On Robust Parameter Estimation in Brain-Computer Interfacing

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

[†]Machine Learning Group, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[‡]Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
[♯]Advanced Telecommunication Research Institute International, 619-0288 Kyoto, Japan
[§]Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea
[¶]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

E-mail: wojciech.samek@hhi.fraunhofer.de, klaus-robert.mueller@tu-berlin.de

Abstract. *Objective*: The reliable estimation of parameters such as mean or covariance matrix from noisy and high-dimensional observations is a prerequisite for successful application of signal processing and machine learning algorithms in Brain-Computer Interfacing (BCI). This challenging task becomes significantly more difficult if the data set contains outliers, e.g., due to subject movements, eye blinks or loose electrodes, as they may heavily bias the estimation and the subsequent statistical analysis. Although various robust estimators have been developed to tackle the outlier problem, they ignore important structural information in the data and thus may not be optimal. Typical structural elements in BCI data are the *trials* consisting of a few hundred EEG samples and indicating the start and end of a task. *Approach*: This work discusses the parameter estimation problem in BCI and introduces a novel hierarchical view on robustness which naturally comprises different types of outlierness occurring in structured data. Furthermore, the class of minimum divergence estimators is reviewed and a robust mean and covariance estimator for structured data is derived and evaluated with simulations and on a benchmark data set. *Main results*: The results show that state-of-the-art BCI algorithms benefit from robustly estimated parameters. *Significance*: Since parameter estimation is an integral part of various machine learning algorithms, the presented techniques are applicable to many problems beyond BCI.

1. Introduction

Parameter estimation is one of the key tasks in statistics, signal processing and machine learning and has a substantial influence on the performance of algorithms in these fields. The robustness of an estimator is of central importance as data are not only noisy but often also contaminated by outliers. Although the sample mean and covariance estimators are known to be vulnerable to outliers [1], they are integral part of many popular algorithms. Various robust alternatives (e.g., [1, 2, 3, 4]) have been developed to improve parameter estimation in the presence of outliers, however, these works do not consider the application to *structured data*, i.e., data that can be divided into meaningful units consisting of groups of samples.

One example of structured data are pooled data sets. Here samples coming from different sources (e.g., recordings sites, subjects) make up meaningful units which may largely vary in quality. In this case not only individual samples may be regarded as outliers, but all data from a specific recording site or subject may need to be discarded if this data source systematically biases the parameter estimation.

Another example of structured data are EEG recordings from a Brain-Computer Interfacing (BCI) (e.g. [5, 6]) experiment. In these experiments subjects are asked to perform certain tasks such as motor imagination over a limited period of time (i.e., a *trial*) and the mental state present in the EEG is decoded in real-time. Each trial consists of a few hundred EEG samples and is a meaningful unit in the data, because it indicates the start and end of a task. This structure naturally leads to a multi-scale definition of outlierness which is illustrated in Figure 1. Two types of outliers can be identified in this example. First, the sample at the beginning of trial 1 has a significantly larger value than all other samples in the data, thus it can be easily identified as outlier sample. Second, also trial 4 can be regarded as outlier, because it lacks the high-variance response in the middle of the trial which is present in all other trials. However, without the structural information one can not identify this trial as outlier, because its samples are not different from the other samples in the data set. This example shows that by grouping samples into larger units one can identify outliers which are fundamentally different from individual outlier samples. Such outlier trials can only be identified after aggregating the information within each trial and comparing the trial distributions (or resp. summary statistics). As shown later sample-based estimators such as [1, 2, 3, 4] are not robust‡ to these second-level outliers.

A proper and robust analysis of EEG data in general and BCI data in particular is a necessary prerequisite from the data analysis side for moving experiments out of a lab environment for general tasks of man-machine interaction [7, 8, 9, 10, 11, 12, 13, 14]. While this review is driven by the general motivation to set the scenes for BCI or more general neuroscience experiments beyond the lab, we will focus here on a selected subset of mathematical aspects of this challenge, hoping that it will be of use for the community. Specifically, we will focus on the class of minimum divergence parameter estimators and will derive novel mean and covariance matrix estimators which are tailored to structured data, demonstrating that a multi-level treatment of outliers is important when estimating parameters in structured data. Naturally, we will draw from own work, hoping that we have appropriately maintained the balance to the literature, providing the necessary pointers for further reading.

