will directly affect the spatial filter computation (e.g. [26] [7] [27]). The increasing number of electrodes used in BCI experiments further complicates the estimation problem. Thus if data is scarce it is almost impossible to reliably estimate the high-dimensional covariance matrices without prior information or regularization. Furthermore, the covariance matrix estimation may be negatively affected by

EEG artifacts like eye blinks or loose electrodes. These artifacts often have much more signal power than BCI related activity thus if not properly removed (e.g. [15] [3] [16]) they may dominate the covariance matrix estimation and lead to overfitted CSP solutions.

Another class of problems results from variations of the extracted features. We call this the *non-stationarity problem*. Note that we use the term non-stationarity to denote changes in the feature distribution, irrespectively whether these changes occur within an experimental session, between sessions or between different subjects. In the following we comment on these different types of non-stationarity.

Within-session changes in the signal are very common and may occur on different time scales [18]. For instance, artifacts like loose electrodes, muscle movements, blinking, swallowing, teeth crunching or sudden shifts of attention usually affect one or few trials whereas effects of tiredness, changes in impedance of the electrodes or learning effects are only visible on larger time scales. Note that these changes may not only corrupt the covariance estimation but also lead to overfitted CSP solutions and increase the variability of the extracted features (c.f. [18] [19] [28]). Since many of these changes can not be avoided, the application of robust algorithm becomes crucial for successful BCI operation.

Between-session non-stationarity can be often observed in BCI experiments (c.f. [17] [29] [30]). There are several reasons why data recorded in one session is different from data recorded in another session (which may be on a different day), e.g. the calibration of the system may be different, the state of mind of the subject may differ and the position of the electrodes may not be exactly the same. Furthermore we often observe significant changes when moving from calibration phase to feedback application [17]. These changes may be due to addition processing induced by the visual or auditory feedback which is often lacking when calibrating the system. Users may also change the strategy to control the BCI when knowing the result of the classification. Another scenario where CSP may produce suboptimal results is when it focuses on discriminative but not motor imagery related activity. For instance assume we use a visual cue (arrow pointing to the left or right) in the training phase to indicate the motor imagery class. A subject may involuntarily perform tiny eye movements when observing the cue, i.e. move the eyes to the direction of the arrow. These ocular movements can induce task-related activity that will be captured by the CSP spatial filters. However, this activity is not related to motor imagery thus it becomes meaningless and may deteriorate classification performance if the cue is lacking in subsequent sessions [31].

Finally non-stationarity can also be defined in terms of differences between subjects (we should rather use "heterogeneity" though). This kind of variations may be not relevant when training a single-subject system, but certainly play a role when aiming for user-independent BCIs or shorter calibration times [32]. Several authors also proposed to utilize other subjects' data to improve the spatial filter computation when calibration data is scarce (e.g. [33] [34]). Differences in the signal distribution of different subjects may have many reasons. They may be due to differences in the electrode positions or the user's state of mind, but also anatomical differences, e.g. size of the head, may play a significant role. It is usually advisable to weight the contributions from other subjects according to their relevance (c.f. [35] [34]).

D. *State-of-the-Art CSP Variants*

In this subsection we review some of the recently proposed spatial filter computation algorithms. Note that we mainly focus on CSP-like methods and do not include spatio-spectral algorithms like CSSP [36] or adaptation strategies like [17] [37]. Furthermore we ignore the BCI-related work on artifact identification and removal e.g. by using Independent Component Analysis. Figure 1 gives an overview over the presented CSP variants.

Robust Estimation
Several strategies have been proposed to improve the estimation of the covariance matrix when applying CSP. For instance, the authors of [27] and [38] robustly estimate the covariance matrices by using M-estimators. Regularization of the covariance matrix e.g. [26] [35] [39] [7] is also one common approach to increase robustness, especially in small-sample settings. Other authors [40] [41] [42] propose to improve the CSP solution by performing channel selection or enforcing sparsity on the spatial filters. The idea of computing spatial filters in a region of interest was used in [43] [44] [45]. The authors of [46] [28] propose a maxmin approach to robustify the CSP algorithm. Many other variants of the algorithm use some kind of regularization in order to incorporate a priori information [47], avoid overfitting [7] or reduce ocular artifacts [48]. Other methods robustify the variance estimation in CSP by applying L_p-norms [49] [50]. A generative CSP model using the robust Student-t distribution was proposed in [51]. The authors of [16] apply trial pruning in order to separate signal from noise and [5] discusses several methods for minimum noise estimation. A novel robust CSP algorithm based on beta divergence was proposed in [52].

Stationary Features
Recently, the development of methods compensating for non-stationarities has gained increased attention in many application fields of machine learning including Brain-Computer Interfacing (see e.g. [53] [54]). The stationary CSP approach [18] regularizes the CSP solution towards stationarity in a data-driven manner. The authors of [19] use the same idea but apply Kullback-Leibler divergence to measure the changes in the data. Two-step approaches [55] [56] [57] 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. Furthermore a second-order baseline was used [58] to robustify the algorithm against time and subject related variations. Robust feature extraction methods have been proposed for reducing between-session non-stationarities (e.g. [31] [30] [59]). Some approaches [32] [29] [60] utilize data collected in previous sessions for this task, others [61]

update the trained model using adaptation.

Multi-Subject Methods
Many recent algorithms improve the CSP solution by incorporating data from other subjects. Such approaches are especially important when aiming for a subject independent BCI system or reducing the calibration time of the system. The authors of [34] jointly train the spatial filters of several subjects by applying a multi-task learning algorithm. A Bayesian method for subject-to-subject information transfer has been proposed in [62]. Data from other users have also been used as regularization target by [33] and [35]. A recently proposed method [63] incorporates information from other subjects by applying Multiple Kernel Learning.

Other approaches
Some CSP variants improve the quality of the solution by explicitly considering the temporally local structure of observed samples [64] [65] [66]. Other algorithms were specifically designed for multi-class problems and optimize the solution by using information theory [67], joint approximate diagonalization [68] [69] or Kullback-Leibler Divergence [70]. Recently, a spatial filtering method directly linking to Bayes classification error was proposed in [71]. A Wavelet CSP method for asynchronous BCI systems was proposed in [72] whereas the authors of [73] improve the discriminative capability of CSP by taking into account both the amplitude and phase components of the EEG signal. A CSP variant directly optimizing the discriminativity of the features was proposed in [74]. A recently proposed approach [75] learns spatial filters by considering signal propagation and volume conduction effects.

Note that many existing methods try to improve the classification step rather than the CSP computation. For instance, the method presented in [76] incorporates information from other subjects by applying multi-task learning whereas the authors of [77] [78] propose different adaptation strategies to cope with non-stationarity. Some authors omit the CSP computation step and suggest to jointly perform feature extraction and classification (e.g. [79] [80]). Other approaches [81] [82] omit spatial filtering by directly performing classification on the manifold of covariance matrices.

III. CSP AS DIVERGENCE MAXIMIZATION PROBLEM

This section introduces the divergence-based framework (divCSP) for computing spatial filters. Note that we extend our previous contribution [52] in several ways. First, we introduce and compare two different optimization algorithms for divCSP, a subspace approach that optimizes the objective function in the whole subspace and a deflation method that applies optimization in a sequential manner. Furthermore we propose three different regularization schemes for tackling the non-stationarity problems in BCI, moreover, we also show that subject-independent spatial filters can be extracted with our method by jointly optimizing the divergence problems of multiple users. Finally we discuss the effects of using beta divergence in our optimization framework and relate the proposed methods to state-of-the-art algorithms.

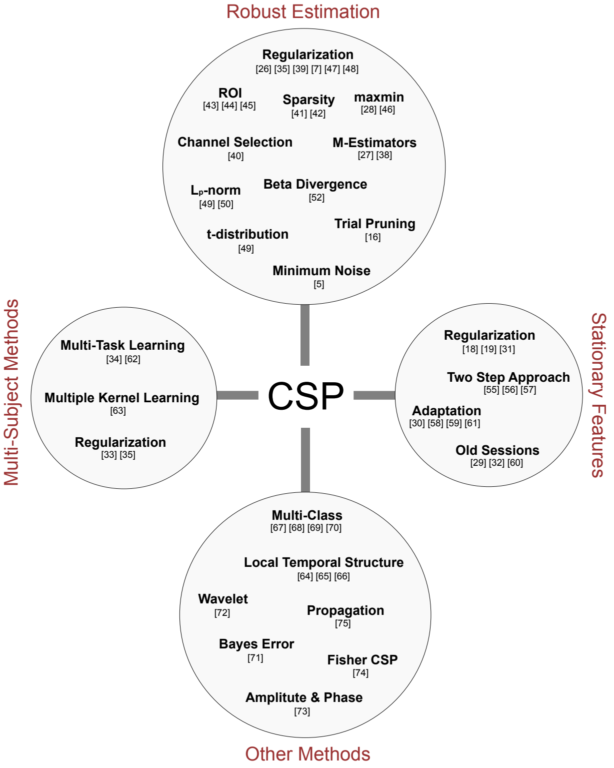

Fig. 1. Overview over different state-of-the-art CSP variants.

A. Divergence-Based Framework

Many machine learning algorithms, e.g. Independent Component Analysis [83] or Stationary Subspace Analysis [84], can be cast into the framework of information geometry [85] and formulated as divergence optimization problems. In our previous conference paper [52] we showed that the Common Spatial Patterns (CSP) algorithm can also be interpreted from this perspective, in particular we showed that the subspace extracted by CSP maximizes the symmetric Kullback-Leibler (KL) divergence between the distributions of both classes. Note that the symmetric Kullback-Leibler Divergence $\tilde{D}_{kl}$ between distributions $f(x)$ and $g(x)$ is defined as

$$\int f(x) \log \frac{f(x)}{g(x)} dx \;+\; \int g(x) \log \frac{g(x)}{f(x)} dx. \tag{4}$$

It can be interpreted as distortion measure between two probability distributions, thus it is always positive and equals zero if and only if $g = f$. Note that in this paper we always compute divergences between zero mean Gaussian distributions. The following relation exists between the spatial filters extracted by CSP and the symmetric Kullback-Leibler divergence.

Theorem: Let $\mathbf{W} \in \mathbb{R}^{D \times d}$ be the top d (sorted by α_i) spatial filters computed by CSP and let $\mathbf{V}^\top = \tilde{\mathbf{R}}\mathbf{P} \in \mathbb{R}^{d \times D}$ be a matrix that can be decomposed into a whitening projection $\mathbf{P} \in \mathbb{R}^{D \times D}$ with $(\mathbf{P}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{P}^\top = \mathbf{I})$ and an orthogonal projection $\tilde{\mathbf{R}} = \mathbf{I}_d\mathbf{R} \in \mathbb{R}^{d \times D}$ with $\mathbf{I}_d \in \mathbb{R}^{d \times D}$ being the identity matrix truncated to the first d rows and $\mathbf{R}^\top\mathbf{R} = \mathbf{I} \in \mathbb{R}^{D\times D}$. Then

$$\begin{aligned} \mathrm{span}(\mathbf{W}) &= \mathrm{span}(\mathbf{V}^*) &(5)\\ \text{with } \mathbf{V}^* &= \underset{\mathbf{V}}{\mathrm{argmax}}\, \tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}_1\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}_2\mathbf{V}\right) &(6)\end{aligned}$$

Here $\mathrm{span}(\mathbf{M})$ means the subspace spanned by the columns of the matrix $\mathbf{M}$.

Proof: See Appendix.

The theorem says that the CSP filters $\mathbf{W}$ project the data to a subspace with maximum discrepancy, measured by symmetric Kullback-Leibler divergence, between the d-dimensional Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{W}^\top\mathbf{\Sigma}_1\mathbf{W})$ and $\mathcal{N}(\mathbf{0}, \mathbf{W}^\top\mathbf{\Sigma}_2\mathbf{W})$. Thus instead of computing spatial filters with CSP we obtain an equivalent solution (up to linear transformations within the subspace) when maximizing Eq. (6). Note that [70] has provided a proof for the special case of one spatial filter, i.e. for $\mathbf{V} \in \mathbb{R}^{D\times 1}$. In the following we present two approaches for divergence maximization, namely the subspace method and the deflation algorithm. Note that our optimization framework is based on the work [86] [87] [57].

B. Optimization Algorithms

Subspace Method

Let us first describe the *subspace* approach (see Algorithm 1). The first step of the method consists of the computation of a whitening matrix $\mathbf{P} \in \mathbb{R}^{D\times D}$ that projects the data onto the unit sphere, i.e. $\mathbf{P}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{P}^\top = \mathbf{I}$. This whitening transformation is applied to the class covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ followed by a (random) rotation with $\mathbf{R}_0 \in \mathbb{R}^{D\times D}$. Note that the rotation matrix satisfies $\mathbf{R}_0^\top\mathbf{R}_0 = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The optimization process then consists of finding a rotation matrix $\mathbf{R} \in \mathbb{R}^{D\times D}$ that maximizes the symmetric KL divergence in the first d sources. More precisely we optimize the following objective function

$$\begin{aligned} \mathcal{L}_{kl}(\mathbf{R}) &= \tilde{D}_{kl}\left(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d^\top \;||\; \mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d^\top\right) &(7)\\ &= \frac{1}{2}\mathrm{tr}\Big((\mathbf{I}_{\mathrm{d}}\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_{\mathrm{d}}^\top)^{-1}(\mathbf{I}_{\mathrm{d}}\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_{\mathrm{d}}^\top) \\ &+ (\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d^\top)^{-1}(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d^\top)\Big) - d,\end{aligned}$$

where $\tilde{\mathbf{\Sigma}}_1$ and $\tilde{\mathbf{\Sigma}}_2$ denote the whitened covariance matrices and $\mathbf{I}_d \in \mathbb{R}^{d\times D}$ is the identity matrix truncated to the first d rows. Note that although $\mathbf{R}$ is a $D \times D$ rotation matrix, we only evaluate the first d rows of it, i.e. we only evaluate the divergence in a d-dimensional subspace.

The optimization is performed by gradient descend on the manifold of orthogonal matrices. More precisely, we start with an orthogonal matrix $\mathbf{R}_0$ and find an orthogonal update $\mathbf{U}$ in the k-th step such that $\mathbf{R}_{k+1} = \mathbf{U}\mathbf{R}_k$. This ensures that we stay on the manifold of orthogonal matrices at each step. Note that the update matrix can be written as a matrix exponential of a skew-symmetric matrix $\mathbf{M} = -\mathbf{M}^\top$. We find a search direction $\mathbf{H} = -\mathbf{H}^\top$ in the set of skew symmetric matrices by computing the gradient of the loss function w.r.t. $\mathbf{M}$ at $\mathbf{M} = \mathbf{0}$ and determine the optimal step size t along this gradient by line search (see [88] and [87] for details). Finally we represent the update matrix as $\mathbf{U} = e^{t\mathbf{H}}$. Since the objective function in Eq. (7) is invariant to rotations within the d-dimensional subspace[2], we rotate the projection matrix $\mathbf{V}$ in the last step of the algorithm with a matrix $\mathbf{G}$, so that it maximally separates the classes along the projection directions (as is the case with CSP). The spatial filters can be rearranged so that they capture the class differences with decreasing strength (α_i sorting).

