Decomposable Common Spatial Patterns: Applying Graphical Model Selection in the Brain-Computer Interface Domain

Jannik Frank Faarkrog, Mikkel Holm Søgaard and Nikolaj Andersen

Department of Computer Science, Aalborg University

***Abstract*—When working with Electroencephalography (EEG) in the Brain-Computer Interface (BCI) domain, variability both subject-to-subject and session-to-session, results in the need to acquire labeled calibration data. This process can be time consuming, and as the number of channels grows, so does data requirements due to the curse of dimensionality.**

We address this problem by using model selection in undirected Gaussian graphical models, to reduce the number of parameters requiring estimation. The model is represented as a sparse precision matrix, where zeros are introduced by model selection, representing conditional independence between channels. We adapt the established method Common Spatial Patterns (CSP) into Decomposable CSP (DCSP) which is derived using precision matrices, instead of covariance matrices.

The approach is evaluated on both an existing and a novel dataset, and is compared to CSP. We find that DCSP outperforms CSP, with relative improvements of up to 50% in error rates Furthermore, we find that simply converting to use precision as features constitutes an improvement, and conjecture this is due to robustness against outliers. The improvement from the reduced number of parameters is more significant as the number of channels grow.

I. Introduction

Brain-Computer Interfaces (BCIs) have become an increasingly popular research area, since they were introduced as a concept in 1973 by Vidal [1]. As a result, different kinds of BCIs have emerged, such as non-invasive (functional Magnetic Resonance Imaging (fMRI) [2], Electroencephalography (EEG) [3]), partially invasive (Electrocorticography (ECoG) [4], [5]) and invasive [6] BCIs. EEG in particular, is becoming increasingly available to the general public because of its low cost and its non-invasive nature.

However, the variability both from one user to another and from session-to-session, means that most BCI solutions, have to employ a calibration phase, where labeled data is required in order to achieve good results. This process can be time-consuming and, as the number of EEG sensors increases for these low-cost devices, the data analysis requirements grow as well, due to the curse of dimensionality.

To reduce the need for calibration data, a number of approaches have been explored in the literature, such as (i) including prior knowledge about the brain, as in beamforming [7], (ii) using information from other subjects to construct prior knowledge [8], or (iii) reducing the dimensionality of the data. The dimensionality reduction can be done in different ways; by channel selection, given prior knowledge about the brain regarding the different brain lobes [9], or by source separation through use of either Independent Component Analysis (ICA) [10][11] or Principal Component Analysis (PCA) [12].

This paper will address the curse of dimensionality in the BCI domain, by adapting the Common Spatial Patterns (CSP) algorithm into using precision matrices, instead of covariance matrices. We name this solution Decomposable CSP (DCSP). Changing CSP to use precision matrices instead of covariance matrices, allows exploitation of conditional independence. Channels that are conditionally independent, will result in zeros in the precision matrix, and therefore reduce the number of parameters to estimate. Zeros in the covariance matrix, however, will result in marginal independence between channels. In terms of modeling the relationship between measurements on the scalp, conditional independence makes more sense. It allow channels that are not directly connected, to influence each other through intermediate channels, while marginal independence means that they can not influence each other at all.

To generate and estimate the precision matrices required by the adaption, stepwise forward-selection is used to find the optimal graphical model, in terms of goodness of fit contra complexity. By using Fisher's Linear Discriminant [13], the obtained recordings are then classified into binary classes. It should be mentioned, that any classification method can be used.

We perform tests on two different datasets, where one is novel, created during the writing of this paper. In these tests we find a relative increase in accuracy of up to 50% when comparing DCSP to CSP, and that model selection had a larger impact, when more channels were used. We also find that using precision matrices instead of covariance matrices is an improvement by itself, in all of the performed tests.

Section II will describe the necessary background in relation to BCIs and model selection. Then in Section III, DCSP will be derived and its implementation described. Section IV will describe the datasets that were used in the experiments. The results can be found in Section V, while Section VI will evaluate on the conducted experiments. Finally, we conclude on the suggested method in Section VII and discuss future work in Section VIII.

II. Background

A. Brain-Computer Interfaces

A BCI acts as a medium between the brain and some external device. This medium translates the electrical signals, that are generated by neurons firing inside the brain, into commands that the external device can interpret and act upon.

A BCI using EEG, like the Emotiv Epoc [14] used in the HumanSensing dataset (see Section IV-B), measures the electrical signals on the scalp by voltage differences caused by ionic current from the neurons [15]. The number of electrodes used in a BCI varies, depending on whether the BCI is for consumer or clinical use. Consumer BCIs tend to have fewer electrodes (e.g. 1 [16] and 14 [14] + references) and are connected to a computer wirelessly, while clinical BCIs (EEG caps) often have a higher number, ranging from 16 and up to 256 electrodes [17].

As a rule of thumb; a BCI with many electrodes allows for a higher spatial resolution, but is often more expensive (in terms of price and computation of data) and takes more time to place on a subject.

1) Electrode Placement: The human brain can be divided into six lobes [18]: The frontal, temporal, parietal, occipital, limbic lobes, and the insular cortex, of which the last two are located within the center of the brain. Each of these lobes have their own set of functions, which have been researched over time and resulted in the Brodmann Atlas [19], that is based on the Brodmann areas [20].

Depending on what one wants to measure with EEG, the electrodes must be placed accordingly, in order to obtain the relevant signals. To help standardize the placement of the electrodes, Jasper [21] published the international 10-20 system.

The naming convention of the electrodes consists of a letter followed by a number, as can be seen in Figure 1.

The letters (F, T, C[1], P, O) denote the brain lobe they are placed on, and the numbers (z(ero), and 1 through 8) denote how close the electrodes are placed to the center line (going from the nasion[2] to the inion[3]). Even numbers corresponds to the right hemisphere of the brain, and uneven numbers corresponds to the left hemisphere of the brain.

An extended system exists, called the 10-10 system [23], that supports more electrodes than the 10-20 system. The Emotiv Epoc headset utilizes the 10-10 system, as it has two additional electrodes (AF3 and AF4), found only in that system.

2) Motor Imagery: The frontal lobe can be divided into sub-cortices [24, p. 7]: The prefrontal-, motor- and premotor cortex. At the back end of the frontal lobe, the premotor cortex (Brodmann Area 6 [19]) is responsible for organizing movement [24, p. 4].

This means that before any limb of the body is moved physically (using the motor cortex in Brodmann Area 4), electrical activity can be measured in the pre-motor cortex.

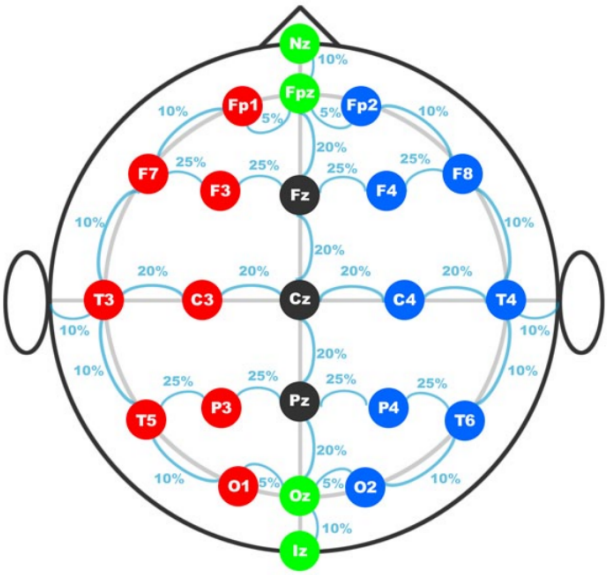


Figure 1: *The 10-20 system for placing EEG electrodes correctly on the scalp. The letter denotes what brain lobe an electrode is placed on, while the number denotes how far the electrode is from the center line. Figure retrieved from [22].*

In the BCI domain, this is called motor imagery, and allows for reading the intent of an action, i.e. lifting an arm, without performing the lifting action itself. Using EEG, motor imagery can be detected by Event Related Desynchronization (ERD) and Event Related Synchronization (ERS) [25]; a decrease and increase, respectively, in the band power of the signal. These pre-motor intents are measured within the μ- (8-13 Hz) and β (14-30 Hz) frequency bands [26], [27].