The remaining paper is organized as follows: At first, we briefly review robust methods in BCI. Then we formulate parameter estimation as divergence minimization problem and show that using particular classes of divergences results in robust estimates. In Section 4, we present the minimum β-divergence estimators for the Gaussian and Wishart distribution model and discuss the application to structured data. In Section 5 we investigate the advantages and limitations of our estimators using simulations and discuss their relations to state-of-the-art estimators. We evaluate and compare the estimators on a large BCI data set with 80 subjects.

‡ Note that Riemannian geometry-based estimators tend to be more robust, but are not discussed in this paper.

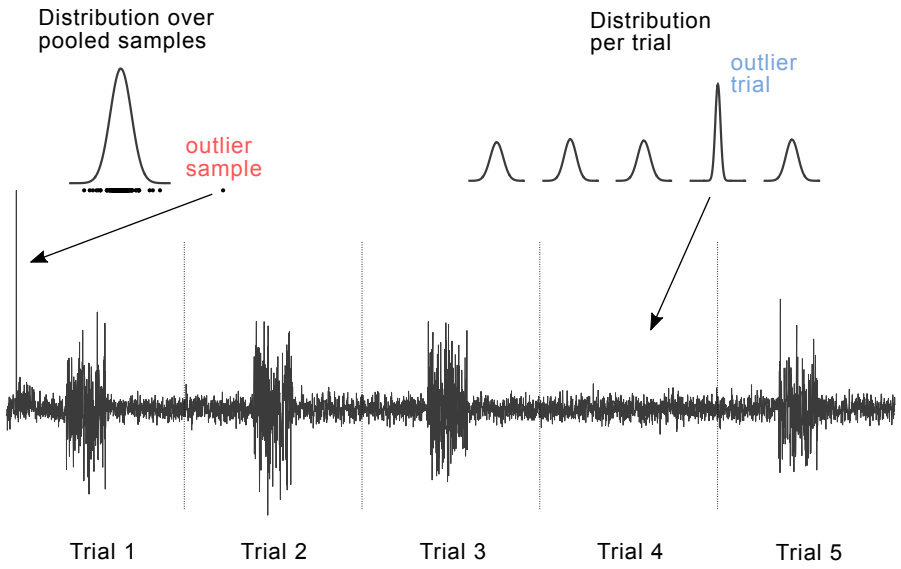

Figure 1: Illustrative example of different types of outlierness in structured data. The outlier sample significantly deviates from all other samples in the data, thus can be identified even when ignoring the trial structure. The outlier trial can only be identified by using summary statistics as its samples do not deviate from the majority of the samples in the data.

Finally, Section 6 concludes this paper with a discussion and an outlook on future work.

2. Robust Methods in Brain-Computer Interfacing

EEG signals are often affected by electrical sources which are unrelated to brain activity, but produce very larger potential differences. These *artifacts* can be of physiological origin, e.g., eye blinks or muscle contractions, or be due to technical reasons such as loose electrodes or power grid noise. In either case artifacts contaminate the recorded EEG signals and heavily bias the estimation of parameters at all stages of the BCI processing pipeline. Since poorly estimated parameters do not well represent the underlying neural processes, artifacts negatively affect BCI performance. This section provides an overview of recent techniques which tackle the outlier problem in BCI.

2.1. Artifact Removal

Removing artifactual components from the recorded data is a common strategy to robustify BCI systems. Most of the artifact removal methods exploit the assumption that artifactual and neural activity are independent. These algorithms first decompose the EEG signal into independent source components (ICs) and then discard the artifactual ICs. Especially, ICs related to ocular artifacts can be easily identified with this approach. Other artifact types produce ICs that are often less consistent, but can be distinguished from neural activity. Various methods have been proposed for automatic or semi-automatic identification of ICs representing EEG artifacts [15, 16, 17, 18, 19, 20, 21]. An automatic and adaptive artifact

detection method based on Riemannian geometry has been proposed in [22]. Surveys on artifact removal can be found in [23, 24] and the effects of different artifact types on motor imagery BCI are investigated in [25, 26].

2.2. Trial and Channel Selection

An alternative to discarding artifactual ICs is to focus on the identification and removal of artifactual trials and channels. This approach is advantageous when individual channels are unreliable (e.g., loose electrode) or when only few trials are contaminated by artifacts (e.g., by subject movements). The authors of [27] proposed a sparsity-aware method to eliminate low-quality trials from a BCI data set. Other researchers used M-estimators [28, 29], robust divergences [30] and Riemannian Geometry [31] for robust estimation of covariance matrices. Since these methods implicitly perform some kind of trial weighting, they also reduce the impact of artifactual trials. The identification of reliable and informative channels is a topic which has received a lot of attention in the BCI community. Various techniques have been suggested to identify the optimal channel configuration [32, 33, 34, 35]. Sparsity enforcing methods (e.g., [36, 37]) have also been successfully applied in this context.