Algorithm 1 Subspace divCSP

```
function SUB-DIVCSP($\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, d$)
    Compute the whitening matrix $\mathbf{P} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-\frac{1}{2}}$
    Initialise $\mathbf{R}_0$ with a (random) rotation matrix
    Whiten and rotate $\mathbf{\Sigma}_{\{1/2\}} = (\mathbf{R}_0\mathbf{P})\mathbf{\Sigma}_{\{1/2\}}(\mathbf{R}_0\mathbf{P})^\top$
    repeat
        Compute the gradient matrix
        Determine the optimal step size
        Update the rotation matrix $\mathbf{R}_{k+1} = \mathbf{U}\mathbf{R}_k$
        Apply the rotation to data $\mathbf{\Sigma}_{\{1/2\}} = \mathbf{U}\mathbf{\Sigma}_{\{1/2\}}\mathbf{U}^\top$
    until convergence
    Let $\mathbf{V}^\top = \mathbf{I}_d\mathbf{R}_{k+1}\mathbf{P}$
    Compute the eigenvectors $\mathbf{G} \in \mathbb{R}^{d\times d}$ of $\mathbf{V}^\top\mathbf{\Sigma}_1\mathbf{V}$
    Let $\mathbf{V}^* = \mathbf{V}\mathbf{G}$ and rearrange filters ($\alpha_i$ sorting)
    return $\mathbf{V}^*$
end function
```

Deflation Method

A further interesting algorithm is the *deflation* method. It does not extract the whole subspace at once, but performs the optimization in a sequential manner (see also deflation FastICA [89]). More precisely, the algorithm reduces the dimensionality of the data space by one in each step. This provides a sorting of the spatial filters that is analogous to CSP, i.e. the first solution is the most discriminative one and so on. The different steps of the method are described in Algorithm 2. In the first steps of the algorithm we apply the whitening transformation $\mathbf{P}$ to the class covariance matrices and initialize a matrix $\mathbf{B}$ that represents the basis of the subspace in which the spatial filters are computed. Then we repeat the following procedure d times. We calculate the best spatial filter by applying the subspace divCSP algorithm described in Algorithm 1 with parameter $d = 1$. Note that we skip the whitening step as it has already been performed. After obtaining the spatial filter $\mathbf{w}$ we compute its corresponding orthogonal complement and project the class covariance matrices to this subspace. This step ensures that the spatial filters computed in subsequent steps will be orthogonal to the current ones. Since the i-th spatial filter $\mathbf{w}$ has been computed in the subspace with basis $\mathbf{B}$ its representation in the original coordinate system is $\mathbf{v}_i = \mathbf{B}\mathbf{w}$. In the last step of the loop we update the basis matrix $\mathbf{B}$. The final solution consists of the spatial filters $\mathbf{v}_i$ with $i = 1 \ldots d$ and is already sorted according to α_i.

[2] $\tilde{D}_{kl}(\mathbf{C}_1 \;||\; \mathbf{C}_2) = \tilde{D}_{kl}\left(\mathbf{G}^\top\mathbf{C}_1\mathbf{G} \;||\; \mathbf{G}^\top\mathbf{C}_2\mathbf{G}\right)$ for a square matrix $\mathbf{G}$ with $|\mathbf{G}| \neq 0$.

The deflation algorithm optimizes the following objective function in the i-th step

$$\tilde{D}_{kl}\left(\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{w}_i \;\;||\;\; \mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{w}_i\right) = \frac{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{w}_i}{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{w}_i} + \frac{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{w}_i}{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{w}_i} \quad (8)$$

$$s.t. \quad \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall j \in 1 \dots i-1. \quad (9)$$

Note that this objective function can be written as $f(z) = z + \frac{1}{z}$ with $z = \frac{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{w}_i}{\mathbf{w}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{w}_i}$ and one can prove easily that this function is maximized at the border. Thus it is maximized either for the largest z or for the smallest one (largest $\frac{1}{z}$). This solution corresponds to the i-th CSP spatial filter (sorted by α_i). Thus both methods, subspace and deflation, provide the same spatial filters, namely the CSP ones, when applied to the objective function of divCSP (see Eq. (6)). However, when applied to the regularized version of divCSP (described in next section) the solution of the subspace and deflation method will not necessarily coincide. This is because the objective function consists of a sum of divergences and only the subspace method considers (changes in) correlations between sources extracted by different spatial filters[3]. We discuss this difference between both optimization schemes in Section V using simulations.

Algorithm 2 Deflation divCSP

```
function DEF-DIVCSP($\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, d$)
    Compute the whitening matrix $\mathbf{P} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-\frac{1}{2}}$
    Apply whitening $\tilde{\boldsymbol{\Sigma}}_{\{1/2\}} = \mathbf{P}\boldsymbol{\Sigma}_{\{1/2\}}\mathbf{P}^\top$
    Initialize basis $\mathbf{B} = \mathbf{I} \in \mathbb{R}^{D \times D}$
    for i=1...d do
        Compute $\mathbf{w} \in \mathbb{R}^{(D-i+1) \times 1}$ by sub-divCSP
        Compute $\mathbf{W}^\perp \in \mathbb{R}^{(D-i+1) \times (D-i)}$ the orthogonal
          complement of $\mathbf{w}$
        Project $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ to subspace by $\mathbf{W}^\perp$
        Reproject $\mathbf{w}$ to original space by $\mathbf{v}_i = \mathbf{B}\mathbf{w}$
        Update basis $\mathbf{B} = \mathbf{B}\mathbf{W}^\perp \in \mathbb{R}^{D \times (D-i)}$
    end for
    Let $\mathbf{V}^* = \mathbf{P}[\mathbf{v}_1 \dots \mathbf{v}_d]$
    return $\mathbf{V}^*$
end function
```

IV. Unifying CSP Framework

In this section we extend divCSP by adding different regularization schemes to the objective function. Furthermore we introduce the beta divergence variant of the algorithm. Finally we show that our novel framework unifies many of the state-of-the-art CSP variants in a principled manner. Figure 2 gives an overview over different application scenarios of our framework.

A. Invariance Through Regularization

Above we proved that CSP can be formulated in a divergence maximization framework, however, maximizing the band power ratios may not be the only objective for feature extraction. For instance, imposing stationarity on the extracted features is also of high interest in Brain-Computer Interfacing (e.g. [18] [31] [19]). A natural way of regularizing the extracted spatial filters towards stationarity is to combine the objective function of divCSP with a divergence term that accounts for the stationarity of the features. We propose to tackle the different non-stationarity problems by adding different such regularization terms. Since the optimization process is not affected by changing the way how stationarity is measured (as long it is a divergence), our framework integrates several stationary CSP variants and permits utilizing information from other subjects.

The objective function of the proposed regularized divCSP method can be written as

$$\mathcal{L}(\mathbf{V}) = \underbrace{(1-\lambda)\tilde{D}_{kl}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_1 \mathbf{V} \;\;||\;\; \mathbf{V}^\top \boldsymbol{\Sigma}_2 \mathbf{V}\right)}_{\text{CSP Term}} - \underbrace{\lambda \Delta}_{\text{Reg. Term}} \quad (10)$$

where Δ is the regularization term that can be arbitrarily defined, depending on the type of invariance that we want to achieve, and λ is a regularization parameter trading-off the influence of the CSP objective function and the regularization term. Note that the objective functions of all algorithms presented in this work can be written as weighted sum of divergences and the goal is to find a projection to a d-dimensional subspace that maximizes this sum. In the following we will discuss four different regularization terms.

Within Session Stationarity (divCSP-WS): In order to reduce the influence of artifacts or shifts that are present in the training data we divide the data into a set of smaller epochs. The epochs consist of concatenated recordings of one or several subsequent trials of the same class. The non-stationarity of the extracted features is measured as average divergence between the data distribution of the epochs and the whole data distribution for each class separately (see [57]). More precisely, we compute

$$\boldsymbol{\Delta} = \frac{1}{2N} \sum_{c=1}^{2} \sum_{i=1}^{N} D_{kl}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_c^i \mathbf{V} \;\;||\;\; \mathbf{V}^\top \boldsymbol{\Sigma}_c \mathbf{V}\right), \quad (11)$$

where N denotes the number of trials and $\boldsymbol{\Sigma}_c^i$ stands for the estimated covariance matrix of class c and epoch i. Note that we use the Kullback-Leibler divergence (and not its symmetric version) for capturing the changes; the reasons for that will be explained in Section V. Adding the regularization term $\boldsymbol{\Delta}$ to Eq. (10) reduces the within-class variability of the extracted training features.

Between Session Stationarity (divCSP-BS): The purpose of the next regularization term is to reduce the shift between the data distribution in calibration and feedback phase (e.g. [17] [77]). Since we may assume that feedback data is not available at the time of computing the spatial filters we utilize information from other subjects to estimate these changes. Note that this approach implicitly assumes that the between-session non-stationarities are similar among different users, e.g. because they are induced by the change in experimental

[3] Note that due to whitening there is no correlation between the sources when only considering the divergence between the average class covariance matrices.

paradigm (no feedback vs. visual feedback). It is based on the recently proposed idea of transferring non-stationary information between subjects [31]. For our experiments we consider the following regularization term

$$\mathbf{\Delta} = \frac{1}{2K} \sum_{c=1}^{2} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \mathbf{\Sigma}_{tr,c}^{k} \mathbf{V} \;\;||\;\; \mathbf{V}^\top \mathbf{\Sigma}_{te,c}^{k} \mathbf{V} \right), \qquad (12)$$

where K stands for the number of other subjects and $\mathbf{\Sigma}_{tr,c}^{k}$ and $\mathbf{\Sigma}_{te,c}^{k}$ denote the class covariance matrices estimated on training and test data of subject k.

Across Subject Stationarity (divCSP-AS): If the goal is to reduce differences between subjects, i.e. because one assumes that the underlying processes governing motor imagery are very similar between users, then one may use the changes between the average data of the subject of interest ℓ and the data of other subjects k as regularization term

$$\mathbf{\Delta} = \frac{1}{2K} \sum_{c=1}^{2} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \mathbf{\Sigma}_{tr,c}^{\ell} \mathbf{V} \;\;||\;\; \mathbf{V}^\top \mathbf{\Sigma}_{tr,c}^{k} \mathbf{V} \right). \qquad (13)$$

Multi-Subject CSP (divCSP-MS): Instead of combining the discriminativity term with a regularization term that captures non-stationarity, we can also combine it with the divCSP objective functions of other subjects. This permits extracting a more subject-independent feature space. In this case we need to invert the sign of Δ as we aim to maximize this regularization term

$$\mathbf{\Delta} = -\frac{1}{K} \sum_{k=1}^{K} \tilde{D}_{kl} \left(\mathbf{V}^\top \mathbf{\Sigma}_{1}^{k} \mathbf{V} \;\;||\;\; \mathbf{V}^\top \mathbf{\Sigma}_{2}^{k} \mathbf{V} \right). \qquad (14)$$

Many other forms of regularization, e.g. considering multiple classes or containing priori information, can be easily integrated into our framework.

B. Robustness Through Beta Divergence

Once a divergence formulation is established it becomes possible to define the same algorithm using other divergences (cf. [52]). The application of beta divergence [90] [91], a generalization of Kullback-Leibler divergence, is especially promising for BCI application as it robustly averages the terms in Eq. (10) by downweighting the influence of outlier terms [52]. Eguchi and Kanno [90] discussed that Bregman divergence [92] including beta and KL divergences share the same property when the function controlling sample weights is properly designed.

Beta divergence D_β between distributions $g(x)$ and $f(x)$ has been proposed in [90] [91] and is defined (for $\beta > 0$) as

$$\frac{1}{\beta} \int (g^\beta(x) - f^\beta(x)) g(x) dx \;\; - \qquad (15)$$
$$\frac{1}{\beta + 1} \int (g^{\beta+1}(x) - f^{\beta+1}(x)) dx,$$

where $g(x)$ and $f(x)$ are two probability distributions. Note that [93] extends the definition of beta divergence to $\beta \in \mathbb{R}$, however, since our algorithms require that g and f are Gaussian we will use $\beta \geq 0$ (or $\beta > c$ with c being a small negative data-dependent value; see derivation in appendix for more details). One can show easily that beta divergence and Kullback-Leibler divergence coincide as $\beta \to 0$. Thus beta divergence can be seen as a generalization of Kullback-Leibler divergence. The symmetric beta divergence between two zero mean d-dimensional Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ can be written in explicit form [52] as

$$\gamma \Bigg(|\mathbf{\Sigma}_1|^{-\frac{\beta}{2}} + |\mathbf{\Sigma}_2|^{-\frac{\beta}{2}} - \qquad (16)$$
$$(\beta+1)^{\frac{d}{2}} \left(\frac{|\mathbf{\Sigma}_2|^{\frac{1-\beta}{2}}}{|\beta \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2|^{\frac{1}{2}}} + \frac{|\mathbf{\Sigma}_1|^{\frac{1-\beta}{2}}}{|\beta \mathbf{\Sigma}_2 + \mathbf{\Sigma}_1|^{\frac{1}{2}}} \right) \Bigg),$$

with $\gamma = \frac{1}{\beta} \sqrt{\frac{1}{(2\pi)^{\beta d} (\beta+1)^d}}$. Since the gradient can also be represented explicitly we can directly apply Algorithm 1 and Algorithm 2 for computing subspaces with maximum (sums of) beta divergence.

A robust CSP method using beta divergence has been proposed in [52]. Note that this method can be easily incorporated into our framework by setting $\lambda = 1$ and using

$$\mathbf{\Delta} = -\sum_{i=1}^{N} \tilde{D}_\beta \left(\mathbf{V}^\top \mathbf{\Sigma}_1^i \mathbf{V} \;\;||\;\; \mathbf{V}^\top \mathbf{\Sigma}_2^i \mathbf{V} \right), \qquad (17)$$

where $\mathbf{\Sigma}_1^i$ and $\mathbf{\Sigma}_2^i$ denote the covariance matrix estimated from i-th trial of class 1 and 2 (assuming both classes have the same number of trials), respectively. Since this method has been extensively evaluated in [52], we will not consider it in the current work.

Instead we will show that using beta divergence adds a degree of freedom to the regularization approaches presented above, more precisely it allows to specify (by changing the β parameter) the type of non-stationarity we want to become invariant to. For instance, using beta divergence with small β in divCSP-WS penalizes single extreme events (as they are not downweighted thus will dominate $\mathbf{\Delta}$) with large deviation from the average activity, e.g. electrode artefacts, whereas larger β parameters penalize more stable variations that occur throughout the experiment. We will discuss this property of beta divergence in the next section in more detail.

Note that also the multi-subject algorithm divCSP-MS may profit from using beta divergence because integrating data from several subjects usually requires subject selection, e.g. [33] [34], as different users may have very different signal properties due to differences in head size, electrode montage, state of mind etc. With increasing[4] β parameter we implicitly perform this kind of subject selection as the influence of "outlier subjects" that have very different signal characteristics will be reduced. On the other hand in some applications we are interested in the similarity between subjects and want to identify "outlier subjects". When using our framework with a small β parameter the focus of the optimization shifts from identifying activity that is common to all subjects to identifying extreme activity that only occurs in one or few users. This property makes beta divergence a valuable tool that can not only be used to fine tune the type of invariance

[4] Note that the β parameter has the opposite effect in divCSP-MS than in divCSP-WS due to the inversion of the sign in $\mathbf{\Delta}$.

but can also be very helpful for exploratory analysis.

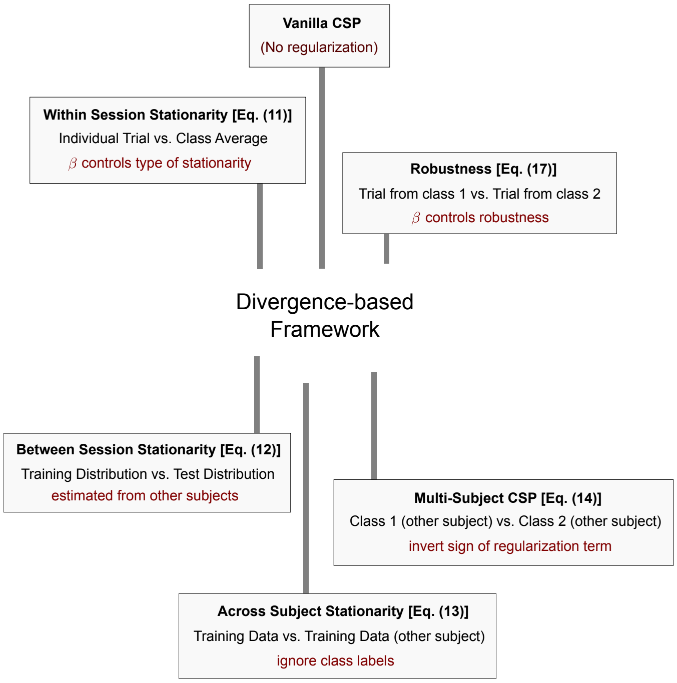
Vanilla CSP
(No regularization)
Within Session Stationarity [Eq. (11)]
Individual Trial vs. Class Average
β controls type of stationarity
Robustness [Eq. (17)]
Trial from class 1 vs. Trial from class 2
β controls robustness
Divergence-based Framework
Between Session Stationarity [Eq. (12)]
Training Distribution vs. Test Distribution
estimated from other subjects
Multi-Subject CSP [Eq. (14)]
Class 1 (other subject) vs. Class 2 (other subject)
invert sign of regularization term
Across Subject Stationarity [Eq. (13)]
Training Data vs. Training Data (other subject)
ignore class labels

Fig. 2. Unifying divergence-based CSP framework. Each box shows a particular application scenario of the framework. It also describes how to compute the regularization term $\boldsymbol{\Delta}$ for this scenario, e.g. in order to achieve within session stationarity we need to compute the divergences between individual trials and the class average (see Eq. (11)).