3) Event-related Potentials: Another research area exists within the BCI domain, called Event-Related Potentials (ERP) [28]. This area investigates how the brain responds to external sensory-, and cognitive stimuli. As an example of use, ERP can be used for classifying when the brain detects new or unusual stimuli [29], [30].

Like motor imagery, ERP can be measured with EEG electrodes, in accordance to the 10-20 system, on the Fz, Cz and Pz positions. The difference is that ERP can be acquired at a larger temporal resolution [31], allowing brain activity to be measured down to one millisecond, depending on the hardware being used.

An ERP signal is a waveform that can contain a number of components, in the form of peaks or troughs. Depending on when these peaks or troughs occur after the stimuli onset, they will have different meanings [32]: Components peaking within the first 100 milliseconds after some provided stimuli, are said to be sensory components, like the P50 wave which

[1](C)entral lobe does not exist - it is only for identification purposes
[2]The depressed area right above the bridge of the nose
[3]The lowest point of the skull, from the back of the head

represents sensory gating[4] [33]. Components peaking later are said to be cognitive components, like the P300 wave which indicates evaluation of the stimuli [34], [35].

4) BCI Pipeline: BCIs can be constructed in several ways depending on which methods are used in the pipeline. Basically, the BCI can be considered to consist of three steps: Preprocessing, feature extraction and classification. After classification, the system may give the subject feedback, but that depends on the system in question.

a) Preprocessing: The raw signal is filtered to reduce noise, and also filtered into certain frequencies, perhaps re-referenced, and then divided into trials.

b) Feature Extraction: Each trial is processed and relevant features are extracted. Common techniques include source separation and spatial filtering.

c) Classification: The features are classified and assigned a label that can serve as a control signal for a feedback application, or be used to analyze the mental state of the subject.

The pipeline may be tailored to the specific domain for which the BCI is built, in this case for classifying intended/imagined movement; motor imagery.

5) Spatial filtering: A spatial filter is a vector that describes a spatial weighting of the channels. It is used as a feature transformation to linearly combine channel values into more discriminating signals. If there are fewer spatial filters than channels, the filtering will reduce the dimensions of the data. Spatial filters can also be used to isolate sources at particular points, re-reference channels, or attenuate noisy ones. A number of spatial filtering methods exists, where Common Average Reference (CAR) [36], Surface Laplacian Spatial Filtering [36], Common Spatial Patterns (CSP) [37], [38], Principal Component Analysis (PCA) [12] and Independent Component Analysis (ICA) [10], [11] are often used methods.

The main reason for using spatial filters, is the smearing of signals that happen as they propagate from the brain to the surface of the skull. A simulation study [39], that explored the electric conductance of the brain, found that only 50% of the signal originated from within 3 cm of the electrode. This smearing means that a single source, or area of interest inside the head, may register across several electrodes, and similarly have uninteresting neighboring sources mixed in.

B. Common Spatial Patterns

CSP finds the spatial filters that maximize the variance for one condition, while minimizing the variance for another condition. The filtering is applied as a linear transformation on the signal: $\boldsymbol{w}^\top \boldsymbol{X}$, with $\boldsymbol{w}$ as the spatial filter and $\boldsymbol{X}$ as the signal. If the signal is filtered to a specific frequency band, and has a zero mean, the calculation of the transformed variance can be simplified[5] to: $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$. The maximization problem for finding the most discriminative filter is:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}^{(+)} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}^{(-)} \boldsymbol{w}}, \tag{1}$$

where $\boldsymbol{\Sigma}^{(c)}$, $(c \in \{+, -\})$ are the estimated covariance matrices for the two conditions, respectively. Typically more than one filter is selected, as using more filters will give better discriminative power.

CSP can be solved by simultaneous diagonalization [38], formulated as the solution to:

$$\begin{aligned} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{(+)} \boldsymbol{W} &= \boldsymbol{\Lambda}^{(+)} \\ \boldsymbol{W}^\top \boldsymbol{\Sigma}^{(-)} \boldsymbol{W} &= \boldsymbol{\Lambda}^{(-)}, \end{aligned} \tag{2}$$

where $\boldsymbol{W}$ consists of the $\boldsymbol{w}$ filters as columns (see Equation 1), and $\boldsymbol{\Lambda}^{(c)}$ are diagonal matrices with the transformed variance for each filter $\boldsymbol{w}$ along the diagonal. The problem of finding the filters, can be solved through the generalized eigenvalue decomposition problem [40]:

$$\boldsymbol{\Sigma}^{(+)} \boldsymbol{w} = \lambda \boldsymbol{\Sigma}^{(-)} \boldsymbol{w} \tag{3}$$

However, since an eigenvector $\boldsymbol{w}$ can be scaled arbitrarily and still be a solution, a constraint on the scaling is introduced:

$$\boldsymbol{\Lambda}^{(+)} + \boldsymbol{\Lambda}^{(-)} = \boldsymbol{I}, \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix.

In terms of the generalized eigenvalue problem, each column of $\boldsymbol{W}$ is a generalized eigenvector and each element on the diagonal of $\boldsymbol{\Lambda}^{(c)}$ is an eigenvalue such that $\lambda_j = \boldsymbol{\Lambda}_{jj}$ and:

$$\lambda_j^{(+)} + \lambda_j^{(-)} = 1 \tag{5}$$

From the above relation between eigenvalues, it can be observed that the larger the eigenvalue is for one condition, the smaller it is for the other condition. This means that the worst filter for maximizing the difference in variance for the first condition, will be the best for the second condition, and vice versa. This leads to that the eigenvalue decomposition only needs to be performed once, as filters for both conditions can be obtained from the solution to either problem. Also, as the eigenvalues can not be negative, and can be interpreted as variance of the spatially filtered signal, the transformed signal will have an average variance constrained between 0 and 1. Consequently, the original optimization problem in Equation 1 is maximized in the following way:

$$\max \frac{\lambda_j^{(+)}}{\lambda_j^{(-)}} \tag{6}$$

by either increasing the value for the first condition, or decreasing the value for the second condition.

[4]The ability to selectively attend to relevant stimuli, and ignore repetitive stimuli, in order to protect the brain from information overload. This is also known as the cocktail party effect.

[5] As $\boldsymbol{w}^\top \boldsymbol{X}$ is a vector, the variance can be calculated as: $\boldsymbol{w}^\top \boldsymbol{X} (\boldsymbol{w}^\top \boldsymbol{X})^\top = \boldsymbol{w}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{w}$, scaled with $\frac{1}{T}$ to get: $\frac{1}{T} \boldsymbol{X} \boldsymbol{X}^\top = \boldsymbol{\Sigma}$.

C. Undirected Gaussian Graphical Models

Given a covariance matrix $\boldsymbol{\Sigma}$, it is possible to interpret several characteristics from the observed data, such as the marginal dependence between channels on the off-diagonal elements, as well as the band power of individual channels along the diagonal. Marginal dependence is a measure of how much two channels interact regardless of the other channels. Thus, attempting to reduce the number of parameters in the covariance matrix (by constraining them to zero), will introduce marginal independence. Having marginal independence between two channels, means that changes in the value of one channel does not influence the other channel. This also means that changes cannot propagate through others channels, as such marginal independent channels must be completely disjoint.