2.3. Robust Spatial Filtering

Common Spatial Patterns (CSP) [38, 39] is a popular algorithm for optimizing spatial filters in motor imaginary BCIs. Since the algorithm relies on class covariance matrices which have to be estimated on the calibration data, it can be severely affected by artifacts. Trial and channel selection can significantly improve the performance of CSP in the presence of outliers. Another approach to increase robustness, especially in small-sample settings, is based on regularization of the covariance matrices [40, 41, 42, 43, 44]. Researchers have also formulated CSP as a maxmin optimization problem [45], proposed a CSP variant based on Student-t distribution [46], applied trial pruning [47] and used generalized norms [48] to robustify the algorithm. Other work increases robustness by adding regularization to the CSP objective [44, 49, 50] or by formulating the algorithm as divergence maximization problem [51, 52] and utilizing the robustness property of particular divergences.

2.4. Nonstationarity & Robust Classification

A significant fraction of errors in BCI can be attributed to the nonstationarity of the EEG [53, 54, 55] which leads to a changing feature distributions and compounds the classification problem. Various adaptation strategies (e.g., [56, 57]) have been proposed to cope with nonstationarity in BCI. Researchers have also tackled the nonstationarity problem by regularizing the spatial filters [50, 58] towards stationarity, i.e., by trading-off the discriminativity and stationarity of the extracted features, and by applying an importance-weighted covariance estimator [59]. Other approaches project the signals into a stationary subspace prior to spatial filter computation by using the stationary subspace analysis (SSA) algorithm [60, 61], it's geometry-aware extension [62] or a methods based on principal

components [63]. The authors of [64, 43] jointly perform feature extraction and classification, others [65, 66, 67] directly perform classification on the manifold of covariance matrices. Shrinkage is a common approach to robustify the covariance matrix estimation and the Linear Discriminant Analysis (LDA) classifier [68, 69, 70]. Recent works apply deep neural networks [71, 72, 73, 74] for classification of EEG motor imagery signals. These powerful nonlinear methods have recently become interpretable [75, 74, 76], which is an important aspect in BCI research because it allows to make sure that the model relies on neurophysiological features.

3. Parameter Estimation based on Divergence Minimization

In the following we will lay out the basis for robust estimation methods that implements a further direction beyond nonstationarity and regularization, essentially providing the basis for all discussed algorithmic directions: if the parameters (e.g., mean, covariance matrix) are estimated better, many of the variants discussed above can improve.

3.1. Minimum Divergence Estimator

In parameter estimation a common assumption is that the observations $\mathcal{D} = \{\mathbf{x}_i : i = 1 \dots n\}$ come from an underlying statistical model q with unknown parameter θ^*. A standard procedure to estimate this parameter is to maximize the log-likelihood function $\mathcal{L}(\theta \,|\, \mathcal{D})$ of the parameter given observations

$$\mathcal{L}(\theta \,|\, \mathcal{D}) \;=\; \log\left(\prod_{i=1}^{n} q_\theta(\mathbf{x}_i)\right) \;=\; \sum_{i=1}^{n} \ell(\mathbf{x}_i;\, \theta). \tag{1}$$

In an alternative formulation of parameter estimation, also known as the minimum disparity estimation or minimum divergence estimation (MDE) [4], one aims to minimize the *divergence* $D(p \,||\, q_\theta)$ between the empirical distribution p of the observations and the model distribution q_θ. Note that a divergence [77] is a non-negative measure from information geometry (c.f. [78, 79]) used to quantify the disparity between two probability distributions. A divergence is in general asymmetric and has value zero iff the distributions coincide. When using a specific divergence, namely the Kullback-Leibler divergence§, the minimum divergence estimator coincides with the maximum likelihood estimator (MLE) [4], i.e.,

$$\hat{\theta} \;=\; argmax_\theta \mathcal{L}(\theta \,|\, \mathcal{D}) \;=\; argmin_\theta D_{kl}(p \,||\, q_\theta). \tag{2}$$