C. Relations to State-of-the-Art Methods

In the following we comment on the relations between our novel framework and several state-of-the-art CSP algorithms.

Relation to Stationary Subspace Analysis
Stationary Subspace Analysis (SSA) [86] is an algorithm that decomposes a multivariate signal into a stationary and non-stationary part. Different variants of the algorithm have been used to extract stationary CSP features [86] [57] [56]. All these approaches consist of two steps, namely the removal of non-stationarities and the computation of spatial filters in the stationary subspace. This is an important difference to the proposed framework that simultaneously optimizes the CSP objective and stationarity of the features. Note that two-step approaches may be suboptimal [80] as information that is relevant for the second step can be removed in the first step. Two-step approaches can be regarded as hard regularization methods as they remove some part of the data in the preprocessing step. The divCSP approaches proposed in this work on the other hand are soft regularization methods regularizing the filters towards stationarity with subject dependent strength. The SSA+CSP variant proposed in [56] aims to ensure that the removed non-stationarities do not contain discriminative information, however, the discriminativity of the subspace is measured using Kullback-Leibler (KL) whereas our framework uses the symmetric KL divergence for this task. Since the symmetric KL divergence has a direct relation to CSP it is more appropriate. Our framework is also more generic than SSA+CSP as it e.g. considers multiple types of non-stationarity, permits utilizing data from other users, offers a deflation mode of optimization and can be used with other divergences.

Relation to Stationary CSP
The work on stationary CSP (sCSP) [18] originally introduced the idea of regularizing spatial filters towards stationarity. The method tackles the within-session non-stationarity problem, thus it is related to divCSP-WS. Stationary CSP has been shown to robustify the solution against artifacts in the data and to increase stationarity of the features, thus from a conceptual point of view both methods are very similar. The penalty matrix of stationary CSP is computed by using a heuristic, namely flipping the sign of negative eigenvalues, and one can show (see Section V) that this heuristic may fail. Our novel divCSP-WS method on the other hand measures non-stationarity in a principled way by using divergences. Both divCSP-WS and sCSP perform soft regularization, however, sCSP does not capture the changes in correlation between the extracted spatial filters, thus it can be regarded as a deflation method. In contrast to our method sCSP does not require gradient descent optimization, thus it is computationally more efficient.

Relation to Kullback-Leibler CSP
Recently, [19] proposed a method (KLCSP) that combines the CSP objective function with the non-stationarity term proposed in [57]. Although this method is very similar to divCSP-WS, one can identify some important differences. The KLCSP approach is a deflation method i.e. it first extracts the most important filter, then the second most one and so on. Our method, when optimized with the subspace algorithm, evaluates the objective function on the whole subspace, thus it also measures non-stationarities in the correlations between the extracted sources which are not captured in [19]. We will show later that optimizing the whole subspace at once may be useful when integrating data from different subject as it permits extracting the optimal subspace irrespectively of differences in source correlation between users. Furthermore our method uses a consistent formulation that can be fully interpreted as divergence maximization i.e. it has an information geometric [85] [93] interpretation. Thus we can easily change the divergence function and with it the properties of the solution. On the other hand KLCSP heuristically combines the CSP objective with a divergence-based penalty term, therefore it lacks this information geometric interpretation. KLCSP has only been applied to one kind of non-stationarity, namely the within-session changes, whereas we propose a generic CSP framework that can be used with different types of regularization and different types of divergences. Both methods also differ in the way the optimization is performed.

Relation to Stationary Subspace CSP
The Stationary Subspace CSP method (ssCSP) [31] estimates the changes between the calibration and feedback session by using other subjects' data. Thus from a conceptual point of view it is very similar to divCSP-BS. However, in contrast to divCSP-BS it performs hard regularization, i.e. it completely

TABLE I
PROPERTIES OF THE DIFFERENT STATIONARY CSP VARIANTS

Method	Non-Stationarity	Regularization	Mode	Optimization
SSA+CSP	Within-Sess.	Hard	Sub.	Gradient
sCSP	Within-Sess.	Soft	Defl.	Eigenvalue
KLCSP	Within-Sess.	Soft	Defl.	Newton
ssCSP	Between-Sess.	Hard	Defl.	Eigenvalue
mtCSP	Multi-Subj.	Soft	Defl.	Newton

removes some of the non-stationary directions (by using a large regularization constant). Furthermore it only considers class unrelated changes whereas divCSP-BS evaluates the variations for each class separately. The non-stationary subspace in ssCSP is constructed by computing the eigenvectors with largest absolute eigenvalues of the difference between the training and test covariance matrix. Although these vectors span a subspace with maximum symmetric KL divergence between those covariance matrices[5], the final non-stationary subspaces of ssCSP and divCSP-BS do not coincide as both methods differ in the way information from different subjects is integrated.

Relation to Multi-Subject CSP Methods

A method (multi task CSP or mtCSP) that is related to divCSP-MS has been proposed by [34]. The authors decompose the spatial filters into a subject specific and a general part and formulate the CSP optimization for all subjects as multi-task learning problem. Our divCSP-MS does not perform a joint learning of all spatial filters for all subjects, but rather regularizes a subject specific set of spatial filters so that they are as useful as possible for the other subjects. Furthermore our method can learn the whole set of spatial filters at once (subspace scheme) whereas mtCSP is a deflation method. Note that the subspace approach may be superior for the multi-subject problem as it permits extracting the common activity even when subjects have different correlations e.g. due to differences in the electrode montage or head size. We will illustrate this point in the simulations. Above all by using beta divergence we perform implicit subject selection by downweighting the contributions from "outlier subjects". The authors of [34] propose a clustering-based approach for this task.

In the context of related methods we also want to mention the work of [70]. The author proposes a divergence-based method for solving the CSP multi-class problem. We are not aware of any work that is similar to divCSP-AS. The properties of the different stationary CSP variants are summarized in Table I.

V. SIMULATIONS

This section investigates situations in which some of the state-of-the-art CSP methods fail to extract the optimal spatial filters. Since the proposed framework is very flexible it can be adapted to give the correct solution in all these cases.

A. One step vs. Two step methods

Two-step methods perform hard regularization, thus once relevant information is removed in the first step these methods are bound to fail [80]. The following example shows a situation where the recently proposed two step methods [55] [57] fail to extract the true spatial filters.

Consider the observed signal $\mathbf{x}(t) \in \mathbb{R}^{10}$ generated as mixture of 10 sources $\mathbf{s}(t) = [s_1(t) \dots s_{10}(t)]^\top$ with a random orthogonal mixing matrix $\mathbf{A} \in \mathbb{R}^{10\times 10}$

$$\mathbf{x}(t) \quad = \quad \mathbf{A}\mathbf{s}(t). \tag{18}$$

Assume sources s_1 and s_2 are non-stationary. The signal of the first source is sampled from $\mathcal{N}(0,\sigma_1)$ for class 1 and $\mathcal{N}(0,\sigma_2)$ for class 2, whereas the signal of source s_2 is sampled from $\mathcal{N}(0,\sigma_3)$ irrespectively of class. All the other sources generate normally distributed data with zero mean and unit variance. Now let $\sigma_1 = 1.2 + \epsilon_1$, $\sigma_2 = 0.8 + \epsilon_1$ and $\sigma_3 = 1 + \epsilon_2$ with $\epsilon_1 \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ and $\epsilon_2 \sim \mathcal{N}\left(0, \frac{1}{3}\right)$ be the variance parameters that are non-stationary, i.e. they are resampled for each trial. In summary we have constructed a data set with one discriminative and non-stationary source and nine non-discriminative sources from which one source is also non-stationary. We sample 100 trials per condition, each trial contains 200 ten-dimensional samples, and repeat the experiment 100 times. The goal is to find a spatial filter that recovers the discriminative source s_1.

Figure 3 shows the angle between the spatial filter computed by divCSP-WS (one-step method) or SSA+CSP (two-step method) and the true projection to the discriminative source s_1. One can clearly see that the two-step method removes the discriminative information in the first step, i.e. the median angle is over 50° when projecting out one or more dimensions. In other words the two-step method projects out the activity related to source s_1 simply because it is non-stationary. Thus two-step methods rely on the assumption that the discriminative subspace is stationary. If this assumption does not hold they may fail. On the other hand when applying divCSP-WS we can control the strength of regularization. That means we can trade-off stationarity and discriminativity; in real applications some amount of variation will always be present even when projecting to the sources that represent motor imagery related activity. Consequently the simultaneous optimization of stationarity and discriminativity is not only more natural but also allows to fine tune the amount of stationarity and discriminativity (soft regularization).

B. The sCSP heuristic

The following example[6] shows that the heuristic used by sCSP to construct the penalty matrix may fail, i.e. it does not capture the true non-stationarities in the data. Assume we have the following matrices

$$\mathbf{\Sigma}_1 \quad = \quad \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \; \mathbf{\Sigma}_2 \quad = \quad \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix} \tag{19}$$

$$\mathbf{\Sigma}_1^1 \quad = \quad \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \; \mathbf{\Sigma}_1^2 \quad = \quad \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix} \tag{20}$$

[5] Note that the CSP problem (maximization of a quotient) can be log-transformed into a difference maximization problem without changing the optimum. In [52] we proved that the CSP filters span a subspace with maximum symmetric KL divergence.

[6] This example comes from private communication with the authors of [19].

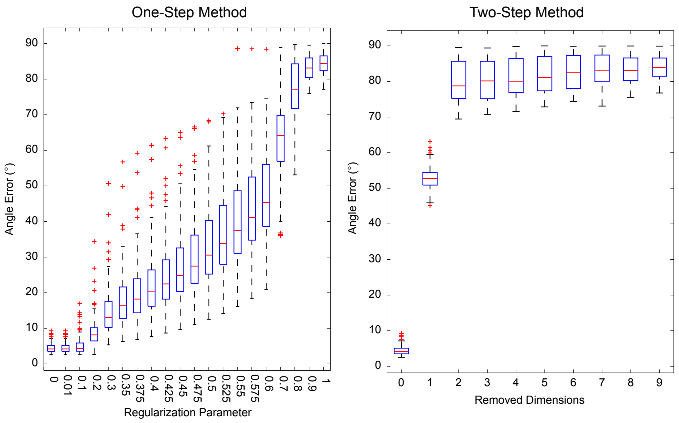

Fig. 3. Comparison of one-step approach (divCSP-WS) and two-step method (SSA+CSP). The boxplots show the distribution of the angle between the true discriminative projection and the one provided by divCSP-WS and SSA+CSP for different regularization parameters. One can clearly see that the two-step method removes discriminative information in the first step, thus will provide poor classification accuracy. Our divCSP-WS approach on the other hand performs soft regularization thus permits determining the right trade-off between stationarity and discriminativity.

where $\mathbf{\Sigma}_c$ denotes the average covariance matrices of class c and $\mathbf{\Sigma}_1^1$ and $\mathbf{\Sigma}_1^2$ stand for the covariance matrices estimated from trial 1 and 2 of class 1, respectively. Note that we only assume class 1 to be non-stationary, i.e. the trial covariance matrices of class 2 coincide with $\mathbf{\Sigma}_2$. If we aim to maximize the ratio between the variance of class 1 and 2 and simultaneously want to minimize non-stationarity then the optimal spatial filter is $\mathbf{w} = [w_1 \quad w_2]^\top = [1 \quad 0]^\top$. Considering the class differences in the off-diagonal elements of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ leads to a higher Rayleigh quotient (therefore it is preferred by CSP), but introduces variability to the extracted features. The penalty matrix $\mathbf{\Delta}$ of sCSP can be computed as

$$0.5 \cdot \mathcal{F}\left(\mathbf{\Sigma}_1^1 - \mathbf{\Sigma}_1\right) + 0.5 \cdot \mathcal{F}\left(\mathbf{\Sigma}_1^2 - \mathbf{\Sigma}_1\right) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad (21)$$

where $\mathcal{F}$ is the operator that flips the negative eigenvalues of a matrix. Since adding this matrix to the denominator of the Rayleigh quotient (as done in sCSP) will not penalize the off-diagonal elements, sCSP will not extract the filter $\mathbf{w} = [1 \quad 0]^\top$. In other words the flipping sign heuristic fails in this example. Our divergence-based approach (as well as SSA+CSP and KLCSP) penalizes the off-diagonal terms because it does not rely on heuristics but rather evaluate non-stationarity in a principled manner using KL divergence. The divCSP-WS method uses the following regularization term $\mathbf{\Delta}$

$$\sum_{i=1}^{2} D_{kl}\left(\mathbf{w}^\top \mathbf{\Sigma}_1^i \mathbf{w} \;\; || \;\; \mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}\right) \quad (22)$$

$$= \frac{1}{2}\left(\frac{\mathbf{w}^\top \mathbf{\Sigma}_1^1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}} + \log\left(\frac{\mathbf{w}^\top \mathbf{\Sigma}^1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_1^1 \mathbf{w}}\right) + \frac{\mathbf{w}^\top \mathbf{\Sigma}_1^2 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}} + \log\left(\frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_1^2 \mathbf{w}}\right)\right) - 2 \quad (23)$$

$$= \frac{1}{2}\left(2 + \log\left(\frac{0.9w_1^2 + 0.3w_1w_2 + 0.1w_2^2}{0.9w_1^2 + 0.1w_1w_2 + 0.1w_2^2}\right) + \log\left(\frac{0.9w_1^2 + 0.3w_1w_2 + 0.1w_2^2}{0.9w_1^2 + 0.5w_1w_2 + 0.1w_2^2}\right)\right) - 1. \quad (24)$$

This divCSP-WS penalty term is zero if and only if $w_1 w_2 = 0$, i.e. when disregarding the off-diagonal terms. Thus divCSP-WS finds the optimal trade-off between stationarity and discriminativity.

C. Deflation vs. Subspace algorithms

In the following let us apply the multi-subject algorithm divCSP-MS to data of five simulated subjects. As before we use a mixture model with random orthogonal mixing matrix $\mathbf{A}$ to generate the data $\mathbf{x}^j(t)$ of each subject j

$$\mathbf{x}_c^j(t) \quad \sim \quad \mathcal{N}\left(\mathbf{0}, \mathbf{A}^\top \mathbf{\Sigma}_c^j \mathbf{A}\right). \quad (25)$$

Let $\mathbf{\Sigma}_c^j = \begin{bmatrix} \mathbf{\Gamma}_c^j & 0 \\ 0 & \mathbf{\Delta}_c^j \end{bmatrix} \in \mathbb{R}^{12\times 12}$ denote the source covariance matrix of class c and subject j with $\mathbf{\Gamma}_c^j \in \mathbb{R}^{2\times 2}$ being the covariance matrix of discriminative sources common to all subjects and $\mathbf{\Delta}^j \in \mathbb{R}^{10\times 10}$ the corresponding subject specific matrix. Let the first two sources of all subjects be discriminative but have different correlations. In other words we simulate the case where the projections that reconstruct the two (independent) discriminative sources of subject i will reconstruct a linear mixture of the discriminative sources of subject j. Thus the discriminative sources of subject i and j lie in the same subspace but have different correlations. This may happen when e.g. the mixing matrix of subject i is a rotated version of the mixing matrix of subject j, e.g. because of tiny differences in electrode position or head size. For simplicity let us assume the mixing matrix is fixed for all subjects, but the correlations between the sources differ. The goal of the multi-subject algorithm is to extract discriminative activity common to all subjects, i.e. to extract the first two sources.

Let the first two sources of subject 1 be generated by a zero mean Gaussian with variance 1.5 and 0.5 for condition one and variance 0.5 and 1.5 for condition two, i.e. $\mathbf{\Gamma}_1^1 = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ and $\mathbf{\Gamma}_2^1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.5 \end{bmatrix}$. The covariance matrices $\mathbf{\Gamma}_c^j$ for the other subjects show the same structure as for subject one, but are rotated by a (random) rotation matrix with angle $\alpha \in [-\frac{\pi}{4} \;\; \frac{\pi}{4}]$. The first row of Figure 4 shows a possible data distribution of the first two sources for three subjects.