The precision matrix $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}$ on the other hand, reveals the partial correlation coefficients between channels, which shows direct (in)dependence, under the condition that the values of the other channels are known. If an element in the precision matrix is zero, $p_{i,j} = 0,\ p_{i,j} \in \boldsymbol{P}$, the two channels are independent, given the rest of the channels, written as: $\boldsymbol{c}_i \perp\!\!\!\perp \boldsymbol{c}_j \mid \mathcal{C} \setminus \{\boldsymbol{c}_i, \boldsymbol{c}_j\}$. This type of relationship can be represented in an undirected graph $\mathcal{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$, where $\boldsymbol{V}$ is a set of vertices corresponding to the channels $\mathcal{C}$, $\boldsymbol{E}$ is a set of edges, and the graph contains an edge from channel c_i to c_j when $p_{i,j} \neq 0$.

More generally for a graphical model, two subsets of channels are *conditionally independent* if the global Markov property is obeyed [41]; that is, if for any triple $(\mathcal{A}, \mathcal{B}, \mathcal{S})$ of disjoint subsets of $\boldsymbol{V}$ such that $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$ in $\mathcal{G}$:

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S} \tag{7}$$

The graph $\mathcal{G}$ is said to be decomposable if $\boldsymbol{V} = \mathcal{A} \cup \mathcal{B} \cup \mathcal{S}$, such that (i) $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$, (ii) $\mathcal{S}$ is a complete subset of $\boldsymbol{V}$, and (iii) $\mathcal{A}$ is a clique or $\mathcal{G}_{\mathcal{A}\cup\mathcal{S}}$ is decomposable and similar for $\mathcal{B}$.

A clique is a maximal complete subset, meaning no more vertices can be added, while remaining complete. A decomposable model can be represented as a set of cliques and a set of separating subsets, which will be referred to as separators.

Given a decomposition, the precision matrix can be calculated from the subsets and thereby reduce the problem into smaller sub-problems. For a graph decomposed into $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{S}$, the precision is calculated as:

$$\boldsymbol{P} = [\boldsymbol{P}_{\mathcal{A}\cup\mathcal{S}}]^{\mathcal{C}} + [\boldsymbol{P}_{\mathcal{B}\cup\mathcal{S}}]^{\mathcal{C}} - [\boldsymbol{P}_{\mathcal{S}}]^{\mathcal{C}}, \tag{8}$$

where $[\boldsymbol{P}_{\mathcal{A}}]^{\mathcal{C}}$ means the elements in the sub-matrix $\boldsymbol{P}_{\mathcal{A}}$ are assigned to the positions specified by $\mathcal{A}$ in a zero matrix of size $|\mathcal{C}| \times |\mathcal{C}|$. This operation ensures that the subsets are added to the correct places in $\boldsymbol{P}$.

D. Model Selection

The selection of edges in the graphical model is based on which edge would increase the fitness with the sample precision matrix the most. In general, a fully connected model will always result in the best fit. However, estimating a fully connected model requires more data, given the higher amount of parameters, and runs the risk of overfitting the training data. Therefore, sparse models are desired in such a way, that the less relevant parameters get excluded.

Methods exist which penalize more complex models. Some of these methods are the *Akaike Information Criterion (AIC)* [42] and the *Bayesian Information Criterion (BIC)* [43]. Both of these methods are based on log-likelihood calculations, and are used to calculate the relative fitness between two models. Typically the fitting process starts either with a fully connected, or fully disjoint, model and either add or remove edges, based on the selected *Information Criterion* (IC) [41].

The difference between AIC and BIC lies in the penalty term:

$$\begin{aligned} \mathrm{IC}(k) &= -2\log L_{\mathcal{G}} + k\,\dim(\mathcal{G}) \\ \mathrm{AIC} &= \mathrm{IC}(2) \\ \mathrm{BIC} &= \mathrm{IC}(\log(T)), \end{aligned} \tag{9}$$

where $L_{\mathcal{G}}$ is the maximum likelihood of the model $\mathcal{G}$, $\dim(\mathcal{G})$ is the number of free parameters in the model $\mathcal{G}$ and T is the number of observations. For both methods, the penalty is based on the number of free parameters of the model, multiplied by a parameter k, for a trade-off between fitness and model complexity.

III. Methods

The effect of model selection in graphical models can be illustrated, by comparing a fully connected graphical model (see Figure 2) and a sparsely connected graphical model (see Figure 3). In the sparse model, the least correlated channels will not be connected. As physically distant channels are less likely to be correlated, there are fewer connections across the model. The model in Figure 3 is both sparse and decomposable which gives it a number of desirable properties, as discussed in Section II-C.

If the model had to experience marginal independence, the independent channels would not be allowed to be connected, not even through intermediate channels. Thus the graph would have to be divided into disjoint subgraphs. A model of disjoint subgraphs would translate to having two areas of the brain completely disconnected from each other, with no correlation between them. Conditional independence provides properties that are better aligned with the structure of the brain. Therefore, decomposable models can be used in BCI to construct precision matrices that, more precisely, model the relations between measurements on the scalp.

A. Adapting CSP Into DCSP

CSP can be shown to produce equivalent solutions regardless of whether it is based on covariance or precision matrices. Observe the factorization of $\boldsymbol{W}$ derived in a solution to simultaneous diagonalization, as given by Fukunaga [40, p. 31]:

$$\boldsymbol{W} = \boldsymbol{\Phi}\boldsymbol{\Theta}^{-\frac{1}{2}}\boldsymbol{\Psi}, \tag{10}$$

where both $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are orthonormal matrices and Θ is a diagonal matrix.

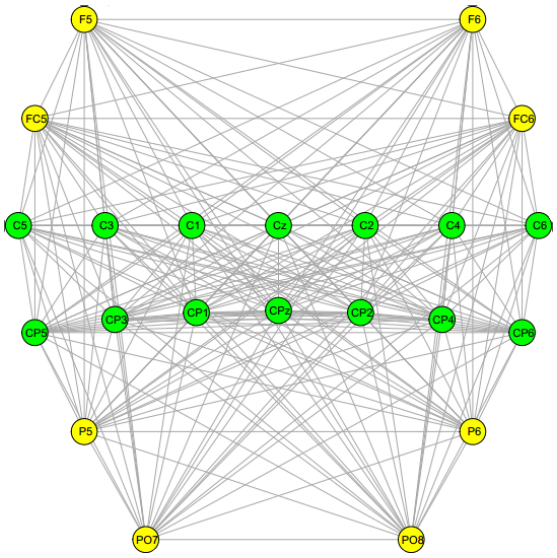


Figure 2: *A fully connected graphical model. Here, all 22 channels have a connection to each other, meaning that each channel directly affects all of the other channels, regardless of their location. The position of channels in the figure, corresponds to their physical placement on the scalp, with the top of the figure being the front of the head.*

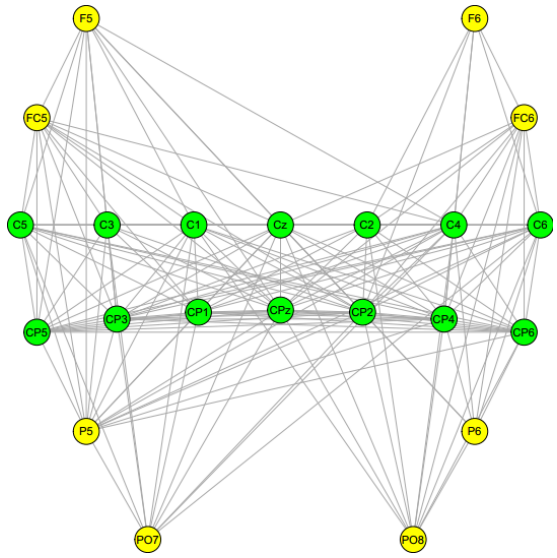


Figure 3: *A sparsely connected graphical model. Here, the graphical model selection has removed edges between some of the channels, meaning that they do not affect each other directly. Note: In this model, 4 trials were randomly selected as training set, and penalty parameter $k = 8$ was used in the model selection.*