The formulation of the parameter estimation problem in terms of divergence minimization has one important advantage, namely it allows to impose additional properties such as robustness on the estimator by using specific divergences [80, 52]. For instance, in the case of β-divergence (with $\beta > 0$)

$$D_\beta\left(p \,||\, q\right) = \frac{1}{\beta}\int (p^\beta - q^\beta) p dx - \frac{1}{\beta+1}\int (p^{\beta+1} - q^{\beta+1}) dx$$

§ The Kullback-Leibler divergence between distribution p and q is defined as $D_{kl}\left(p(x) \,||\, q(x)\right) = \int p(x) \log \frac{p(x)}{q(x)} dx$

one can prove [81] increased robustness of the estimator, i.e., reduced vulnerability to outliers in the data. Note that for $\beta \longrightarrow 0$ the β-divergence coincides with the KL-divergence and loses its robustness property. Beta divergence is an instance of a general class of divergences termed Bregmann divergences [82] and has many interesting properties (see [77] for more details).

Minimum divergence estimators have also been applied with other measures of disparity such as Hellinger distance [83], power divergences [84] or γ-divergence [85]. Note that although we limit our discussion to β-divergence in this paper, the proposed ideas are applicable to all other measures of disparity in probability distributions.

3.2. Iterative Algorithm

The minimization of a divergence $D(p \mid\mid q_\theta)$ between the empirical and model distribution with respect to parameter θ is in general a non-convex problem and can be solved (up to local optimality) iteratively by using a fixed point algorithm. For a specific class of divergences, termed Ψ-divergences, the estimating equation reduces to (see [81, 86])

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\ell(\mathbf{x}_i;\,\theta^{(k)}))S(\mathbf{x}_i;\,\theta^{(k+1)}) = E[\psi(\ell(\mathbf{x};\,\theta^{(k+1)}))S(\mathbf{x};\,\theta^{(k+1)})], \tag{3}$$

where $\psi(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\Psi(\mathbf{x})$, $S(\mathbf{x};\,\theta) = \frac{\partial}{\partial \theta}\ell(\mathbf{x};\,\theta)$ and $E[\,\cdot\,]$ denotes the expectation over the whole input space. Ψ is assumed to be monotonically increasing, convex and differentiable scalar function. Note that Eq. (3) is related to the update equation of M-estimators [1] and the parameter $\theta^{(k+1)}$ is determined iteratively as solution of Eq. (3) starting from an initial value $\theta^{(0)}$. For the function

$$\Psi_\beta(z) = \frac{\exp(\beta z)}{\beta} \tag{4}$$

the fix point equation (3) minimizes the β-divergence between the empirical and model distribution as Ψ-divergence reduces to β-divergence (see [81] for more details).

The steps of the minimum divergence estimator (MDE) can be summarized as follows. First, we initialize the parameter $\theta^{(0)}$ by either a random value or the value obtained when applying the MLE. Then, we use Eq. (3) to iteratively compute the parameter $\theta^{(k+1)}$. Note that fix point equation (3) will be different for different underlying models q_θ and Ψ-functions. In this work we use the Gaussian and Wishart model for q_θ and the Ψ-function defined in Eq. (4). We stop the iteration after k_{max} steps or when a stopping criterion, e.g., relative change in estimate below a threshold, has been reached. Note that this algorithm converges to a local optimum [81], thus several restarts may be required in practice. Among the several restarts one has to select the solution, e.g., by preferring parameters with a specific property such as minimum determinant, by using cross-validation, or by applying stability related selection criteria. The joint estimation of several parameters such as mean and covariance matrix can be performed easily by keeping one parameter fix at a time.

3.3. Robustness Property

The goal of robust estimators is to reliably estimate the parameter of the underlying model when data is heavily contaminated. Formally, we aim to estimate the density $q_{\hat{\theta}}(\mathbf{x})$ with $\hat{\theta} \approx \theta^*$ while observing data generated from

$$p_{\theta^*}(\mathbf{x}) = (1-\eta)q_{\theta^*}(\mathbf{x}) + \eta\epsilon(\mathbf{x}) \tag{5}$$

where $\epsilon(\mathbf{x})$ is the contamination probability density and $q_{\theta^*}(\mathbf{x})$ is the true probability density. A common assumption is that an outlier sample $\mathbf{x}_{out}$ has a very low probability to be generated by the true model, i.e., $q_{\theta^*}(\mathbf{x}_{out}) \approx 0$. This implies that the log-likelihood term becomes extremely small for this sample, i.e., $\ell(\mathbf{x}_{out}; \theta^*) \approx -\infty$, consequently the maximum likelihood method will not estimate the true parameter θ^*, but a parameter $\hat{\theta}$ with a significantly larger log-likelihood term $\ell(\mathbf{x}_{out}; \theta^*) << \ell(\mathbf{x}_{out}; \hat{\theta})$ for the outlier sample. Instead of ignoring the outlier $\mathbf{x}_{out}$, the maximum likelihood estimator tries to compensate for its very small log-likelihood term. Thus, the outlier introduces a huge bias in the estimation.