Note that the first two sources are discriminative for all subjects, thus they should be recovered by multi-subject CSP

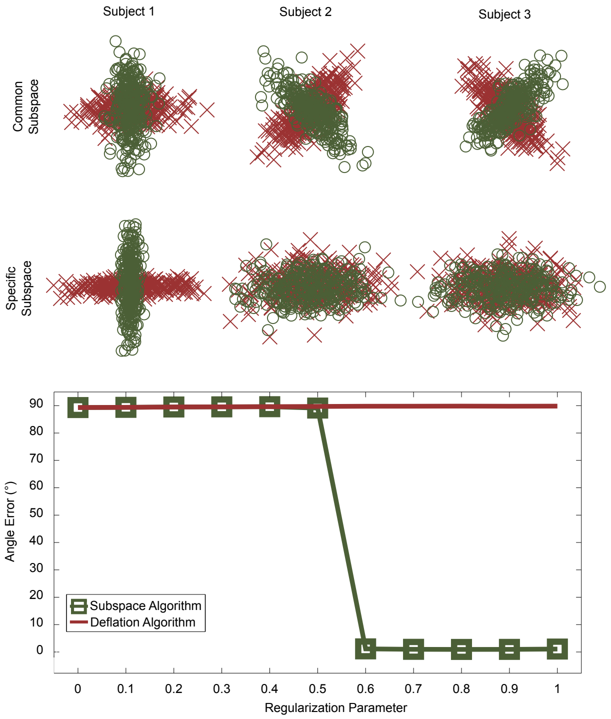

Fig. 4. Comparison of deflation and subspace version of the divCSP-MS algorithm. The first row shows a distribution that is discriminative for all three subjects when considering the whole subspace, but is not when performing the computation in deflation mode. Thus the top row represents the common discriminative activity. The changes in correlation may be due to differences in electrode montage or head size etc. The second row shows a distribution that is only discriminative for the first subject. The plot at the bottom shows that by using the subspace version of divCSP-MS we can identify the common discriminative activity whereas when optimizing one filter at a time we always prefer the specific solution.

algorithms. However, when applying divCSP-MS in deflation mode the (single) filter that separates class one and two for subject 1 may not separate the classes for the other subjects as their source activity is rotated (see first row of Figure 4). Only when extracting the whole subspace, i.e. sources one and two, the algorithm "realizes" that these subspaces are equivalently discriminative for all subjects. Thus only a subspace method helps for these kind of data integration problems. Note that the constructed example is equivalent to the well-known XOR problem [94] in feature selection literature.

Now let us assume that every subject has two other discriminative sources with variance 1.6 / 0.4 and 0.4 / 1.6 in condition one and two, respectively. However, these sources are at random positions in $\mathbf{\Delta}^j$, i.e. they are not necessarily at the same position for all subjects. The second row of Figure 4 illustrates a case where the sources are discriminative for subject 1, but the subspace is not discriminative for the other subjects.

The plot at the bottom of Figure 4 shows the results of applying divCSP-MS (100 repetitions) in deflation (red line) and in subspace (green line) mode to the data of subject 1. With increasing regularization parameter the algorithms utilize information from the four other subjects. The plot shows the median of the largest principal angles between the true filters capturing the activity of the first two sources and the filters computed by divCSP-MS. One can clearly see that for small regularization parameters λ (i.e. when only using data from subject 1) both methods do not reconstruct the activity of the common subspace (first row). This is because the subject specific activity (second row) is simply more discriminative, the subject specific sources show a variance ratio (between both classes) of 1.6 / 0.4 compared to 1.5 / 0.5. However, with increasing regularization i.e. when taking into account other subjects' data the subspace method "realizes" that there is a subspace that is discriminative for all users, thus it is preferred and the angle error decreases to 0. On the other hand the deflation method does not reconstruct the common subspace because it is not able to utilize the XOR-like structure.

D. *Effects of beta divergence*

In the following we investigate the effects of the β parameter on the type of stationarity achieved by divCSP-WS. Let us consider the two types of changes shown in Figure 5, namely gradual changes and abrupt changes. The first row of Figure 5 shows the data distributions of five epochs that change gradually. The second row of Figure 5 shows four relatively stable (stationary) distributions and one extreme change. We measure both types of variations as average divergence between the data distribution in the first epoch (reference) and the four subsequent distributions.

The bottom row of Figure 5 plots the ratio of the divergence terms computed on the examples in the second row and the first row for different β parameters with logarithmic scaling. Note that we change the scale of the x-axis at 0 (the reason for this sudden drop) as setting β to lower values than -0.0115 results in numerical problems (see derivation of beta divergence in appendix). Note that if the ratio of the divergences is above 1 then the abrupt change is regarded as more non-stationary than the gradual change; the opposite holds if the value is below 1. Thus by using beta divergence we have an additional degree of freedom, namely we can shift the focus from gradual changes that are relatively stable over the data set to strong abrupt events like electrode artifacts. Thus we can easily match the types of non-stationarities that are present in the data and compute invariant features. This flexibility can also be utilized for exploratory analysis, i.e. identification of gradual changes.

E. *KL divergence vs. Symmetric KL divergence*

In the following we want to touch upon the difference between the symmetric KL divergence and the KL divergence. The KL divergence between two zero-mean Gaussians with covariances $\mathbf{A}$ and $\mathbf{B}$ can be written in explicit form as

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbf{A} \,||\, \mathbf{B}\right) = \log\left|\mathbf{A}^{-1}\mathbf{B}\right| + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right), \tag{26}$$

whereas its symmetric counterpart is

$$\tilde{\mathrm{D}}_{\mathrm{KL}}\left(\mathbf{A} \,||\, \mathbf{B}\right) = \mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right). \tag{27}$$

From linear algebra the following relation is known

$$\log|A| = \mathrm{tr}(\log(\mathrm{A})). \tag{28}$$

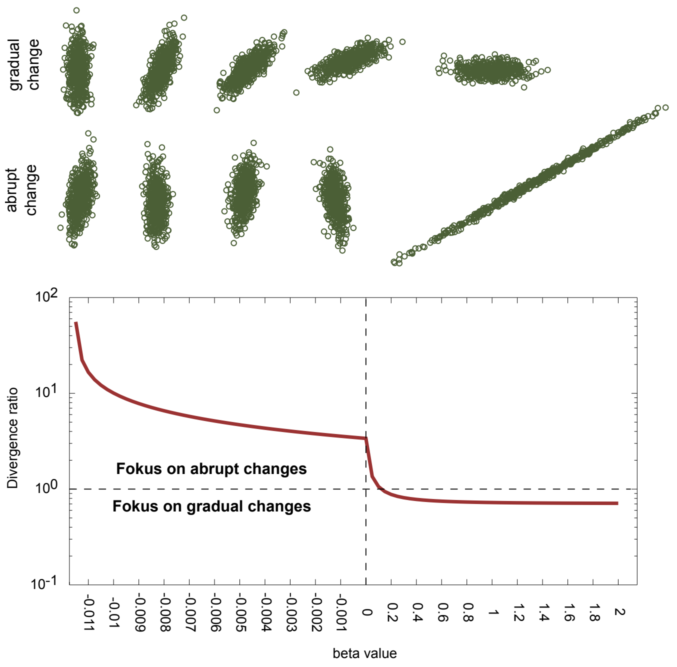

Fig. 5. The first row shows five data distributions that change gradually whereas the second row shows four distributions that are relatively stable and one that is very different, i.e. it illustrates an abrupt change in distribution. When using beta divergence we are able to differentiate between both types of non-stationarity. The bottom plot shows the ratio of the regularization terms (measured as average symmetric Kullback-Leibler divergence between the first epoch and the other ones) of the two above sets of distributions in log scale. Thus if the curve is above 1 (= 10^0) then abrupt changes are preferred, i.e. the regularization term computed for the lower set of distributions is higher than the one for the upper set, whereas if it is below 1 we give higher regularization to the gradual change. Thus by changing the beta parameter we can shift the focus from abrupt changes to gradual changes.

Using this relation we can rewrite the KL divergence objective function as

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbf{A} \mid\mid \mathbf{B}\right) = \mathrm{tr}\left(\log\left(\mathbf{A}^{-1}\mathbf{B}\right)\right) + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right) \qquad (29)$$

Thus the difference between both divergences is the log operator inside the first trace term. This log operation downweights the influence of the $\mathrm{tr}\left(\log\left(\mathbf{A}^{-1}\mathbf{B}\right)\right)$ term compared to $\mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right)$ when the eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ are very large. The question is when does such an operation make sense ?

When $\mathbf{A}$ is ill-conditioned it may have eigenvalues close to zero. In this case the term $\mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right)$ becomes very large, consequently it will dominate the solution obtained by using symmetric divergence $\tilde{\mathrm{D}}_{\mathrm{KL}}$. Using the (non-symmetric) KL divergence significantly reduces the influence of the ill-conditioned matrix $\mathbf{A}$. Thus using the log operator makes perfectly sense in the divCSP-WS algorithm as it operates on trial-wise covariance matrices that may be poorly estimated. In this case the KL divergence should be preferred. On the other hand when using average matrices as in divCSP-BS, divCSP-AS or divCSP-MS there is no reason to downweight one term of the divergence, thus the symmetric divergence should be applied.

VI. Experimental Evaluation

This section evaluates the proposed framework using real EEG recordings from 80 subjects.

A. Dataset and Experimental Setup

The data set [13] used for the evaluation comes from a joint study of TU Berlin with University Tübingen and contains EEG recordings from 80 healthy inexperienced volunteers performing motor imagery tasks with the left and right hand or with the feet. The subjects performed motor imagery first in a calibration session and then in a feedback operation in which they had to control a 1D cursor application. 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 bandpass from 0.05 to 200 Hz. Three runs with 25 trials of each motor condition were recorded in the calibration session, then the two best classes were selected and the subjects performed feedback with three runs of 100 trials. Both sessions were recorded on the same day.

For the offline analysis we manually select 62 electrodes densely covering the motor cortex, filter the data in 8-30 Hz with a 5-order Butterworth filter and extract a time segment from 750ms to 3500ms after the trial start. We do not apply any manual or automatic rejection of trials or electrodes and use six spatial filters for feature extraction. As classifier we apply Linear Discriminant Analysis (LDA) after computing the logarithm of the variance on the spatially filtered data. We measure performance as rate of misclassification and normalize the covariance matrices by dividing them by their traces. The parameter λ is selected from the set of 11 candidates $\{0, 0.1, 0.2 \ldots 1\}$ by 5-fold cross-validation on the calibration data using minimal error rate as selection criterion. We compare the different divCSP approaches to several state-of-the-art algorithms and also select their regularization parameters by cross-validation. The following methods are used for comparison:

- SSA+CSP [57] is a two-step method that projects the data to a stationary subspace prior to CSP computation. The regularization parameter is the number of removed directions in the first step. We select it from $0 - 22$.
- The covCSP [33] method regularizes the class-covariance matrices towards the average covariance matrix of other subjects. The klcovCSP [35] approach applies the same idea but weights the contributions of other subjects by similarity (measured as inverse KL divergence). We use the regularization parameters $\{0, 10^{-5}, \ldots, 10^{-1}, 0.2, \ldots, 0.9, 1\}$ for both methods. Both methods rely completely on other subjects' data if the parameter is 1 and they correspond to standard CSP if it is 0.
- The ssCSP method [31] extracts the directions of common change between training and test data from other subjects' data. For that it extracts the $l = 1 \ldots 10$ most non-stationary directions for each subject and constructs a $\nu = 0 \ldots 10$ dimensional subspace from them. These directions of common change are then penalized in the spatial filter computation.

- The KLCSP method [19] optimizes the same objective as deflation divCSP-WS with $\beta = 0$. Thus we use our implementation with $\lambda = \{0, 0.1, 0.2 \ldots 1\}$ to compute the KLCSP spatial filters.

B. Reducing Within-Session Non-Stationarity

In the first experiment we aim to increase the stationarity of the training features by applying divCSP-WS. In order to capture different kinds of variations, both single extreme events and common slow changes, we test our algorithm with different beta parameters. We use $\beta = 0, 0.5, 1$ and the minimal possible[7] negative value from $-0.0005, -0.0010, -0.0015, \ldots$. We select the best of these four β values for each subject by applying cross validation. Figure 6 shows the error rates of all subjects for the subspace and deflation divCSP-WS and compares them to standard CSP (first row), SSA+CSP (second row) and KLCSP (last row). Note that we did not reimplement the original KLCSP algorithm, but use the deflation divCSP-WS algorithm with $\beta = 0$ as both algorithms optimize the same objective. Each circle in the scatter plot represents the error rate of one subject and the red number in the lower right corner denotes the p-value when applying the one-sided Wilcoxon sign-rank test. The error rate of our approach is represented on the y-axis i.e. if the circle is below the solid line then our method outperforms the baseline for this subject. The null hypothesis of the Wilcoxon test is that the median of the error rate differences (our method (y-axis) - baseline method (x-axis)) is greater or equal to zero. For $p < 0.05$ we reject this null hypothesis, thus we say that our method significantly outperform the baseline.

One can see from the plot that the deflation divCSP-WS outperforms the subspace method. It significantly decreases classification error rates compared to CSP (p=0.0481); the subspace approach does not show any improvement. The subspace method performs poorly as it considers changes in correlations between different spatial filters. These correlations are ignored in the feature extraction and classification process, thus should not be considered when computing the spatial filters. One can also see from the plots that the simultaneous optimization of two objectives, discriminativity and stationarity in this case, is superior to the sequential optimization as done by the two-step SSA+CSP approach. The improvement of the deflation divCSP-WS algorithm is very close to being significant (p=0.0526). This observation is in line with the simulations performed in the last section and with previous work [80]. The fact that two-step methods may remove information in the first step that is important for the second step is a significant disadvantage of these approaches. We will comment on this in the next paragraph. The scatter plots at the bottom show the advantage of using the beta divergence version of our algorithms. The results show that the Kullback-Leibler divergence algorithm (as in [19]) performs worse than our deflation divCSP-WS method and the difference between both algorithms is close to being significant (p=0.0750). The improvement is due to the additional flexibility of beta divergence; it can capture a whole range of different non-stationarities. On the other hand one can see that KLCSP significantly outperforms the subspace divCSP-WS method (p=0.9814). Thus the additional flexibility of using different beta values does not compensate for the disadvantage of considering non-stationarities in correlation.

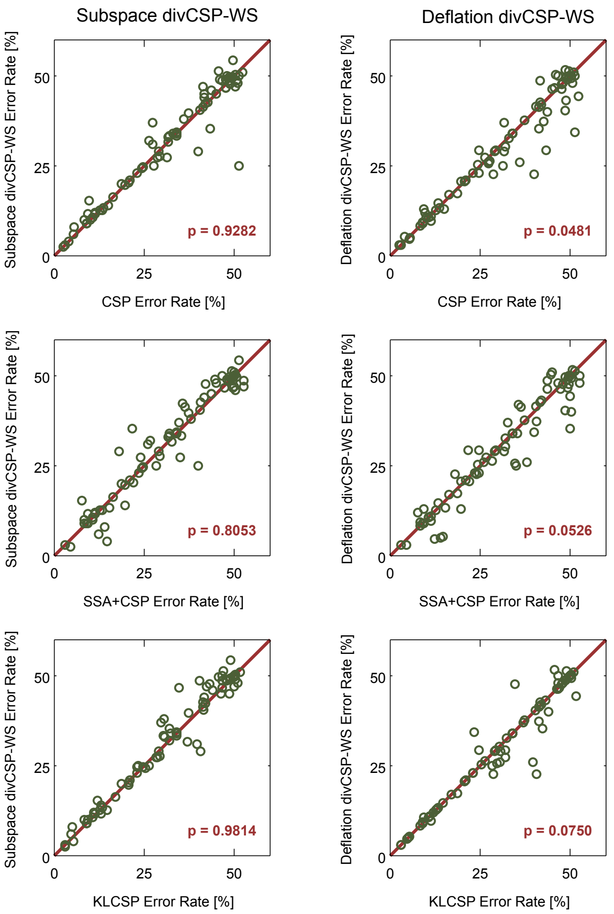

Fig. 6. Scatter plots showing error rates of divCSP-WS and three baseline methods. The left column shows the error rates of subspace divCSP-WS, the right one of the deflation algorithm. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is shown in the right bottom corner.