By substituting Equation 10 into Equation 2, $\mathbf{\Lambda}^{(+)}$ and $\mathbf{\Lambda}^{(-)}$ can be derived:

$$\begin{aligned}\mathbf{\Psi}^\top \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Phi}^\top \mathbf{\Sigma}^{(+)} \mathbf{\Phi} \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Psi} &= \mathbf{\Lambda}^{(+)} \\ \mathbf{\Psi}^\top \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Phi}^\top \mathbf{\Sigma}^{(-)} \mathbf{\Phi} \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Psi} &= \mathbf{\Lambda}^{(-)}\end{aligned} \tag{11}$$

Inverting Equation 11 will then provide the diagonalization for the precision matrices:

$$\begin{aligned}\mathbf{\Psi}^\top \mathbf{\Theta}^{\frac{1}{2}} \mathbf{\Phi}^\top \boldsymbol{P}^{(+)} \mathbf{\Phi} \mathbf{\Theta}^{\frac{1}{2}} \mathbf{\Psi} &= \boldsymbol{D}^{(+)} \\ \mathbf{\Psi}^\top \mathbf{\Theta}^{\frac{1}{2}} \mathbf{\Phi}^\top \boldsymbol{P}^{(-)} \mathbf{\Phi} \mathbf{\Theta}^{\frac{1}{2}} \mathbf{\Psi} &= \boldsymbol{D}^{(-)},\end{aligned} \tag{12}$$

where $(\mathbf{\Lambda}^{(c)})^{-1} = \boldsymbol{D}^{(c)}$.

It is now shown that a diagonalization for the precision matrices exists. Furthermore, as both $\boldsymbol{D}^{(c)}$ and $\mathbf{\Lambda}^{(c)}$ are diagonal matrices, the ratio between eigenvalues in the two conditions is preserved, albeit inverted:

$$\frac{\lambda^{(+)}}{\lambda^{(-)}} = \frac{(\lambda^{(-)})^{-1}}{(\lambda^{(+)})^{-1}} \tag{13}$$

This means that the problem remains the same, but the optimization has been reversed. However, the matrices that diagonalize the covariance and precision matrices are not the same. As it can be seen in Equations 10, 11 and 12, the diagonalizing matrices are as follows:

$$\begin{aligned}\boldsymbol{W} &= \mathbf{\Phi} \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Psi} \\ \boldsymbol{V} &= \mathbf{\Phi} \mathbf{\Theta}^{\frac{1}{2}} \mathbf{\Psi},\end{aligned} \tag{14}$$

where $\boldsymbol{W}$ is the transform for the covariance matrix, and $\boldsymbol{V}$ is the transform for the precision matrix. These are not equal as in general $\Theta^{\frac{1}{2}} \neq \Theta^{-\frac{1}{2}}$. It is, however, possible to convert the solution for one simultaneous diagonalization to the other, allowing the filters from DCSP to be used for CSP:

$$(\boldsymbol{V}^\top)^{-1} = \mathbf{\Phi} \mathbf{\Theta}^{-\frac{1}{2}} \mathbf{\Psi} = \boldsymbol{W} \tag{15}$$

One important difference is due to the fact that inversion is not distributive. This influences the constraints, as:

$$(\mathbf{\Lambda}^{(+)})^{-1} + (\mathbf{\Lambda}^{(-)})^{-1} \neq (\mathbf{\Lambda}^{(+)} + \mathbf{\Lambda}^{(-)})^{-1} \tag{16}$$

Hence, the constraint from Equation 4 is not preserved through inversion, and a new constraint have to be chosen for DCSP:

$$\boldsymbol{D}^{(+)} + \boldsymbol{D}^{(-)} = \boldsymbol{I} \tag{17}$$

This constraint will result in precision features with an average between 0 and 1, similar to how CSP is constrained for variance. The difference is then, that an interval $]0, 1[$ in precision is equal to an interval $]1, \infty[$ in variance, since precision and variance are inverse of each other. This means the filters will have different constraints depending on the whether they are derived for CSP or DCSP. Even if the conversion in Equation 15 is used, they will obey the constraint from the method for which they were derived. Thus, although the ratio in Equation 13 is preserved, the distribution of features will be different for the two methods.

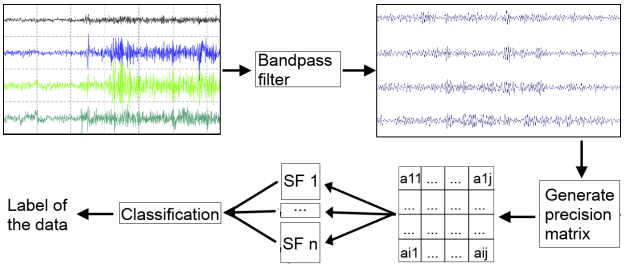


Figure 4: *The pipeline for DCSP, on runtime. The raw EEG input signal is filtered with a bandpass filter, to attenuate artifacts and irrelevant frequencies, and then divided into trials. Using the selected model, a sparse precision matrix is generated from the trials and then spatially filtered to extract the features. The classifier then labels each trial.*

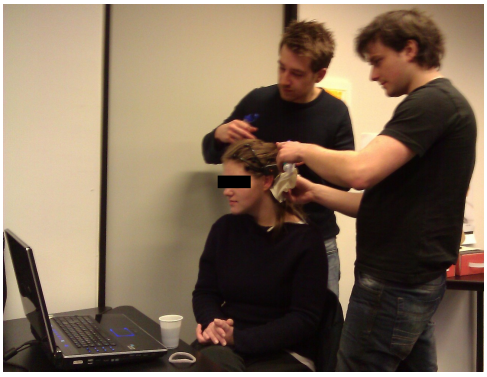


Figure 5: *A test subject is equipped with the Emotiv Epoc headset.*

Another difference for DCSP is the application of filters. As DCSP is used in combination with graphical model selection, it can not be applied as a linear transformation on the raw signal. Instead, the data must follow the relationships described by the selected graphical model, which are enforced when estimating each trial's precision matrix following Equation 8. When the precision matrices are acquired, the filters can be applied as:

$$\boldsymbol{V}^\top \boldsymbol{P} \boldsymbol{V} = \boldsymbol{D} \tag{18}$$

B. Pipeline

To test the effect of using DCSP in place of CSP, a pipeline was constructed by using the methods mentioned in Section II, and the runtime part is illustrated in Figure 4.

a) Preprocessing: The input signals are filtered with a bandpass filter, such that artifacts and activity in irrelevant frequencies are attenuated. Trials are extracted from the data and divided into a test set and a training set.

b) Feature Extraction CSP: From the training set, spatial filters are calculated and the 3 best filters for each condition are chosen, which should provide a reasonable accuracy [38].

c) Feature Extraction DCSP: Filters need to be chosen as in CSP, but here they are derived from sparse precision matrices instead of covariance matrices. However, to estimate the sparse precision matrices, a model must be selected as well. Model selection constructs a single model using the training set of both conditions. This model is then used to estimate the precision matrices both for calculating the spatial filters, and during the extraction of features. To select a model, a stepwise forward model selection is used. The model selection will only add edges, which will result in a decomposable graph. This restriction reduces the search space of possible models.

d) Classification: For classification, Fisher's Linear Discriminant (FLD) is used, although some other classifier could be used. It is trained on the features extracted from the training set, each trial from the test set is given a label, and the accuracy of the pipeline is calculated.

IV. Experimental setup

This section will describe the setup for two datasets, that have been used in this paper. First, a dataset from the BCI Competition III will be described, followed by a collaborative experiment conducted between two departments at Aalborg University.