The authors of [81] have shown that minimizing β-divergence is equivalent (up to a normalizing constant) to maximizing the Ψ_β-likelihood, i.e., $\sum_{i=1}^{n} \Psi_\beta(\ell(\mathbf{x}_i; \theta))$. Since Ψ_β is an exponential function (for $\beta > 0$), it reduces the influence of outliers to zero, i.e., $\Psi_\beta(\ell(\mathbf{x}_{out}; \theta^*)) \approx 0$. This property makes the minimum β-divergence estimator very robust, because extreme unlikely samples are being effectively discarded. For more formal discussion of robustness we refer to [81, 4, 1, 87].

4. Robust Parameter Estimators for Structured Data

Samples in a data set are often naturally grouped into meaningful units. This type of structured data can be analyzed in two ways, namely with respect to the individual samples or the groups. Consequently, we can define robustness with respect to both levels of analysis. Figure 2 visualizes the difference between sample-level and group-level estimation. In the former case parameters are estimated directly from the samples, whereas in the latter approach, parameters are estimated from summary statistics that have been computed from the groups. In the following we present robust mean and covariance matrix estimators for both the sample- and the group-level view. We assume that all samples except outliers are generated from a Gaussian distribution.

4.1. Sample-level Estimators

When estimating parameters on sample-level we discard structural information in the data and assume that all observations come from a Gaussian distribution q with unknown parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$. For minimum β-divergence estimator in such data generation model, the iteration formula in Eq. (3) reduces to

$$\boldsymbol{\mu}^{(k+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n} \psi_\beta(\ell(\mathbf{x}_i;\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}))\mathbf{x}_i}{\frac{1}{n}\sum_{i=1}^{n} \psi_\beta(\ell(\mathbf{x}_i;\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}))} \tag{6}$$

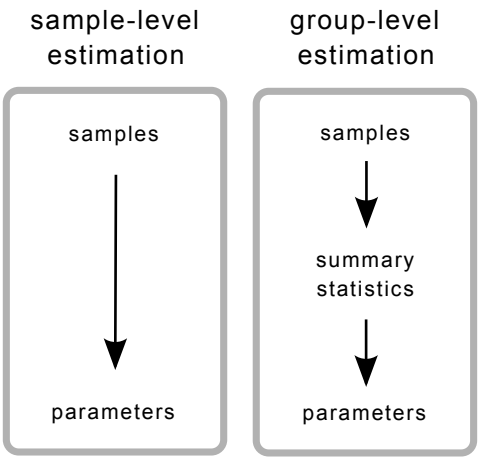

Figure 2: Two ways of robustly estimating parameters from data: (i) direct estimation of parameters from the samples and (ii) estimation of parameters from summary statistics that have been computed from the samples and capture the structural information in the data.

$$\boldsymbol{\Sigma}^{(k+1)} \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}))(\mathbf{x}_i-\boldsymbol{\mu}^{(k+1)})(\mathbf{x}_i-\boldsymbol{\mu}^{(k+1)})^\top}{\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}))\;-\;\beta/(\beta+1)^{D/2+1}} \tag{7}$$

Note that $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ stand for the parameter estimates in kth iteration step and

$$\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}))\;=\;e^{-\frac{1}{2}\beta(\mathbf{x}_i-\boldsymbol{\mu}^{(k)})^\top(\boldsymbol{\Sigma}^{(k)})^{-1}(\mathbf{x}_i-\boldsymbol{\mu}^{(k)})} \tag{8}$$

is a factor downweighting the influence of outlier samples $\mathbf{x}_i$. From the formula we can see that if the sample $\mathbf{x}_i$ is an outlier, i.e., it is very unlikely that it has been generated by a Gaussian with parameters $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$, then its influence on the update of the parameters is very small due to vanishing $\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}))$. Since $\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}))$ is a monotonically decreasing function (for $\beta > 0$), it limits the influence of extreme outliers. Note that for $\beta \to 0$ these estimators reduce to the maximum likelihood estimators $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top$, i.e., all samples have uniform weight $\psi_\beta(\ell(\mathbf{x}_i;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)})) = 1$. As mentioned before we can estimate both parameters simultaneously by alternating equations (6) and (7). Note that this estimator has been proposed in [81]. We refer to it as Gaussian-MDE or $\mathcal{G}$-MDE.