Above we discussed situations where two-step methods may provide suboptimal performance. Figure 7 shows the boxplot of the difference in error rate between SSA+CSP [57] and CSP. In SSA+CSP we remove $0 - 22$ dimensions from the data prior to CSP computation. One can see from the figure that the classification performance of SSA+CSP drops with increasing number of removed dimensions. This means that the directions removed in the first step of SSA+CSP contain increasing amount of discriminative information (that is required for the second step). The two scalp plots visualize the activity patterns corresponding to the removed directions for two subjects. One can clearly see that the upper scalp plot shows activity over the left motor and temporal cortex. Since such activity contains motor imagery related information (right hand class) it is not advisable to remove it. Since SSA only evaluates the amount of non-stationarity and does not take into account the information content it removes this activity,

[7] See derivation in appendix.

thus the corresponding subject shows a significant increase in error rate, namely from 9.3 % to 18.3 %. The lower scalp plot shows a subject that improves classification accuracy (from 40 % to 18 %) by applying SSA preprocessing. One can see that some temporal activity is removed from this subjects data. Since this information is not motor imagery related it can be safely removed. These examples shows that two-step methods may fail in practice. Although the authors of [56] propose to trade-off non-stationarity and discriminativity when using SSA, we emphasize the limits of applying two step approaches for feature extraction in BCI.

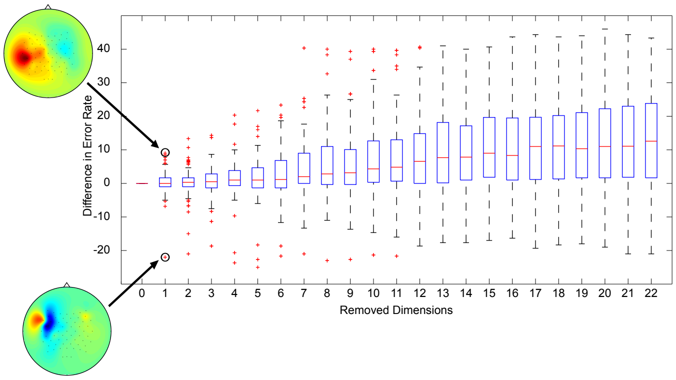

Fig. 7. The boxplot shows the distribution of the differences in error rate when removing $0 \ldots 22$ dimensions from the data by applying SSA. One can see that the error rates significantly increase as more dimensions are removed. The two scalp plots show the activity patterns corresponding to the removed direction. One can clearly see that for the subject with increasing error rate the (upper) scalp plot shows activity related to motor imagery. Since this information should not be removed, SSA+CSP increases the error rate for this subject.

The effects of different beta parameters can be studied in Figure 8. The upper panel shows a subject's EEG signal with a strong artifact in electrode FFC6. The three scalp plots at the bottom panel show the activity patterns of the first spatial filter extracted by divCSP-WS with $\lambda = 0.5$ and different beta values. One can clearly see that if $\beta = 0$ (left scalp plot) the regularization of divCSP-WS has no effect on the solution. The pattern focuses on the activity in FFC6 (due to the strong artifact) and does not capture motor imagery related information. Thus the error rate is over 40 %. If using a beta value of 1 (right scalp plot) there is an improvement i.e. a right hand motor imagery pattern emerges, however, the focus on the electrode FFC6 is still present. This is because larger β values downweight the influence of the artifactual trial in the penalty term $\boldsymbol{\Delta}$, thus the regularization does not penalize strong extreme events like the artifact in FFC6. The situation changes if $\beta < 0$ as then we enhance the extreme values in the penalty term $\boldsymbol{\Delta}$ computation i.e. the artifact dominates the penalty term thus is much more strongly penalized in the optimization process. The effect of this penalty is that a true motor imagery related pattern emerges and the focus on electrode FFC6 disappears. The error rate of this pattern also largely decreases to 13.8%. We showed a similar effect in the toy simulations in last section. This additional degree of freedom makes our method(s) much more flexible than e.g. KLCSP [19].

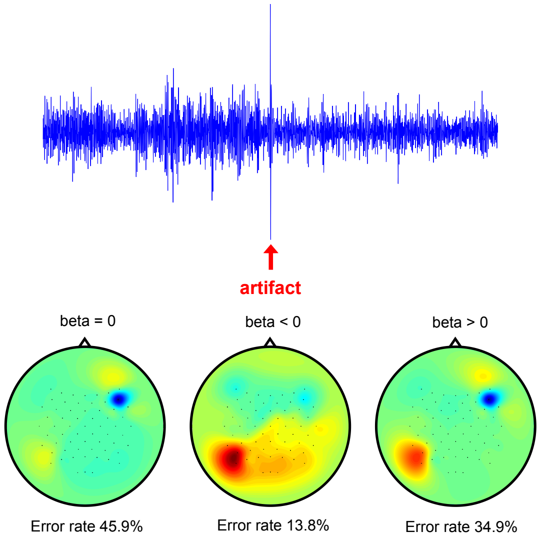

Fig. 8. Effects of different beta values on the artifact penalization of subject 74. The upper plot shows an artifact in the signal of the FFC6 electrode. The lower panel shows the activity patterns computed by divCSP-WS with $\lambda = 0.5$. One can see that the regularization (minimizing the effect of FFC6 on the solution) only works properly if $\beta < 0$ as this enhances the artifactual activity and thus increases its relative penalty.

C. Reducing Between-Session Shifts

In this subsection we describe several between-session experiments using divCSP-BS. As before we apply the subspace and deflation algorithm and use the beta values 0, 0.5 and 1. We compare the results to standard CSP, to the recently proposed stationary subspace CSP (ssCSP) method [31] and to divCSP-BS with $\beta = 0$. Note that we only integrate information from other subjects with the same motor imagery classes and select the regularization parameters by minimizing test error on the other subjects' data. The first row of Figure 9 shows a performance improvement of the deflation divCSP-BS method over CSP. Although there is a trend the difference is not statistically significant (p=0.0938). This confirms the observation of [31] that information about shifts between sessions can be transferred across subjects. In contrast to the within-session non-stationarities presented in the last subsection we are not so much interested in single extreme non-stationarities between training and test sessions, but rather in changes that are stable over subjects. By using $\beta > 0$ we penalize these (common) changes between calibration and feedback. However, it seems that using different beta parameters does not have a large impact on the results (see bottom row). The second row of Figure 9 compares divCSP-BS to ssCSP. Although we compute the shift between calibration and feedback session for each class separately, our method does not outperform ssCSP which does not use class labels (p=0.5377). This suggests that the non-stationarities between calibration and feedback session are not class-dependent.

The upper plot in Figure 10 shows the median (over subjects) KL divergence differences between CSP (no regularization) and deflation divCSP-BS with increasing regularization. The divergence is computed between the calibration and feedback feature distribution when applying the filters

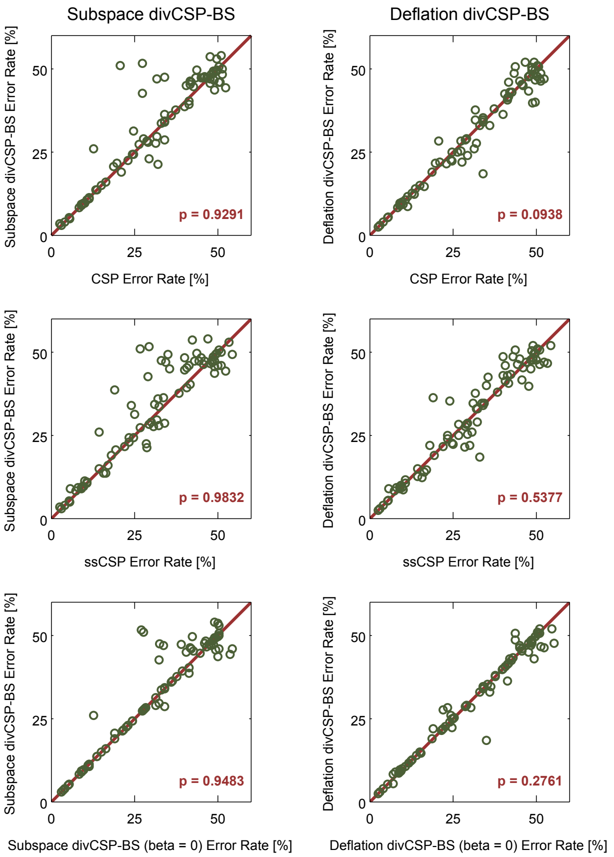

Fig. 9. Scatter plots showing error rates of divCSP-BS and three baseline methods. The left column shows the error rates of subspace divCSP-BS, the right one of the deflation algorithm. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is shown in the right bottom corner.

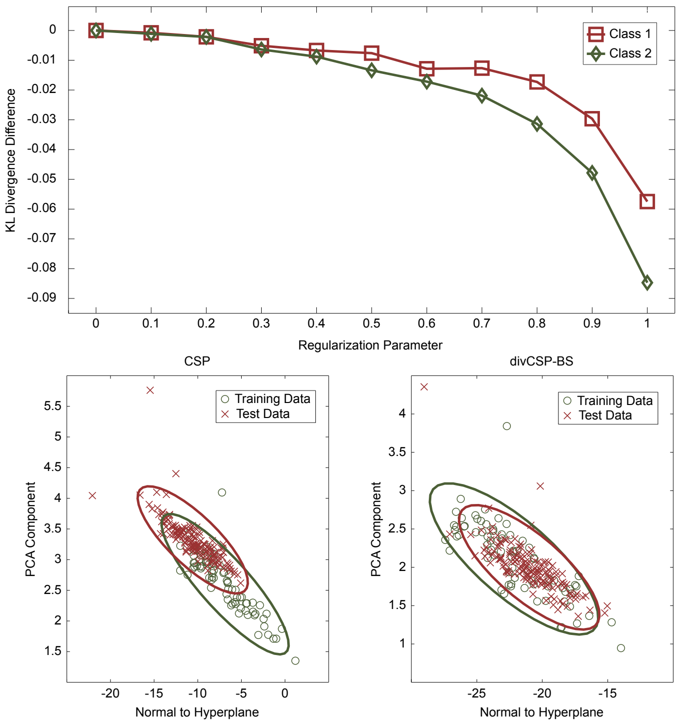

Fig. 10. Effect of regularizing the solution towards between-session stationarity. The upper plot shows the median KL divergence difference between CSP and deflation divCSP-BS with increasing regularization. The divergence is computed between the calibration and feedback feature distribution. One can see that the divergence decreases with increasing regularization. This confirms that the non-stationarities are similar between different users. The lower plots show the 'left hand' feature distribution of training data (green circles) and test data (red crosses) when applying CSP and divCSP-BS. The features are projected to the largest PCA component and the normal vector to the classification hyperplane. One can clearly see that the divCSP-BS solution provides much more stationary feature distributions than CSP.

computed by divCSP-BS with increasing λ. One can see from the plot that incorporating information from other users about the shift between calibration and feedback constantly reduces this shift on the subject of interest. This confirms our observation from [31] that non-stationarities are similar between different subjects. The lower panel of the figure shows the effect of applying divCSP-BS. It shows the feature distribution of the 'left hand class' train data (green circles) and the corresponding test data (red crosses) of subject 13. The six dimensional feature distribution is projected to two dimensions by using the largest PCA component and the normal vector to the classification hyperplane. One can see that when applying CSP there is a large shift in the distribution between training and test. If on the other hand incorporating information form other subjects one obtains a stationary distribution with no significant shift between training and test.

D. Stationarity Across Subjects

This subsection discusses the results of divCSP-AS; as before we use the beta values 0, 0.5 and 1. Figure 11 shows the results of both the subspace and deflation algorithm and compares them to standard CSP, covCSP and klcovCSP. One can see (first row) that divCSP-AS significantly outperforms CSP ($p < 10^{-4}$), i.e. regularizing the feature distribution towards the feature distribution of the other subjects seems to have a strong effect on the quality of the spatial filters. This regularization effect is stronger than when regularizing the covariance matrices towards other subjects as done by covCSP (p=0.0626) and klcovCSP (p=0.1120).

Figure 12 evaluates the improvement of subject 74, the user with largest decrease in error rate. The lower boxplot shows the distribution of the KL divergence between subject 74 and the other subjects when applying the first spatial filter computed by divCSP-AS with increasing regularization parameters. One can see that there is a large gap when moving from $\lambda = 0.2$ to $\lambda = 0.3$, i.e. the feature distribution of subject 74 becomes similar to the distribution of other subjects. Above the boxplot we visualize the activation patterns of the first spatial filter computed with divCSP-AS. One can clearly see the electrode artifact in FFC6 (see also Figure 8). Since this activity is not present in the other subjects it is penalized when applying divCSP-AS. For regularization parameters larger than $\lambda = 0.3$ it completely vanishes. In other words regularizing the feature distribution towards other subjects helps to remove this kinds of anomalies. Note also that for $\lambda > 0.5$ the activity patterns show strong activation in motor imagery related areas. This activation is captured by divCSP-AS as it is present in all

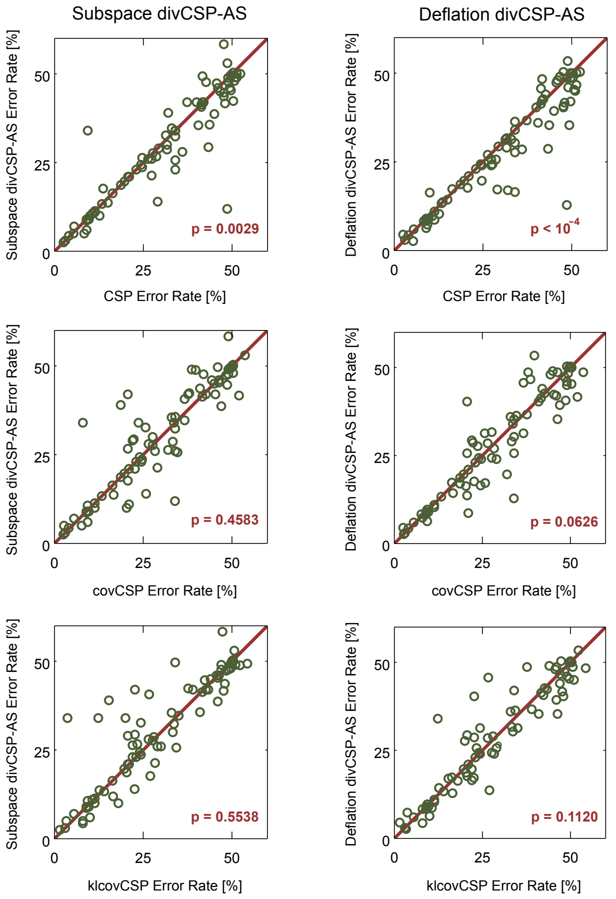

Fig. 11. Scatter plots showing error rates of divCSP-AS and three baseline methods. The left column shows the error rates of subspace divCSP-AS, the right one of the deflation algorithm. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is shown in the right bottom corner.

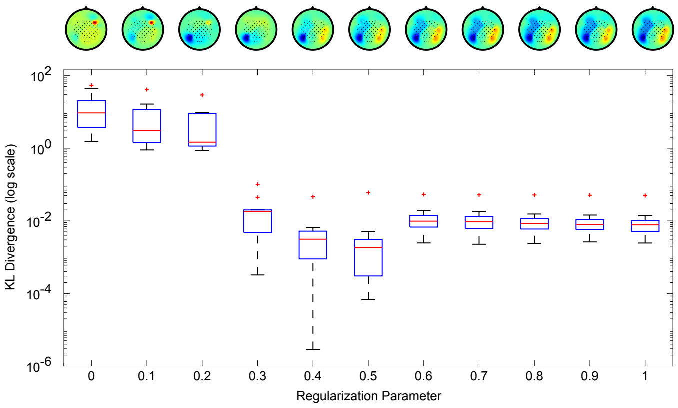

Fig. 12. Effects of applying divCSP-AS. The boxplot shows the similarity of the feature distribution of subject 74 and the other subjects when projected by the first spatial filter computed with divCSP-AS. The similarity is measured as KL divergence. The activity patterns above the boxplot show that the influence of the artifact in electrode FFC6 decreases with increasing regularization.

subjects (having the same classes as subject 74).