A. BCI Competition

To properly estimate the effects of DCSP compared to CSP, a motor imagery dataset was required. The BCI Competition III dataset IVa [44] fulfilled this requirement, as well as having a suitable number of channels, such that model selection would be more meaningful. The dataset consists of 5 test subjects (*aa*, *al*, *av*, *aw*, *ay*), and was recorded using 118 channels with a 1000 Hz sample frequency, which was down-sampled to 100 Hz. The data was recorded over 4 sessions without feedback. During the recording, subjects were instructed to perform motor imagery on right hand, left hand and right foot, though only right hand and right foot were used in the competition. Subjects were presented with a cue for 3.5 seconds, after which there would be a short period of rest. The dataset for each subject consists of 140 trials per condition.

B. HumanSensing study

The study was conducted through collaboration between two departments (Computer Science and Humanistic Informatics) at Aalborg University. It was initially designed to measure skin conductance while subjects were watching short video clips, chosen to have a wide range of emotional impacts, but was extended so that subjects would also wear Emotiv Epoc headsets during the experiment.

Prior to the experiment, the test subjects would be seated in front of a laptop, and equipped with a headset, as seen in Figure 5.

In groups of three, the test subjects entered the experiment room and were seated next to each other and approximately 3 meters from a projector canvas. From this position they watched seven video clips for a total duration of a little over

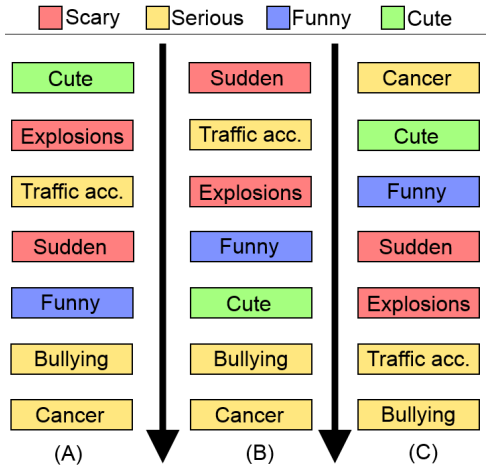


Figure 6: *3 different orders of the six video clips used in the HumanSensing study, where the arrow direction denotes the order from start to end. The video clips were divided into four categories: cute, funny, scary and serious.* ***A)*** *Day 1 before noon,* ***B)*** *Day 1 after noon,* ***C)*** *Day 2 before noon.*

10 minutes. The subjects were instructed to relax during the test, and refrain from moving as much as possible.

A black screen was displayed for 10 seconds in order to prepare the test subjects. This was done before the first video clip started, after the last video clip ended, and in between video clips.

Once the headsets were connected to their respective laptops, the test leader would press a button on each laptop to mark the start of the test.

In total, the recordings of 68 unique subjects were obtained.

The videos used in the HumanSensing study were found on the video-sharing website, YouTube. One video would be cute [45], another video would be funny [46], two videoes would be scary; involving explosions [47] or sudden appearance [48], and three videos would be about serious topics such as cancer [49], bullying [50] or traffic accidents [51].

The order of the six video clips were chosen at random, such that subjects during the same time of the day would see the clips in the same order, as can be seen in Figure 6.

V. Results

For the following tests, the goal was to explore the impact on error rate, when changing the amount of data available for calibration. Because of this, the tests were run with a decreasing amount of trials available in the training set, as well as various values of the penalty parameter, k, for the model selection.

The performance was evaluated by running a modified variation of cross-validation, where instead of using one fold as test set, it was used as training set. This allowed simulation of smaller training sets, to make the effects of overfitting more severe. As an example, a 7-fold test meant that for both conditions, the dataset was randomly split into 7 evenly-sized folds. Each fold was then iteratively used as a training set, while the remaining folds were used as test set. When all folds had been evaluated, the results of the 7 runs were averaged. Thus, by increasing the number of folds, less data would be available during calibration.

Each configuration was repeated 16 times and averaged. A configuration for DCSP is a k-parameter and the number of folds, and for CSP it is only the number of folds as it does not use model selection.

A. BCI Competition

The dataset from the BCI Competition was bandpass filtered in the range 8-30 Hz, as this is the general frequency band for motor imagery, as mentioned in Section II. The k-parameters used were 0, meaning the model was a fully connected, and powers of 2 from 2 to 512 (2, 4, 8, ..., 512).

1) Motor-imagery relevant channels only: A set of 14 channels were selected, that were related to motor-imagery. These channels can be seen in Figure 7.

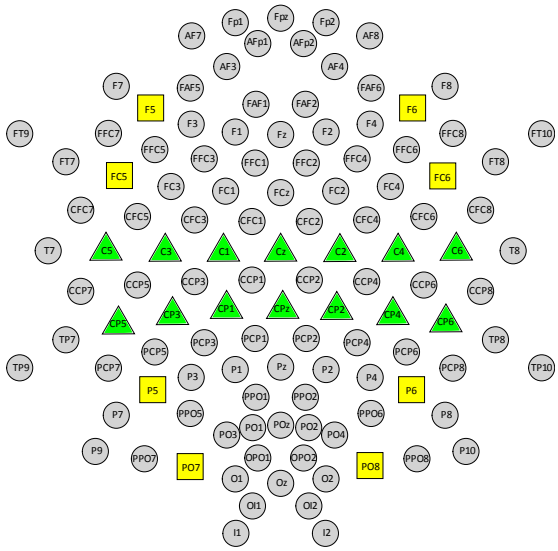
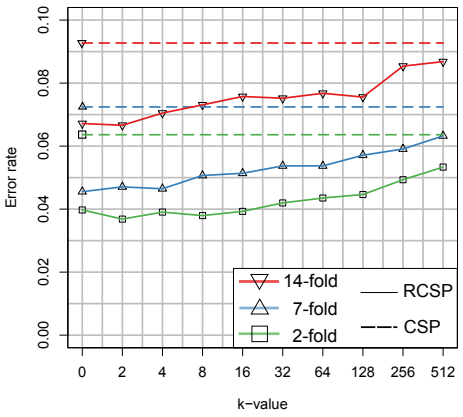


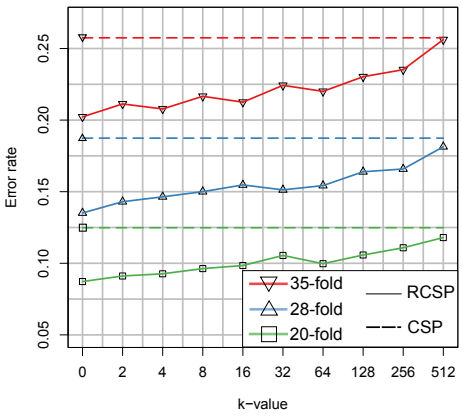
Figure 7: *A visualization of all 118 channels used in the 118-channel tests. The triangular green-colored channels were used in the 14-channel tests, and the squared yellow-colored channels, along with triangular channels, were used in the 22-channel tests.*

For this test, subject *al* was selected, because competitors in the BCI Competition were able to achieve good results with this subject. The results of the 14-channel test on subject *al*, can be seen in Figure 8.

For all tables, the error rate of each fold for the DCSP column corresponds to the k-parameter which gave the best result.

2) Motor-imagery relevant channels + distant irrelevant channels: To investigate if adding noisy channels to the dataset would affect the performance of DCSP, in comparison to CSP, a set of 22 channels was selected, such that 8 channels





Folds	CSP	DCSP	Rel. impr.
2	6.36±1.68	3.68±1.36	42.11
7	7.24±1.87	4.55±1.55	37.13
14	9.27±3.35	6.66±3.09	28.15
20	12.48±5.36	8.72±4.64	30.11
28	18.74±10.40	13.51±8.10	27.89
35	25.75±13.95	20.23±11.91	21.44

Figure 8: *Test results for subject* al*, with the 14-channel dataset - in regards to the error rate. Left figure: Results for 2-, 7-, and 14-folds. Center figure: Results for 20-, 28-, and 35-folds. Right figure: Summary of the best results (error % ± standard deviation), where Rel. impr. describes the relative improvement (percentage-wise) of DCSP's error rate, compared to CSP's.*

were expected not to contain discriminating information. An overview of the channels that were selected, can be seen in Figure 7.

In addition to subject *al*, another subject, *av*, was selected for this test. The reason is that subject *av* was the subject who produced the poorest results for the competitors.

The test results for subject *al* and *av*, can be seen in Figure 9 and 10, respectively.

3) All channels - 118 channels: A test on all 118 channels was also performed on subject *al*, with the same k-parameters as used in the previous tests. The test was repeated 8 times, but due to a bug in the used libraries, records for some configurations were lost. An overview of all the channels used in this test, can be seen in Figure 7.

An alternative model selection algorithm was chosen for this test, since the dimensionality of the problem made it unfeasible to start from an empty model. This model selection was initialized with a minimum spanning forest ensuring the model was not disjoint. As such, the k-parameters are not directly comparable to the other tests.

The results can be seen in Figure 11.

B. HumanSensing - 14 channels

Similar to the tests on the BCI Competition data, the signals were initially bandpass filtered, albeit to a broader spectrum (4-30 Hz), as this study does not rely on motor imagery.

4 trials, of 4 seconds each, were extracted for each subject during a dramatic section of the explosion video clip, and were labeled based on the gender. In total, 20 men and 30 women were chosen, while the remaining subjects were discarded due to missing gender identification in the records.

All 14 channels of the Emotiv Epoc headset were used, and the relationship between available data and model complexity was explored as in previous tests. The fold-values were set to: 2, 4, 8, 10 and 20, while the k-parameters were set to: 0 and powers of 2, from 2 to 128 (2, 4, 8, ..., 128).

The HumanSensing test results can be seen in Figure 12.

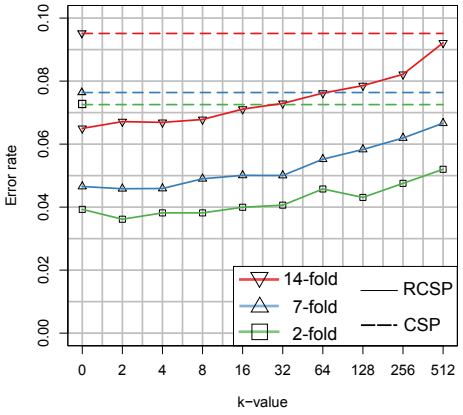
VI. Discussion

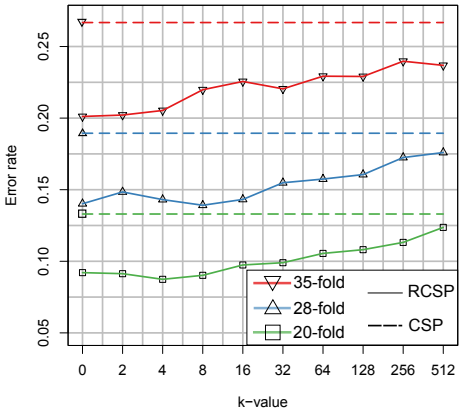
1) General Improvement: A general trend in all the performed tests is that the fully connected (penalty parameter $k = 0$) DCSP outperforms CSP (see Figure 8, 9, 10, 11 or 12). Since DCSP in these cases, estimates the same number of parameters as CSP, the effects of overfitting in the model should be the same. An explanation for the difference in performance, could lie in the different constraints (see Equation 4 and 17), and how this impacts the distribution of features. Exactly why the precision is preferable to band power is an open question. We conjecture that precision is more resistant to outliers and artifacts for the following reasons.

When applying CSP spatial filters, the output variance is constrained such that it is in the interval $]0, 1[$. However, artifacts in EEG will usually lead to outliers with great increases in band power, sometimes orders of magnitude. Because of the constraint, these outliers can affect the density of the feature distribution, such that the majority of features will be concentrated closer to zero.

When using precision, the inversion of high band power outliers will instead result in features that are closer to zero and due to this, the feature distribution will be less affected and become less dense. Accordingly, while variance is vulnerable to high outliers, precision is vulnerable to outliers very close to zero. We expect high outliers in band power to be more frequent than low outliers, and that this is the reason for DCSP's increased performance when using a fully connected model.

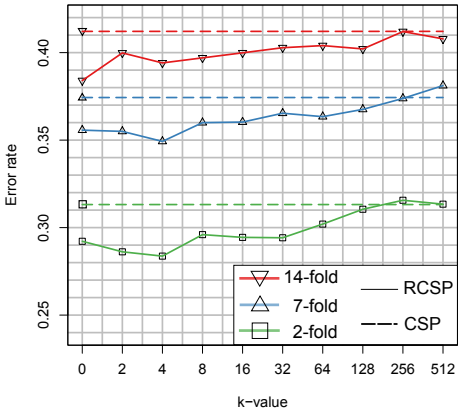
2) Noise: The added noise channels did not have a large effect on the differences between DCSP and CSP (see Figure 8 and 9). This suggest that both methods are fairly robust to spatially distributed noise, and the amount of distinguishing information in the noise channels is minimal. However, when looking at the case where the full channel-set is used, an

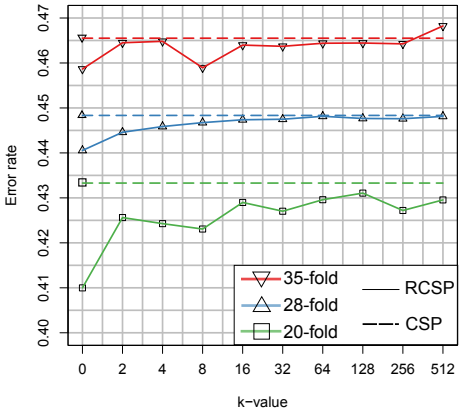




Folds	CSP	DCSP	Rel. impr.
2	7.25±1.79	3.62±1.26	50.15
7	7.64±1.84	4.58±1.37	39.99
14	9.51±3.24	6.50±3.33	31.68
20	13.30±6.05	8.74±4.70	34.30
28	18.94±10.32	13.91±8.79	26.55
35	26.67±14.35	20.11±13.10	24.59

Figure 9: *Test results for subject* al*, with the 22-channel dataset - in regards to the error rate. Left figure: Results for 2-, 7-, and 14-folds. Center figure: Results for 20-, 28-, and 35-folds. Right figure: Summary of the best results (error % ± standard deviation), where Rel. impr. describes the relative improvement (percentage-wise) of DCSP's error rate, compared to CSP's.*





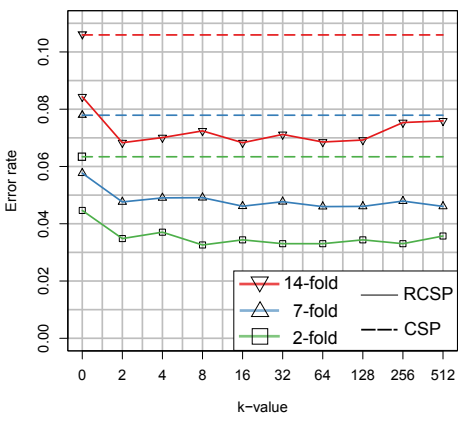
Folds	CSP	DCSP	Rel. impr.
2	31.32±4.26	28.37±3.72	9.41
7	37.43±5.27	34.93±6.10	6.70
14	41.21±6.33	38.40±6.29	6.82
20	43.33±6.41	41.00±7.36	5.38
28	44.83±6.88	44.05±7.16	1.74
35	46.55±6.75	45.87±7.34	1.47