E. Subject-Independent Spatial Filters

In the last subsection we perform regularization towards other subjects, here we aim to use other subjects' data to extract a subject independent feature space. Therefore we apply divCSP-MS, covCSP and klcovCSP with $\lambda = 1$ (i.e. the case only with the regularization term). In other words we estimate the spatial filters by using other subjects' data only. Note that we still use the calibration data to train the LDA classifier, only the spatial filters are "subject independent". As before we apply our algorithm with the three beta parameters 0, 0.5 and 1. Figure 13 compares the error rates of the subspace and deflation divCSP-MS algorithm with three baselines. One can clearly see that both the subspace and deflation divCSP-MS provide better feature spaces than covCSP and klcovCSP. The improvement is statistically significant for the subspace algorithm with p=0.0004 when comparing its performance to covCSP and p=0.0147 for klcovCSP. The subspace method performs significantly better than covCSP (p=0.0193), the improvement over klcovCSP is not significant (p=0.2105). This means that integrating information from other subjects by combining divergence terms that measure motor imagery related activity is superior to combining the covariance matrices i.e. fusing all information. As observed in the above simulations the subspace method is (slightly) better than the deflation approach. The subspace method is not affected by changes in correlation, thus it identifies the common subspace even when differences in correlation of the sources exist between subjects. We can also see (bottom row) that using beta divergence significantly improves the algorithm, the p-value for the subspace approach is 0.0101, for the deflation method it is smaller than 10^{-4}. It seems that beta values larger than zero have a positive effect on performance as they downweight the influence of individual subjects and help to extract common motor imagery related activity.

A direct comparison of the subspace and deflation method for the three beta values is shown in the upper panel of Figure 14. For the case of beta = 0 one can see a clear advantage of the subspace method (p=0.0001). As shown in the simulations (Figure 4) the deflation approach may prefer single-subject solutions as it does not capture common activity if the correlations of the sources vary between subjects. However, the relative gain of these single-subject solutions decreases with increasing beta value (because of downweighting effect), therefore the subspace and deflation algorithms perform similarly for beta = 1. The lower panel of Figure 14 compares the subject independent feature spaces computed by divCSP-MS (after selecting β by cross-validation) to the CSP solution when computed on increasing number of trials. For that we randomly select $n = 2 \ldots 75$ trials per class from the calibration data and compute CSP on this smaller data set. Afterwards we train the LDA classifier on the whole calibration data and apply it to the feedback data. We repeat this 50 times. In the left boxplot one can clearly see that the subject-independent spatial filters computed with the subspace method (green line) perform as well as the filters computed by CSP, even when using all 150 trials for the covariance estimation. The deflation divCSP-MS method shows a similar performance, although it has much higher variance and its 25%

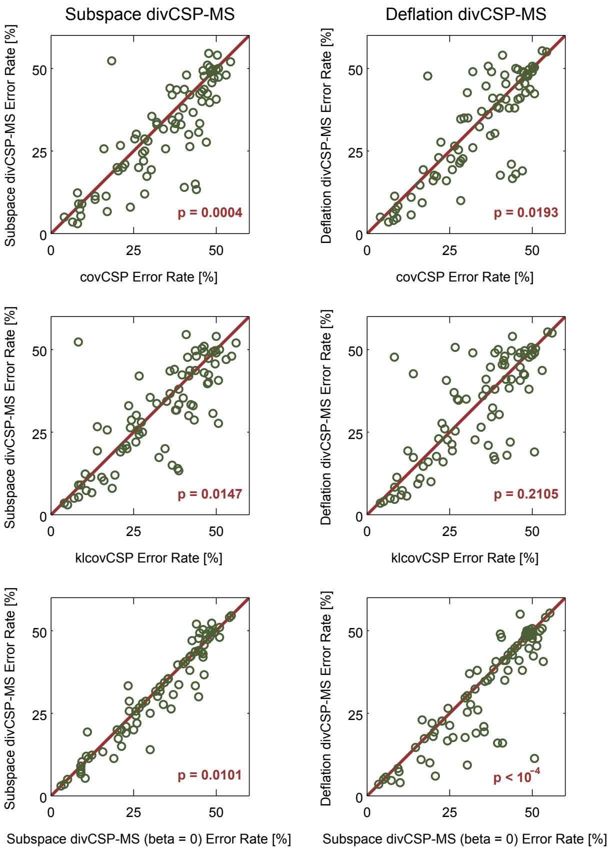

Fig. 13. Scatter plots showing error rates of divCSP-MS and three baseline methods (for $\lambda = 1$). The left column shows the error rates of subspace divCSP-MS, the right one of the deflation algorithm. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is shown in the right bottom corner.

quantile is significantly lower than in the case of the subspace approach. Note that the performance of deflation divCSP-MS is significantly worse than CSP (computed on all trials) when using $\beta = 0$, whereas the performance of the subspace method is (almost) on par with CSP in such a setting. Thus for subject-independent spatial filters we strongly recommend to use the subspace method and the beta divergence algorithm.

VII. Discussion

Common spatial patterns and its variants have established themselves as a de facto standard in EEG analysis in particular for BCI. Since to date numerous papers have been presented, it has become more and more difficult for the user and practitioner to chose between the appropriate algorithm variants. In this work we presented a common divergence-based framework for the CSP family that unifies CSP variants in a principled manner. It can encompass different types of robustness properties, regularization, invariances and also allows to integrate variants of non-stationarity. Moreover we could show that our novel framework can also help to transfer information from one subject to another, and in that manner yield subject independent decoders. The heart of our versatile framework is a reformulation of CSP as a divergence maximization problem. Here we suggested two possible directions to solve the optimization problem: a deflation variant, where the most salient features are found first and a subspace formulation, where the full subspace is extracted. Interestingly, for some problems the deflation optimization is more advantageous, sic robust, over using the subspace approach; for others it is the other way round. Intuitive examples show the limits of existing methods and the power of the novel framework. Finally we provided an extensive numerical comparison study over 80 subjects, allowing for physiological interpretation. With our work we furthermore have hoped to contribute to a common platform that allows a simple and straight forward comparison of new algorithms to the large CSP family encompassed by the novel framework.

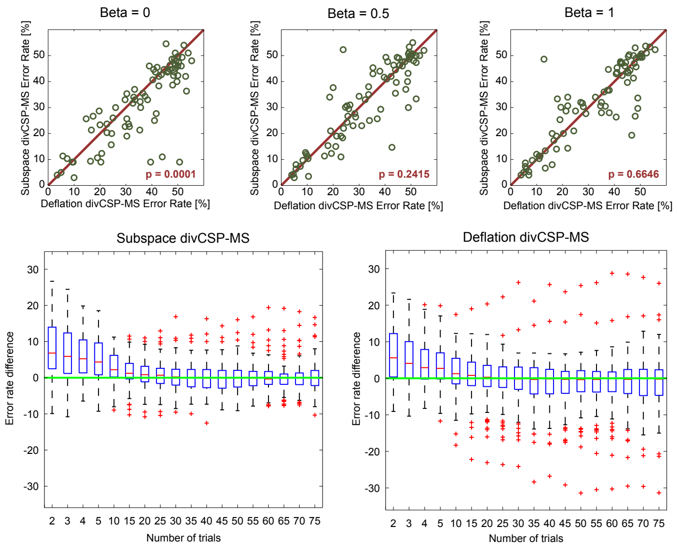

Fig. 14. Comparison of deflation and subspace version of the divCSP-MS algorithm with $\lambda = 1$, i.e. when learning spatial filters on other subjects. The upper plots compare the error rates of the deflation method (x-axis) and the subspace approach (y-axis) for different beta values. For the case of beta = 0 one can see that most of the circles representing the error rate of a particular subject are below the solid line, i.e. the subspace method perform better for these subjects. The relative advantage of the subspace method decreases constantly with increasing beta value. The lower boxplots show the distribution of error rate differences for both the subspace and deflation divCSP-MS approach and CSP computed with different numbers of trials. Both divCSP-MS methods provide significantly better results than CSP when trained on less than 15 trials per class. The CSP performance is poor in this case as the high-dimensional covariance matrices can not be reliably estimated on 15 trials. Although divCSP-MS computes spatial filters by using other subjects' data only, its performance is on par with CSP that use all 75 trials for covariance estimation.

It will be interesting to investigate other divergence measures, especially the class of Bregmann divergences which may further robustify our algorithm. Furthermore we would like to use our information geometric framework for classification purposes directly on the manifold of covariance matrices as done in [82], in the context of kernel machines [95] and apply our methods to multimodal data [96]. Future research will apply our framework in the context of online BCI experiments and generic biomedical data.

References

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

[2] J. Wolpaw and E. W. Wolpaw, Eds., *Brain-Computer Interfaces: Principles and Practice*. Oxford Univ. Press, 2012.

[3] 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*, vol. 25, no. 1, pp. 41–56, 2008.

[4] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 4, pp. 441–446, 1998.

[5] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Recipes for the linear analysis of eeg," *NeuroImage*, vol. 28, pp. 326–341, 2005.

[6] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, 2011.

[7] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355 –362, 2011.

[8] P. Sajda, A. Gerson, K.-R. Müller, B. Blankertz, and L. Parra, "A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 184–185, 2003.

[9] B. Blankertz, K.-R. Müller, G. Curio, T. Vaughan, G. Schalk, J. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The bci competition 2003: progress and perspectives in detection and discrimination of eeg single trials," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1044–1051, 2004.

[10] B. Blankertz, K.-R. Müller, D. Krusienski, G. Schalk, J. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer, "The bci competition iii: validating alternative approaches to actual bci problems," *IEEE Trans. on Neural Syst. and Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, 2006.

[11] 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*, vol. 6, no. 55, 2012.

[12] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for eeg-based communication," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 3, pp. 386 – 394, 1997.

[13] 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*, vol. 51, no. 4, pp. 1303–1309, 2010.

[14] T. P. Jung, S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.

[15] N. Ille, P. Berg, and M. Scherg, "Artifact correction of the ongoing eeg using spatial filters based on artifact and brain signal topographies," *Journal of clinical neurophysiology*, vol. 19, no. 2, pp. 113–124, 2002.

[16] C. Sannelli, M. Braun, and K.-R. Müller, "Improving bci performance by task-related trial pruning," *Neural Networks*, vol. 22, no. 9, pp. 1295–1304, 2009.

[17] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K.-R. Müller, "Towards adaptive classification for BCI," *Journal of neural engineering*, vol. 3, no. 1, pp. R13–R23, 2006.

[18] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[19] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 610–619, 2013.

[20] S. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.

[21] H. Berger, "Über das Elektrenkephalogramm des Menschen II," *Journal für Psychologie und Neurologie*, vol. 40, pp. 160–179, 1930.

[22] G. Pfurtscheller and F. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles." *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[23] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, 5th ed. Lippincott Williams & Wilkins, 2004.

[24] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[25] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. G. de Peralta, "Eeg source imaging," *Clinical Neurophysiology*, vol. 115, no. 10, pp. 2195 – 2222, 2004.

[26] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularized common spatial patterns with generic learning for eeg signal classification," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, 2009, pp. 6599–6602.

[27] M. Kawanabe and C. Vidaurre, "Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices," in *Proc. of IWANN 09, Part I, LNCS*, 2009, pp. 279–282.

[28] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre, "Robust common spatial filters with a maxmin approach," *Neural Computation*, vol. 26, no. 2, pp. 1–28, 2014.

[29] M. Krauledat, "Analysis of nonstationarities in eeg signals for improving brain-computer interface performance," Ph.D. dissertation, Technische Universität Berlin, 2008.

[30] A. Bamdadian, C. Guan, K. K. Ang, and J. Xu, "Online semi-supervised learning with kl distance weighting for motor imagery-based bci," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, 2012, pp. 2732–2735.

[31] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain-computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[32] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PloS one*, vol. 3, no. 8, p. e2967, 2008.

[33] F. Lotte and C. Guan, "Learning from other subjects helps reducing Brain-Computer interface calibration time," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 614–617.

[34] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multi-subject learning for common spatial patterns in motor-imagery bci," *Computational Intelligence and Neuroscience*, vol. 2011, no. 217987, pp. 1–9, 2011.

[35] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *Signal Processing Letters*, vol. 16, no. 8, pp. 683 –686, 2009.

[36] 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*, vol. 52, pp. 1541–1548, 2005.

[37] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Machine-learning-based coadaptive calibration for brain-computer interfaces," *Neural Computation*, vol. 23, no. 3, pp. 791–816, 2011.

[38] X. Yong, R. Ward, and G. Birch, "Robust common spatial patterns for eeg signal preprocessing," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, 2008, pp. 2087–2090.

[39] H. Lu, H.-L. Eng, C. Guan, K. Plataniotis, and A. Venetsanopoulos, "Regularized common spatial pattern with aggregation for eeg classification in small-sample setting," *IEEE Trans. on Biomed. Eng.*, vol. 57, no. 12, pp. 2936–2946, 2010.

[40] T. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support vector channel selection in bci," *IEEE Trans. on Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, 2004.

[41] M. Arvaneh, C. Guan, K. K. Ang, and H.-C. Quek, "Spatially sparsed common spatial pattern to improve bci performance," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 2412–2415.

[42] F. Goksu, N. Ince, and A. Tewfik, "Sparse common spatial patterns in brain computer interface applications," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 533–536.

[43] M. Grosse-Wentrup, K. Gramann, and M. Buss, "Adaptive spatial filters with predefined region of interest for EEG based brain-computer-interface," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 537–544.

[44] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1209 –1219, 2009.

[45] 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*, vol. 8, no. 2, p. 025012, 2011.

[46] M. Kawanabe, C. Vidaurre, B. Blankertz, and K.-R. Müller, "A maxmin approach to optimize spatial filters for eeg single-trial classification," in *Bio-Inspired Systems: Computational and Ambient Intelligence*, ser. LNCS. Springer, 2009, vol. 5517, pp. 674–682.

[47] F. Lotte and C. Guan, "Spatially regularized common spatial patterns for eeg classification," in *Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 3712–3715.

[48] 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*, 2008, pp. 113–120.

[49] J. Park and W. Chung, "Common spatial patterns based on generalized norms," in *Int. Winter Workshop on Brain-Computer Interface*, 2013, pp. 39–42.

[50] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. on Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, 2012.

[51] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A probabilistic framework for learning robust common spatial patterns." *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, vol. 2009, pp. 4658–61, 2009.

[52] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013, pp. 1007–1015.

[53] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

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

[55] P. von Bünau, F. 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)*, 2010, pp. 2810–2813.

[56] W. Samek, K.-R. Müller, M. Kawanabe, and C. Vidaurre, "Brain-computer interfacing in discriminative and stationary subspaces," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, 2012.

[57] W. Samek, M. Kawanabe, and C. Vidaurre, "Group-wise stationary subspace analysis - a novel method for studying non-stationarities," in *Proc. of Int. Brain-Computer Interface Conference*, 2011, pp. 16–20.

[58] B. Reuderink, "Robust brain-computer interfaces," Ph.D. dissertation, University of Twente, 2011.

[59] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Eeg data space adaptation to reduce intersession nonstationarity in brain-computer interface." *Neural Computation*, vol. 25, no. 8, pp. 2146–2171, 2013.

[60] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

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

[62] H. Kang and S. Choi, "Bayesian multi-task learning for common spatial patterns," in *Int. Workshop on Pattern Recognition in NeuroImaging (PRNI)*, 2011, pp. 61–64.

[63] W. Samek, A. Binder, and K.-R. Müller, "Multiple kernel learning for brain-computer interfacing," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, 2013.

[64] H. Wang and W. Zheng, "Local temporal common spatial patterns for robust single-trial eeg classification," *IEEE Trans. on Neural Systems and Rehab. Eng.*, vol. 16, no. 2, pp. 131–139, 2008.

[65] H. Wang, "Discriminant and adaptive extensions to local temporal common spatial patterns," *Pattern Recognition Letters*, vol. 34, no. 10, pp. 1125–1129, 2013.

[66] H. Wang and D. Xu, "Comprehensive common spatial patterns with temporal structure information of eeg data: Minimizing nontask related eeg component," *IEEE Trans. on Biomed. Eng.*, vol. 59, no. 9, pp. 2496–2505, 2012.

[67] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. on Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, 2008.

[68] C. Gouy-Pailler, M. Congedo, C. Brunner, C. Jutten, and G. Pfurtscheller, "Nonstationary brain source separation for multiclass motor imagery," *IEEE Trans. on Biomed. Eng.*, vol. 57, no. 2, pp. 469–478, 2010.

[69] T.-H. Nguyen, S.-M. Park, K.-E. Ko, and K.-B. Sim, "Multi-class stationary csp for optimal feature separation of brain source in bci system," in *Int. Conf. on Control, Automation and Systems (ICCAS)*, 2012, pp. 1035–1039.

[70] H. Wang, "Harmonic mean of kullback-leibler divergences for optimizing multi-class eeg spatio-temporal filters," *Neural Processing Letters*, vol. 36, no. 2, pp. 161–171, 2012.

[71] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an eeg-based brain-computer interface," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1049–1060, 2013.

[72] E. A. Mousavi, J. J. Maller, P. B. Fitzgerald, and B. J. Lithgow, "Wavelet common spatial pattern in asynchronous offline brain computer interfaces," *Biomedical Signal Processing and Control*, vol. 6, no. 2, pp. 121 – 128, 2011.

[73] O. Falzon, K. P. Camilleri, and J. Muscat, "The analytic common spatial patterns method for eeg-based bci data," *Journal of Neural Engineering*, vol. 9, no. 4, p. 045009, 2012.

[74] D. Fattahi, B. Nasihatkon, and R. Boostani, "A general framework to estimate spatial and spatio-spectral filters for eeg signal classification," *Neurocomputing*, vol. 119, no. 0, pp. 165 – 174, 2013.

[75] X. Li, H. Zhang, C. Guan, S. H. Ong, K. K. Ang, and Y. Pan, "Discriminative learning of propagation and spatial pattern for motor imagery eeg analysis," *Neural Comput.*, vol. 25, no. 10, pp. 2709–2733, 2013.

[76] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010*, 2010, pp. 17–24.

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

[78] 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.*, vol. 58, no. 3, pp. 587 –597, 2011.

[79] Y. Li and C. Guan, "An extended em algorithm for joint feature extraction and classification in brain-computer interfaces," *Neural Computation*, vol. 18, no. 11, pp. 2730–2761, 2006.

[80] R. Tomioka and K.-R. Müller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface," *NeuroImage*, vol. 49, no. 1, pp. 415–432, 2009.

[81] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Common spatial pattern revisited by riemannian geometry," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010, pp. 472–476.

[82] ——, "Multiclass brain-computer interface classification by riemannian geometry," *IEEE Trans. on Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, 2012.

[83] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[84] M. Kawanabe, W. Samek, P. von Bünau, and F. Meinecke, "An information geometrical view of stationary subspace analysis," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, ser. LNCS. Springer, 2011, vol. 6792, pp. 397–404.

[85] S. Amari, H. Nagaoka, and D. Harada, *Methods of information geometry*. American Mathematical Society, 2000.

[86] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller, "Finding Stationary Subspaces in Multivariate Time Series," *Physical Review Letters*, vol. 103, no. 21, pp. 214 101+, 2009.

[87] P. von Bünau, "Stationary subspace analysis - towards understanding non-stationary data," Ph.D. dissertation, Technische Universität Berlin, 2012.

[88] M. D. Plumbley, "Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras," *Neurocomputing*, vol. 67, no. 161-197, 2005.

[89] A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[90] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," *Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep*, 2001.

[91] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[92] N. Murata, T. Takenouchi, and T. Kanamori, "Information geometry of u-boost and bregman divergence," *Neural Computation*, vol. 16, pp. 1437–1481, 2004.

[93] A. Cichocki and S. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[94] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[95] G. Montavon, M. Braun, T. Krüger, and K.-R. Müller, "Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 62–74, 2013.

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

[97] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.

[98] K. B. Petersen and M. S. Pedersen, "The Matrix Cookbook," 2008. [Online]. Available: http://matrixcookbook.com/

[99] R. Bhatia, *Matrix analysis*, ser. Graduate Texts in Mathematics. Springer, 1997, vol. 169.

Appendix

The objective function of divCSP $\mathcal{L}(\mathbf{R})$ is a sum of (symmetric) divergences between d-dimensional Gaussian distributions

$$\sum D\left((\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\ ||\ (\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right).$$

We will show that it can be expressed in explicit form when using KL divergence and beta divergence. Let us denote the whitened covariance matrices as $\tilde{\mathbf{\Sigma}} = \mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top$ and the projected ones as $\bar{\mathbf{\Sigma}} = \mathbf{I}_d\mathbf{R}\mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d$. Note that we aim to maximize $\mathcal{L}(\mathbf{R})$ under the orthogonality constraint $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. This can be achieved by using Lie Group Methods [88]. More precisely, we search over the Lie group SO(n) of orthogonal matrices by computing the gradient in the corresponding Lie algebra $\mathfrak{so}(n)$. The gradient in $\mathfrak{so}(n)$ can be calculated as

$$\nabla\mathcal{L} = (\nabla_{\mathbf{R}}\,\mathcal{L})\,\mathbf{R}^\top\ -\ \mathbf{R}\,(\nabla_{\mathbf{R}}\,\mathcal{L})^\top.$$

After finding the search direction in the set of skew symmetric matrices we can compute the orthogonal update matrix by using the exponential map.

Derivation of divCSP using KL Divergence

From information theory [97] it is well known that the KL divergence between two zero mean Gaussians $g_i \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{\Sigma}}_i\right)$ and $f_j \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{\Sigma}}_j\right)$ has the following explicit representation

$$D\left((\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\ ||\ (\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right)$$
$$= \frac{1}{2}\left(\log\left|\bar{\mathbf{\Sigma}}_j\right|\ -\ \log\left|\bar{\mathbf{\Sigma}}_i\right|\ +\ \mathrm{tr}\left[(\bar{\mathbf{\Sigma}}_\mathrm{j})^{-1}\bar{\mathbf{\Sigma}}_\mathrm{i}\right]\ -\ \mathrm{d}\right).$$

Note that the log terms cancel out when using the symmetric divergence, however, an additional trace term (with swapped $\bar{\mathbf{\Sigma}}_i$ and $\bar{\mathbf{\Sigma}}_j$) appears.
The gradient of the divergence with respect to $\mathbf{R}$ can be computed separately for every term in the sum.
Let us rewrite the derivative of the log-determinant term

$$\nabla_{\mathbf{R}}\ \log\left|(\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right| = \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}\ \log\left|\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G}\right|\right]^\top$$

with $\mathbf{G} = \tilde{\mathbf{R}}^\top$ and $\tilde{\mathbf{R}}$ is the $d \times D$ matrix consisting of the first d rows of $\mathbf{R}$. According to [98] this is

$$\mathbf{I}_d^\top\left[2\tilde{\mathbf{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1}\right]^\top\ \text{ or }\ 2\mathbf{I}_\mathrm{d}^\top(\bar{\mathbf{\Sigma}}_\mathrm{j})^{-1}\mathbf{I}_\mathrm{d}\tilde{\mathbf{\Sigma}}_\mathrm{j}\mathbf{R}.$$

The derivative of the other log-determinant term can be computed in an analogous way and gives

$$2\mathbf{I}_d^\top(\bar{\mathbf{\Sigma}}_i)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_i\mathbf{R}.$$

The derivative of the trace term can be computed as follows. Let us rewrite

$$\nabla_{\mathbf{R}}\,\mathrm{tr}\left[((\mathbf{I}_\mathrm{d}\mathbf{R}\mathbf{P})\mathbf{\Sigma}_\mathrm{j}(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_\mathrm{d}^\top))^{-1}((\mathbf{I}_\mathrm{d}\mathbf{R}\mathbf{P})\mathbf{\Sigma}_\mathrm{i}(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_\mathrm{d}^\top))\right]$$
$$= \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}\,\mathrm{tr}\left[\left(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_\mathrm{j}\mathbf{G}\right)^{-1}\left(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_\mathrm{i}\mathbf{G}\right)\right]\right]^\top,$$

with $\mathbf{G}$ being defined as above. According to [98] this is

$$\mathbf{I}_d^\top\left[-2\tilde{\mathbf{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1}\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_i\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1} +\right.$$
$$\left.2\tilde{\mathbf{\Sigma}}_i\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_i\mathbf{G})^{-1}\right]^\top.$$

Thus the derivative of the trace term is

$$-2\mathbf{I}_d^\top\left((\bar{\mathbf{\Sigma}}_j)^{-1}\bar{\mathbf{\Sigma}}_i(\bar{\mathbf{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_j\ -\ (\bar{\mathbf{\Sigma}}_i)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_i\right)\mathbf{R}.$$

Derivation of divCSP using Beta Divergence

Beta divergence between two zero-mean Gaussians $g_i \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{\Sigma}}_i\right)$ and $f_j \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{\Sigma}}_j\right)$ is defined as

$$D_\beta\left((\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\ ||\ (\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right)$$
$$= \left(\frac{1}{\beta(\beta+1)}\int g_i^{\beta+1}(x)dx\ -\ \frac{1}{\beta}\int f_j^\beta g_i(x)dx\ +\right.$$
$$\left.\frac{1}{\beta+1}\int f_j^{\beta+1}(x)dx\right).$$

The integral $\int f_j^{\beta+1}(x)dx$ can be expressed as

$$\begin{aligned}
& \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta+1}{2}}}\int e^{-\frac{1}{2}x^T(\beta+1)\bar{\mathbf{\Sigma}}_j^{-1}x}dx \\
= & \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta+1}{2}}}\int e^{-\frac{1}{2}x^T(\frac{1}{\beta+1}\bar{\mathbf{\Sigma}}_j)^{-1}x}dx \\
\overset{*}{=} & \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta+1}{2}}}(2\pi)^{\frac{d}{2}}\left(\frac{1}{\beta+1}\right)^{\frac{d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{1}{2}} \\
= & \frac{1}{(2\pi)^{\frac{\beta d}{2}}(\beta+1)^{\frac{d}{2}}}|\bar{\mathbf{\Sigma}}_j|^{-\frac{\beta}{2}}
\end{aligned}$$

Note that step * assumes a Gaussian distribution under the integral, i.e. $\beta > -1$. The integral $\int g_i^{\beta+1}(x)dx$ can be computed in an analogous way.
The integral $\int f_j^\beta(x)g_i(x)dx$ is expressed in explicit form as

$$\begin{aligned}
& \frac{1}{(2\pi)^{\frac{\beta d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta}{2}}}\frac{1}{(2\pi)^{\frac{d}{2}}|\bar{\mathbf{\Sigma}}_i|^{\frac{1}{2}}}\int e^{-\frac{1}{2}x^T(\beta(\bar{\mathbf{\Sigma}}_j)^{-1}+(\bar{\mathbf{\Sigma}}_i)^{-1})x}dx \\
\overset{*}{=} & \frac{1}{(2\pi)^{\frac{\beta d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta}{2}}}\frac{1}{(2\pi)^{\frac{d}{2}}|\bar{\mathbf{\Sigma}}_i|^{\frac{1}{2}}}(2\pi)^{\frac{d}{2}}\left|\beta(\bar{\mathbf{\Sigma}}_j)^{-1}+(\bar{\mathbf{\Sigma}}_i)^{-1}\right|^{-\frac{1}{2}} \\
= & \frac{1}{(2\pi)^{\frac{\beta d}{2}}|\bar{\mathbf{\Sigma}}_j|^{\frac{\beta}{2}}|\bar{\mathbf{\Sigma}}_i|^{\frac{1}{2}}}\left|\beta(\bar{\mathbf{\Sigma}}_j)^{-1}+(\bar{\mathbf{\Sigma}}_i)^{-1}\right|^{-\frac{1}{2}} \\
= & \frac{1}{(2\pi)^{\frac{\beta d}{2}}}|\bar{\mathbf{\Sigma}}_j|^{\frac{1-\beta}{2}}\left|\bar{\mathbf{\Sigma}}_j(\beta(\bar{\mathbf{\Sigma}}_j)^{-1}+(\bar{\mathbf{\Sigma}}_i)^{-1})\bar{\mathbf{\Sigma}}_i\right|^{-\frac{1}{2}} \\
= & \frac{1}{(2\pi)^{\frac{\beta d}{2}}}|\bar{\mathbf{\Sigma}}_j|^{\frac{1-\beta}{2}}\left|\beta\bar{\mathbf{\Sigma}}_i+\bar{\mathbf{\Sigma}}_j\right|^{-\frac{1}{2}}
\end{aligned}$$

Note that also here step * assumes that $\beta(\bar{\mathbf{\Sigma}}_j)^{-1}+(\bar{\mathbf{\Sigma}}_i)^{-1}$ is symmetric positive definite. This assumption is always true for $\beta \geq 0$, however, it is violated for $\beta < c$ with c being some negative constant. Therefore we apply very small negative β values for divCSP-WS, more precisely we select the smallest possible β from $-0.0005, -0.0010, -0.0015, \ldots$

When using the symmetric beta divergence some terms cancel out and a simplified explicit representation can be derived (see Eq. (16)). As before we separately compute the gradient of each term of the beta divergence. The gradient of

TABLE II
GRADIENTS FOR THE DIVCSP METHODS USING KL DIVERGENCE.

Method	Gradient $\nabla_{\mathbf{R}}\,\mathcal{L}$
CSP term	$(1-\lambda)\mathbf{I}_d^\top \Big((\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2 \;-\; (\bar{\boldsymbol{\Sigma}}_1)^{-1}\bar{\boldsymbol{\Sigma}}_2(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 \,+$ $(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 \;-\; (\bar{\boldsymbol{\Sigma}}_2)^{-1}\bar{\boldsymbol{\Sigma}}_1(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2\Big)\,\mathbf{R}$
divCSP-WS	CSP term $\;-\;\frac{\lambda}{2N}\sum_{c=1}^{2}\sum_{i=1}^{N}\mathbf{I}_d^\top\Big((\bar{\boldsymbol{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c^i \;-\; (\bar{\boldsymbol{\Sigma}}_c)^{-1}\bar{\boldsymbol{\Sigma}}_c^i(\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c \,-$ $(\bar{\boldsymbol{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c^i \;+\; (\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c\Big)\,\mathbf{R}$
divCSP-BS	CSP term $\;-\;\frac{\lambda}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big((\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c}^k \;-\; (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{te,c}^k(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k \,+$ $(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k \;-\; (\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k(\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c}^k\Big)\,\mathbf{R}$
divCSP-AS	CSP term $\;-\;\frac{\lambda}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big((\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^\ell \;-\; (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k \,+$ $(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k \;-\; (\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k(\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^\ell\Big)\,\mathbf{R}$
divCSP-MS	CSP term $\;+\;\frac{\lambda}{K}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big((\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^k \;-\; (\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\bar{\boldsymbol{\Sigma}}_2^k(\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^k \,+$ $(\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^k \;-\; (\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\bar{\boldsymbol{\Sigma}}_1^k(\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^k\Big)\,\mathbf{R}$

$|\bar{\boldsymbol{\Sigma}}_j|^{-\frac{\beta}{2}}$ with respect to $\mathbf{R}$ can be computed when rewriting

$$\nabla_{\mathbf{R}}\,\left|(\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right|^{-\frac{\beta}{2}} = \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}\,|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta}{2}}\right]^\top$$

with $\mathbf{G} = \tilde{\mathbf{R}}^T$ and $\tilde{\mathbf{R}}$ is the $d\times D$ matrix consisting of the first d rows of $\mathbf{R}$. According to matrix cookbook [98] this is

$$\begin{aligned}&-\beta\mathbf{I}_d^\top|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta}{2}}\cdot\Big(\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G})^{-1}\Big)^\top\\ &= -\beta\mathbf{I}_d^\top|\bar{\boldsymbol{\Sigma}}_j|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_j\mathbf{R}.\end{aligned}$$

The gradient of $|\bar{\boldsymbol{\Sigma}}_i|^{-\frac{\beta}{2}}$ can be derived in an analogous way and gives

$$-\beta\mathbf{I}_d^\top|\bar{\boldsymbol{\Sigma}}_i|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_i\mathbf{R}.$$

Let us rewrite the gradient of $|\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}|\beta\bar{\boldsymbol{\Sigma}}_i+\bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}$ as

$$\begin{aligned}&\nabla_{\mathbf{R}}\,\Big[|(\mathbf{I}_d^\top\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)|^{\frac{1-\beta}{2}}\cdot|\beta(\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)+\\ &(\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)|^{-\frac{1}{2}}\Big]\\ &= \mathbf{I}_d^\top\Big[\nabla_{\mathbf{G}}\,\Big(|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{\frac{1-\beta}{2}}\cdot|\beta\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_i\mathbf{G}\;+\;\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{1}{2}}\Big)\Big]^T\end{aligned}$$

with $\mathbf{G}$ being defined as above. According to the product rule this is

$$\begin{aligned}&-\mathbf{I}_d^\top\Big[(\beta-1)|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta+1}{2}}\cdot|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|\cdot\\ &\Big(\mathbf{G}\tilde{\boldsymbol{\Sigma}}_j(\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G})^{-1}\Big)\cdot|\beta\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_i\mathbf{G}+\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{1}{2}}+\\ &|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{\frac{1-\beta}{2}}\cdot|\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}|^{-\frac{3}{2}}\cdot|\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}|\cdot\\ &((\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}(\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G})^{-1})\Big]^\top\end{aligned}$$

Writing it back gives

$$\begin{aligned}&-\mathbf{I}_d^\top\Big[(\beta-1)|\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_i+\bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}\cdot(\bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_j+\\ &|\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_i+\bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}\cdot(\beta\bar{\boldsymbol{\Sigma}}_i+\bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\Big]^\top\mathbf{R}\end{aligned}$$

PROOF OF THEOREM

Note that [70] has provided a proof for the special case of one spatial filter. Let $\tilde{\mathbf{R}}\in\mathbb{R}^{d\times D}$ denote the orthogonal projection onto a subspace of dimension d and let $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ represent the whitened covariance matrices with $\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2=\mathbf{I}$. Without loss of generality[8] we assume that $\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top=\boldsymbol{\Delta}_1$ and $\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top=\mathbf{I}-\boldsymbol{\Delta}_1$ with $\boldsymbol{\Delta}_1$ are diagonal matrices.