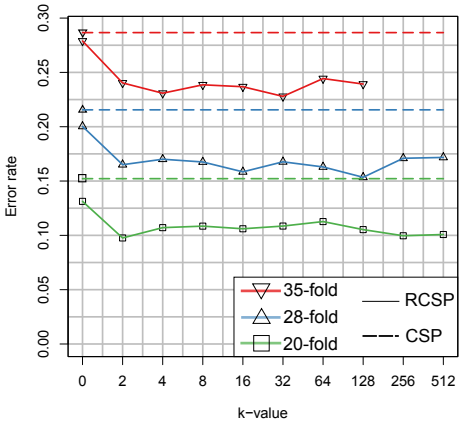
Figure 10: *Test results for subject* av*, with the 22-channel dataset - in regards to the error rate. Left figure: Results for 2-, 7-, and 14-folds. Center figure: Results for 20-, 28-, and 35-folds. Right figure: Summary of the best results (error % ± standard deviation), where Rel. impr. describes the relative improvement (percentage-wise) of DCSP's error rate, compared to CSP's.*

interesting trend can be observed (see Figure 11). Here, any penalty is preferable to the fully connected model and DCSP's relative improvement, compared to CSP, is greater than for the 14 channels experiment (see Figure 11 and 8). The importance of this is that instead of spending time and resources carefully selecting channels, it might be possible to simply use all of them and let the model selection handle the removal of less relevant connections. This might be interesting in BCI domains where there is little or no knowledge of which channels would contain discriminating information.

3) Penalty: It can be observed from the tests, that while some k-parameters result in a better error rate than the fully connected model, there is no fixed penalty which always result in improved error rate. This means that for model selection to become a consistent improvement, hyper parameter selection is required to choose a k-parameter that fits the data. Only the tests with 118 channels showed consistent improvement compared to a full model. This is most likely a result of the fully connected model being overfitted, having over 7000 parameters, where the model selected ones contain from about 1200 to 3000 parameters.

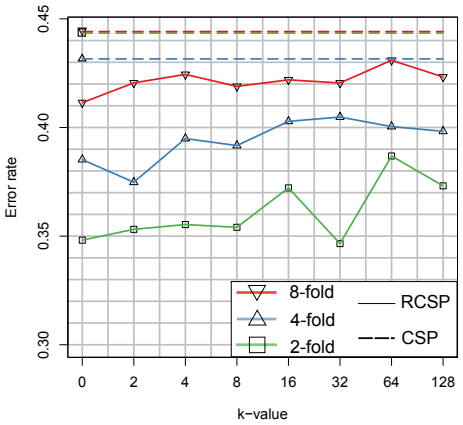
4) Training set size: As the number of samples in the training set decreases (a higher number of folds gives less trials in each fold), the error rate increases. This is no surprise, as in general, having more data should provide a better estimation. Similarly, the standard deviation among the averaged results shows, that the difference in error rate between repetitions also increases when less trials are available. Some single records from the 35-fold tests gave just as low error rates as in the 2-fold tests, they were just less frequent. This highlights the need for representative trials, when using very small training sets.

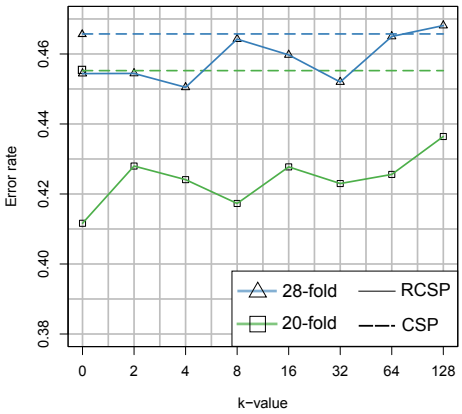




Folds	CSP	DCSP	Rel. impr.
2	6.34±1.71	3.26±1.33	48.59
7	7.79±1.65	4.60±1.84	40.97
14	10.59±5.00	6.83±2.83	35.53
20	15.22±7.73	9.76±5.43	35.87
28	21.55±11.79	15.35±10.00	28.81
35	28.66±16.01	22.79±14.09	20.48

Figure 11: *Test results for subject* al*, with the (full) 118-channel dataset - in regards to the error rate. Left figure: Results for 2-, 7-, and 14-folds. Center figure: Results for 20-, 28-, and 35-folds. Note: Due to a bug in the used libraries, the penalty parameter, $k = 256$ and $k = 512$, are omitted from the 35-fold results. Right figure: Summary of the best results (error % ± standard deviation), where Rel. impr. describes the relative improvement (percentage-wise) of DCSP's error rate, compared to CSP's.*





Folds	CSP	DCSP	Rel. impr.
2	44.34±4.34	34.66±4.96	21.85
4	43.16±3.84	37.48±4.15	13.15
8	44.42±3.91	41.14±4.03	7.37
10	45.52±3.83	41.16±4.73	9.58
20	46.57±4.08	45.05±5.03	3.27

Figure 12: *Test results for the HumanSensing dataset, with the Emotiv Epoc's 14-channel dataset - in regards to the error rate. Left figure: Results for 2-, 4-, and 8-folds. Center figure: Results for 10-, and 20-folds. Right figure: Summary of the best results (error % ± standard deviation), where Rel. impr. describes the relative improvement (percentage-wise) of DCSP's error rate, compared to CSP's.*

5) Subjects: The subjects *al* and *av* were chosen specifically because they were the subjects which achieved best and worst results in their dataset from the BCI competition III. Our results mirror this, having *al* achieve better error rates than *av*, and with DCSP being an improvement over CSP in both cases (see Figure 9 and 10). The tests were performed to ensure that our improvements were not specific to a single subject, but could be applied in general.

6) Runtime: While using DCSP seem to be a general improvement to CSP, there is a trade-off in relation to runtime, both for calibration and application of filters. DCSP requires a decomposable model from which it can estimate precision matrices. As precision matrices require matrix inversion and are needed at runtime, DCSP is significantly slower than CSP, and experiences further slowdown, as the dimensionality of the data grows.

7) HumanSensing: Although our approach is not specifically designed to differentiate gender or reaction to audiovisual stimuli in general, it was possible to achieve better than random results when applying DCSP. Interestingly, DCSP performed much better than CSP (see Figure 12), which never performed much better than almost random. The reason for this might be the same as for the general improvement discussed earlier, since the HumanSensing dataset is very noisy and all available channels were used.

VII. Conclusion

In this paper, we proposed Decomposable CSP (DCSP), an adaptation of CSP, which addresses the issue of the curse of dimensionality. It exploits conditional independence through

model selection, thereby reducing the number of parameters to estimate. The solution was tested with the BCI competition III dataset IVa and a completely new dataset, based on 68 subjects.

Adapting CSP into DCSP, has in all tests resulted in improved error rates, with relative improvements of up to 50%. Using precision features, even without model selection, resulted in improvements over CSP in all tests. The reason for this is conjectured to be because of how BCI signals contain outliers that affect precision less than band power. As EEG outliers tend to be of high values, which when converted to precision become very small, they are less dominant in the average.

Furthermore, when reducing the number of parameters through model selection, the result was highly dependent on choosing the penalty parameter k, since no single k-parameter gave consistently better results. When the dimensionality grew to 118 channels, the impact of model selection was higher, and any of the tested penalty terms were better than none. This indicates that model selection is mainly effective when given a higher number of channels, as having more channels makes it more likely that the model will be overfitted.

VIII. Future works

Execution speed at runtime could be improved significantly if the spatial filters from DCSP could be applied prior to matrix inversion. A solution could be similar to the conversion of filters as in Equation 15, and postponing the inversion until after the spatial filtering, as this would reduce the dimensionality of the inversion operation significantly.

Furthermore, the spatial filters are selected based on how well they discriminate the two conditions, yet the model selection does not reflect this goal. For selecting which edges get added in the graphical model, the edges that best describe the signal are chosen, not the ones that best discriminates the two conditions. Incorporating the principle of CSP in the information criterion, such that the model is evaluated based on its discriminative information, could help discard parameters which are similar in the two conditions and retain the ones that are not.

References

[1] J. J. Vidal, "Toward direct brain-computer communication," *Annual review of Biophysics and Bioengineering*, vol. 2, no. 1, pp. 157–180, 1973.

[2] R. Sitaram, A. Caria, R. Veit, T. Gaber, G. Rota, A. Kuebler, and N. Birbaumer, "FMRI brain-computer interface: a tool for neuroscientific research and treatment," *Computational intelligence and neuroscience*, vol. 2007, 2007.

[3] J. J. Vidal, "Real-time detection of brain events in EEG," *Proceedings of the IEEE*, vol. 65, no. 5, pp. 633–641, 1977.

[4] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain–computer interface using electrocorticographic signals in humans," *Journal of neural engineering*, vol. 1, no. 2, p. 63, 2004.

[5] P. Shenoy, K. J. Miller, J. G. Ojemann, and R. P. Rao, "Generalized features for electrocorticographic BCIs," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 1, pp. 273–280, 2008.

[6] P. R. Kennedy, R. A. Bakay, M. M. Moore, K. Adams, and J. Goldwaithe, "Direct control of a computer from the human central nervous system," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 2, pp. 198–202, 2000.

[7] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain–computer interfaces," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 1209–1219, 2009.

[8] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multi-task learning for Brain-Computer Interfaces," 2009.

[9] M. Naeem, C. Brunner, and G. Pfurtscheller, "Dimensionality reduction and channel selection of motor imagery electroencephalographic data," *Computational intelligence and neuroscience*, vol. 2009, 2009.

[10] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[11] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller, "Spatial filtering and selection of optimized components in four class motor imagery eeg data using independent components analysis," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 957–964, 2007.

[12] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[14] Emotiv, "Epoc neuroheadset." http://www.emotiv.com/apps/epoc/299/. [Webpage] Emotiv's own webpage of the Emotiv Epoc headset. Retrieved on May 14th 2014.

[15] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. LWW Doody's all reviewed collection, Lippincott Williams & Wilkins, 2005.

[16] NeuroSky, "Mindwave." http://store.neurosky.com/products/mindwave-1. [Webpage] Neurosky's own webpage of the NeuroSky Mindwave headset. Retrieved on May 14th 2014.

[17] BioSemi, "EEG headcaps." http://www.biosemi.com/headcap.htm. [Webpage] BioSemi's own webpage of their available EEG headcap. Retrieved on May 14th 2014.

[18] G. C. Ribas, "The cerebral sulci and gyri," *Neurosurgical Focus*, vol. 28, no. 2, p. E2, 2010.

[19] *Brodmann Atlas*. http://www.skiltopo.com/1/index.htm, Oct 2013. [Webpage] The Brodmann Atlas. Updated January 2013. Retrieved on May 23rd 2014.

[20] K. Brodmann, *Brodmann's: Localisation in the Cerebral Cortex*. Springer, 2007. Note: This is the 3rd translation of Brodmann K. "Vergleichende Lokalisationslehre der Grosshirnrinde". Leipzig: Johann Ambrosius Barth, 1909.

[21] H. H. Jasper, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology*, no. 10, pp. 371–375, 1958.

[22] TransCranialTechnologies, *10/20 System Positioning - Manual*. http://www.trans-cranial.com/local/manuals/10_20_pos_man_v1_0_pdf.pdf, 2012. [Manual] Retrieved on May 14th 2014.

[23] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clinical Neurophysiology*, vol. 112, no. 4, pp. 713–719, 2001.

[24] B. L. Miller and J. L. Cummings, *The human frontal lobes: Functions and disorders*. Guilford press, 2007.

[25] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related eeg/meg synchronization and desynchronization: basic principles," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[26] Y. Zhang, Y. Chen, S. L. Bressler, and M. Ding, "Response preparation and inhibition: the role of the cortical sensorimotor beta rhythm," *Neuroscience*, vol. 156, no. 1, pp. 238–246, 2008.

[27] J. A. Pineda, "The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing"," *Brain Research Reviews*, vol. 50, no. 1, pp. 57–68, 2005.

[28] S. J. Luck, *An introduction to the event-related potential technique*. MIT press Cambridge, MA:, 2005.

[29] D. Friedman, Y. M. Cycowicz, and H. Gaeta, "The novelty p3: an event-related brain potential (erp) sign of the brain's evaluation of novelty," *Neuroscience & Biobehavioral Reviews*, vol. 25, no. 4, pp. 355–373, 2001.

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

[31] G. F. Woodman, "A brief introduction to the use of event-related potentials in studies of perception and attention," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2031–2046, 2010.

[32] S. Sur and V. Sinha, "Event-related potential: An overview," *Industrial psychiatry journal*, vol. 18, no. 1, p. 70, 2009.

[33] G. A. Light and D. L. Braff, "Sensory gating deficits in schizophrenia: can we parse the effects of medication, nicotine use, and changes in clinical status?," *Clinical Neuroscience Research*, vol. 3, no. 1, pp. 47–54, 2003.
[34] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965.
[35] J. Polich, "Updating p300: an integrative theory of p3a and p3b," *Clinical neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.
[36] 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.
[37] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 4, pp. 441–446, 2000.
[38] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *Signal Processing Magazine, IEEE*, vol. 25, no. 1, pp. 41–56, 2008.
[39] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "EEG coherency: I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalography and clinical neurophysiology*, vol. 103, no. 5, pp. 499–515, 1997.
[40] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990.
[41] S. Lauritzen, *Graphical Models*. Clarendon Press, 1996.
[42] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.
[43] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
[44] K.-R. Müller, B. Blankertz, and G. Curio, "BCI Competition III - Data set IVa (motor imagery, small training sets)." http://www.bbci.de/competition/iii/desc_IVa.html, Dec. 2004. Retrieved on May 14th 2014.
[45] [budweiser] and Budweiser, "Budweiser Super Bowl XLVIII Commercial –"Puppy Love"." http://youtu.be/uQB7QRyF4p4, Jan 2014. [Video file] Retrieved on May 14th 2014.
[46] [dannotv], "John West Salmon "Bear Fight" ad." http://youtu.be/CVS1UfCfxlU, Nov 2000. [Video file]. This is the original video from 2000, but it was first uploaded to YouTube, on July 26th, 2006. Retrieved on May 14th 2014.
[47] [henryandaaron], H. Inglis, A. McCann, and L. Elliott, "Set Yourself Free." http://youtu.be/STHpMUYeznQ, Jan 2014. [Video file] Retrieved on May 14th 2014.
[48] [AutowayTire], "Snowy roads are scary." http://youtu.be/jGFWEoCGhi8, Nov 2013. [Video file] Retrieved on May 14th 2014.
[49] [solkampagne] and D. skin cancer research campaign, "Kære 15-årige mig." http://youtu.be/JdCDpYMf38M, Jan 2013. [Video file] Retrieved on May 14th 2014.
[50] [CopperCap], "GINGERS DO HAVE SOULS!!." http://youtu.be/EY39fkmqKBM, Jan 2010. [Video file]. The first 36 seconds only. Retrieved on May 14th 2014.
[51] [NZTransportAgency] and N. Z. T. Agency, "Speed ad - Mistakes." http://youtu.be/bvLaTupw-hk, Jan 2014. [Video file] Retrieved on May 14th 2014.