The KL divergence divCSP algorithm ($\lambda=0$) optimizes the following objective function $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ (ignoring constant terms)

$$\begin{aligned}&\operatorname{tr}\Big((\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top)^{-1}(\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top)\Big)\;+\\ &\operatorname{tr}\Big((\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top)^{-1}(\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top)\Big)\\ =\;&\operatorname{tr}(\boldsymbol{\Delta}_1^{-1}(\mathbf{I}-\boldsymbol{\Delta}_1))+\operatorname{tr}((\mathbf{I}-\boldsymbol{\Delta}_1)^{-1}\boldsymbol{\Delta}_1)\\ =\;&\sum_{i=1}^{d}\frac{1-\nu_i}{\nu_i}\;+\;\sum_{i=1}^{d}\frac{\nu_i}{1-\nu_i},\end{aligned}$$

where ν_i is the i-th diagonal element of $\boldsymbol{\Delta}_1$.

Let us decompose $\tilde{\mathbf{R}}=\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ into two matrices $\mathbf{U}\in\mathbb{R}^{k\times D}$ and $\mathbf{V}\in\mathbb{R}^{d-k\times D}$ as follows

$$\mathbf{U}=\left\{\mathbf{r}_i:\frac{1-\nu_i}{\nu_i}>\frac{\nu_i}{1-\nu_i}\right\}\Longrightarrow\nu_i<0.5$$

$$\mathbf{V}=\left\{\mathbf{r}_i:\frac{1-\nu_i}{\nu_i}\le\frac{\nu_i}{1-\nu_i}\right\}\Longrightarrow\nu_i\ge0.5.$$

Thus we can rewrite the objective function $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ as

$$\underbrace{\sum_{i=1}^{k}\frac{1-\nu_i}{\nu_i}\;+\;\frac{\nu_i}{1-\nu_i}}_{\mathbf{U}}\;+\;\underbrace{\sum_{i=k+1}^{d}\frac{1-\nu_i}{\nu_i}\;+\;\frac{\nu_i}{1-\nu_i}}_{\mathbf{V}}.$$

We prove that the top d CSP filters $\mathbf{W}$, i.e. the top d eigenvectors $\mathbf{v}_i$ ($i=1\ldots d$) of $\tilde{\boldsymbol{\Sigma}}_1$ sorted by $\alpha_i=\max\{\mu_i,1-\mu_i\}$ where μ_i denotes the i-th eigenvalue of $\tilde{\boldsymbol{\Sigma}}_1$, maximize $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$. Let us divide $\mathbf{W}$ into $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ as done above.

Case 1: Assume $\tilde{\mathbf{R}}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ and it consists of eigenvectors $\mathbf{v}_i$ of $\tilde{\boldsymbol{\Sigma}}_1$, but there exist $\mathbf{v}_j\in\tilde{\mathbf{R}}$ with $j>d$ (i.e. it is not among the top (according to the above sorting) d eigenvectors). Thus $\mathbf{v}_j\notin\mathbf{W}$ and there exist $\mathbf{w}_l\in\mathbf{W}$ (which

[8] Because the basis in the projected subspace is arbitrary, i.e. the Kullback-Leibler divergence is invariant to right multiplication of any non-singular matrix $\mathbf{G}\in\mathbb{R}^{d\times d}$ with $\mathcal{L}_{kl}(\mathbf{V})=\mathcal{L}_{kl}(\mathbf{VG})$.

TABLE III
GRADIENTS FOR THE DIVCSP METHODS USING BETA DIVERGENCE.

Method	Gradient $\nabla_{\mathbf{R}}\,\mathcal{L}$
CSP term	$(1-\lambda)\gamma\mathbf{I}_d^\top \Big(\beta\lvert\bar{\mathbf{\Sigma}}_1\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1 \;+\; \beta\lvert\bar{\mathbf{\Sigma}}_2\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2 \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_2\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_1 \;+\; \bar{\mathbf{\Sigma}}_2\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2 \;+\; (\beta\bar{\mathbf{\Sigma}}_1 \;+\; \bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_1 \;+\; \tilde{\mathbf{\Sigma}}_2)\Big] \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_1\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_2 \;+\; \bar{\mathbf{\Sigma}}_1\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1 \;+\; (\beta\bar{\mathbf{\Sigma}}_2 \;+\; \bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_2 \;+\; \tilde{\mathbf{\Sigma}}_1)\Big]\Big)\,\mathbf{R}$
divCSP-WS	CSP term $-\;\frac{\lambda\gamma}{2N}\sum_{c=1}^{2}\sum_{i=1}^{N}\mathbf{I}_d^\top\Big(\frac{\beta}{\beta+1}\lvert\bar{\mathbf{\Sigma}}_c^i\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_c^i \;+\; \frac{\beta^2}{\beta+1}\lvert\bar{\mathbf{\Sigma}}_c\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_c \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_c\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_c^i \;+\; \bar{\mathbf{\Sigma}}_c\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_c \;+\; (\beta\bar{\mathbf{\Sigma}}_c^i+\bar{\mathbf{\Sigma}}_c)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_c^i+\tilde{\mathbf{\Sigma}}_c)\Big]^\top\Big)\,\mathbf{R}$
divCSP-BS	CSP term $-\;\frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big(\beta\lvert\bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c}^k \;+\; \beta\lvert\bar{\mathbf{\Sigma}}_{te,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{te,c}^k \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_{te,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_{tr,c}^k \;+\; \bar{\mathbf{\Sigma}}_{te,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_{te,c})^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{te,c} \;+\; (\beta\bar{\mathbf{\Sigma}}_{tr,c}^k \;+\; \bar{\mathbf{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_{tr,c}^k \;+\; \tilde{\mathbf{\Sigma}}_{te,c}^k)\Big] \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_{te,c}^k \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_{tr,c})^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c} \;+\; (\beta\bar{\mathbf{\Sigma}}_{te,c}^k \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_{te,c}^k \;+\; \tilde{\mathbf{\Sigma}}_{tr,c}^k)\Big]\Big)\,\mathbf{R}$
divCSP-AS	CSP term $-\;\frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big(\beta\lvert\bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c}^k \;+\; \beta\lvert\bar{\mathbf{\Sigma}}_{tr,c}^\ell\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c}^\ell \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_{tr,c}^\ell\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_{tr,c}^k \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^\ell\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c}^\ell \;+\; (\beta\bar{\mathbf{\Sigma}}_{tr,c}^k \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_{tr,c}^k \;+\; \tilde{\mathbf{\Sigma}}_{tr,c}^\ell)\Big] \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_{tr,c}^\ell \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_{tr,c}^k \;+\; (\beta\bar{\mathbf{\Sigma}}_{tr,c}^\ell \;+\; \bar{\mathbf{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_{tr,c}^\ell \;+\; \tilde{\mathbf{\Sigma}}_{tr,c}^k)\Big]\Big)\,\mathbf{R}$
divCSP-MS	CSP term $+\;\frac{\lambda\gamma}{K}\sum_{k=1}^{K}\mathbf{I}_d^\top\Big(\beta\lvert\bar{\mathbf{\Sigma}}_1^k\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1^k \;+\; \beta\lvert\bar{\mathbf{\Sigma}}_2^k\rvert^{-\frac{\beta}{2}}(\bar{\mathbf{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2^k \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_2^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_1^k \;+\; \bar{\mathbf{\Sigma}}_2^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2 \;+\; (\beta\bar{\mathbf{\Sigma}}_1^k \;+\; \bar{\mathbf{\Sigma}}_2^k)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_1^k \;+\; \tilde{\mathbf{\Sigma}}_2^k)\Big] \;-$ $(\beta+1)^{\frac{d}{2}}\lvert\bar{\mathbf{\Sigma}}_1^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\mathbf{\Sigma}}_2^k \;+\; \bar{\mathbf{\Sigma}}_1^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1 \;+\; (\beta\bar{\mathbf{\Sigma}}_2^k \;+\; \bar{\mathbf{\Sigma}}_1^k)^{-1}\mathbf{I}_d(\beta\tilde{\mathbf{\Sigma}}_2^k \;+\; \tilde{\mathbf{\Sigma}}_1^k)\Big]\Big)\,\mathbf{R}$
	The constant γ is defined as $\gamma = -\frac{1}{\beta(2\pi)^{\frac{\beta d}{2}}(\beta+1)^{\frac{d}{2}}}$

is among the top d eigenvectors) with $\mathbf{w}_l \notin \tilde{\mathbf{R}}$.

Without loss of generality assume $\mathbf{v}_j \in \mathbf{U}$. In the following we prove

$$\frac{1-\nu_j}{\nu_j} + \frac{\nu_j}{1-\nu_j} < \frac{1-\nu_l}{\nu_l} + \frac{\nu_l}{1-\nu_l},$$

where ν_l and ν_j denote the diagonal element when applying $\mathbf{w}_l$ and $\mathbf{v}_j$, respectively. Note that the function $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (one can show this by taking the derivative).

Assume $\mathbf{w}_l \in \tilde{\mathbf{U}}$. Then $\nu_l < \nu_j < 0.5$ because $\mathbf{w}_l$ is selected before $\mathbf{v}_j$ (remember $\mathbf{v}_j \notin \mathbf{W}$) according to above sorting. Thus $f(\nu_j) < f(\nu_l)$ as $f(\nu)$ is maximized for the smallest argument ν (if $\nu < 0.5$).

Assume $\mathbf{w}_l \in \tilde{\mathbf{V}}$. Then $1 - \nu_l < \nu_j < 0.5$ because $\mathbf{w}_l$ is selected before $\mathbf{v}_j$ according to above sorting. Thus $f(\nu_j) < f(1-\nu_l) = f(\nu_l)$.

Let us define $\mathbf{B}$ as $\tilde{\mathbf{R}}$, but with $\mathbf{w}_l$ instead of $\mathbf{v}_j$. Thus $\mathcal{L}_{kl}(\tilde{\mathbf{R}}) < \mathcal{L}_{kl}(\mathbf{B})$. This is a contradiction to the assumption that $\tilde{\mathbf{R}}$ is the optimal solution.

Case 2: Assume $\tilde{\mathbf{R}}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ and there exist (at least one) $\mathbf{r}_j \in \tilde{\mathbf{R}}$ with $\mathbf{r}_j$ is not an eigenvector of $\tilde{\mathbf{\Sigma}}_1$. Without loss of generality assume $\mathbf{r}_j \in \mathbf{U}$. Let us define a new solution $\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{U}} \\ \tilde{\mathbf{V}} \end{bmatrix}$ as follows:

$\tilde{\mathbf{U}}$ consists of k eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ with smallest eigenvalues. $\tilde{\mathbf{V}}$ consists of $d-k$ eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ with largest eigenvalues.

Let us denote the diagonal elements (eigenvalues) of $\mathbf{U}\tilde{\mathbf{\Sigma}}_1\mathbf{U}^T$ as $\nu_1 < \ldots < \nu_k < 0.5$ and those obtained with $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{U}}^T$ as $u_1 < \ldots < u_k < 0.5$. Note that $u_i = \mu_i$ where $\mu_1 < \ldots < \mu_D$ are the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (because $\tilde{\mathbf{U}}$ consists of the smallest eigenvectors of $\tilde{\mathbf{\Sigma}}_1$). Cauchy's interlacing theorem [99] establishes the following relation between ν_i and u_i, namely $u_i \leq \nu_i$. Note that equality only holds if $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are the same, i.e. if $\mathbf{U}$ consists of the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (irrespectively of permutation). Cauchy's theorem implies that there are no ν_i and ν_j with $u_k < \nu_i < \nu_j < u_{k+1}$. Together with the fact that $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (i.e. for smallest ν in this case) this for all i implies

$$\frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i} \leq \frac{1-u_i}{u_i} + \frac{u_i}{1-u_i},$$

Since $\exists i$ where this relation is strictly positive (because we assumed $\mathbf{r}_j \in \mathbf{U}$), we obtain $\mathcal{L}_{kl}(\mathbf{U}) < \mathcal{L}_{kl}(\tilde{\mathbf{U}})$.

Let us denote the diagonal elements (eigenvalues) of $\mathbf{V}\tilde{\mathbf{\Sigma}}_1\mathbf{V}^T$ as $\nu_1 > \ldots > \nu_{d-k} \geq 0.5$ and those obtained with $\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{V}}^T$ as $u_1 > \ldots > u_{d-k} \geq 0.5$. Note that $u_i = \mu_i$ where $\mu_1 > \ldots > \mu_D$ are the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (because $\tilde{\mathbf{V}}$ consists of the largest eigenvectors of $\tilde{\mathbf{\Sigma}}_1$). Cauchy's interlacing theorem establishes the following relation between the ν_i and u_i, namely $\nu_i \leq u_i$. Note that equality only holds if $\mathbf{V}$ and $\tilde{\mathbf{V}}$ are the same (irrespectively of permutation). Together with the fact that $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (i.e. for largest ν in this case) this implies

$$\frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i} \leq \frac{1-u_i}{u_i} + \frac{u_i}{1-u_i},$$

Thus $\mathcal{L}_{kl}(\mathbf{V}) \leq \mathcal{L}_{kl}(\tilde{\mathbf{V}})$ and consequently $\mathcal{L}_{kl}(\tilde{\mathbf{R}}) = \mathcal{L}_{kl}(\tilde{\mathbf{U}}) + \mathcal{L}_{kl}(\tilde{\mathbf{V}}) < \mathcal{L}_{kl}(\tilde{\mathbf{U}}) + \mathcal{L}_{kl}(\tilde{\mathbf{V}}) = \mathcal{L}_{kl}(\tilde{\mathbf{B}})$. This contradicts the assumption that $\mathbf{R}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$.

Wojciech Samek obtained a Master's degree at the Department of Computer Science, Humboldt University of Berlin, in 2010. During his studies he was scholar of the German National Academic Foundation and was visiting Heriot-Watt University and University of Edinburgh from 2007 to 2008. In 2009 he was working for the Intelligent Robotics Group at NASA Ames in Mountain View, CA. Since 2010 he is pursuing his PhD at Technische Universität Berlin and is a PhD Fellow at the Bernstein Center for Computational Neuroscience Berlin. In 2012 and 2013 he was a guest scientist at ATR Institute International in Kyoto, Japan. His research interests include machine learning, biomedical engineering, neuroscience and computer vision.

Motoaki Kawanabe obtained a Master's degree at the Department of Mathematical Engineering, University of Tokyo, Japan. He studied mathematical statistics and received PhD from the same Department in 1995 where he worked as an assistant professor afterwards. He joined the Fraunhofer Institute FIRST in 2000 as a senior researcher. Until fall 2011 he has led the group for the THESEUS project on image annotation and retrieval there. He stayed at Nara Institute of Science and Technology, Kyoto University and RIKEN in Japan in 2007. Now he is with ATR research in Kyoto, Japan. His research interests include computer vision, biomedical data analysis, statistical signal processing and machine learning.

Klaus-Robert Müller has been a professor of computer science at Technische Universität Berlin since 2006; at the same time he has been the director of the Bernstein Focus on Neurotechnology Berlin. He studied physics in Karlsruhe from 1984 to 1989 and obtained his Ph.D. degree in computer science at Technische Universität Karlsruhe in 1992. After completing a postdoctoral position at GMD FIRST in Berlin, he was a research fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a professor at the University of Potsdam. He was awarded the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, and, in 2006, he received the SEL Alcatel Communication Award. In 2012, he was elected to be a member of the German National Academy of Sciences-Leopoldina. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.