The middle panel of Figure 3 visualizes the downweighting effect of $\mathcal{G}$-MDE. We sample 10 trials with 250 samples each (circles) from a distribution with (almost) the same covariance matrix and 1 trial from a distribution with completely different covariance matrix (crosses). The thick black line represents the estimated covariance matrix at each iteration of the algorithm and the color represents the weight ψ_β (see Eq. (8)) assigned to each sample. Note that red color stands for small weights and gray color represents weights close to 1. One can see that as the outlier samples are more and more downweighted the estimated covariance matrix captures the structure of the clean data. The speed of this downweighting depends on the initialization value and the β parameter. The top panel of Figure 3 visualizes the estimation error (log scale), i.e., Frobenius norm between estimated and true covariance matrix, at different iterations. It should be noted that some of the samples generated by the

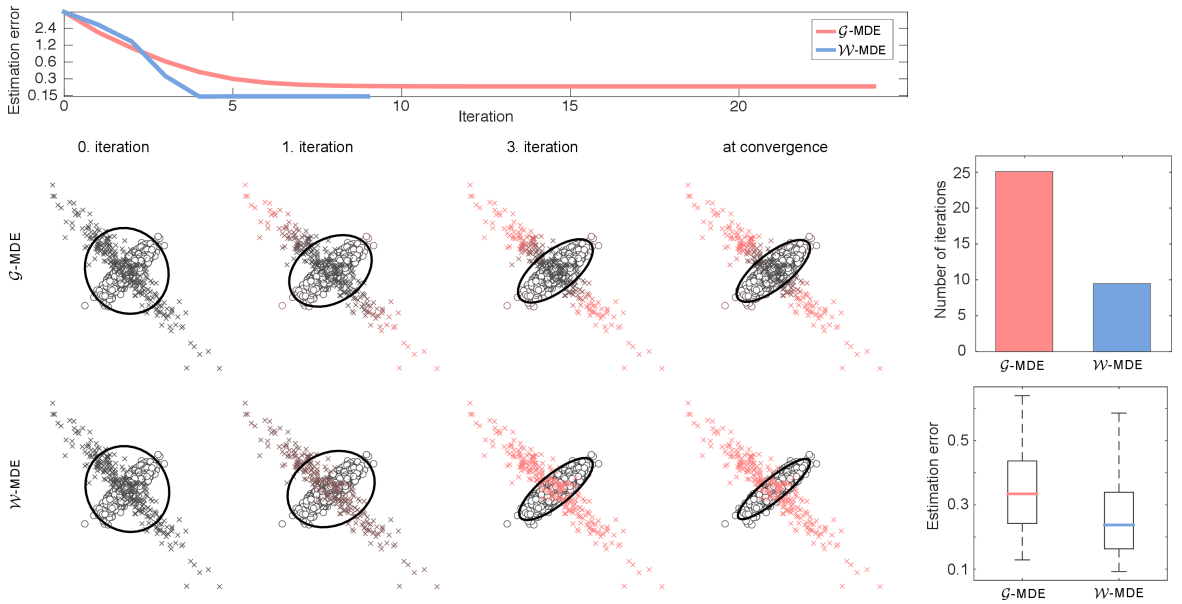

Figure 3: Visualization of the downweighting effect on sample- and trial-level. Top row: Estimation error (log scale), i.e., Frobenius norm between estimated and true covariance matrix, at different iterations. Middle row: The 2500 circles represent samples coming from 10 trials of the model distribution whereas the 250 crosses are samples from an outlier distribution. The ellipse in the 0th iteration represents the sample covariance matrix which is highly affected by the outliers. After several iterations the outlier samples are downweighted (red color), thus the estimated covariance matrix captures the true data distribution. Some of the crosses (i.e., samples generated by the outlier trial) still have gray color and are not downweighted. Bottom row: The same example for the Wishart model. In 3rd iteration the estimated covariance matrix (thick black ellipse) approaches the true covariance matrix as the outlier trial is being effectively discarded (red color). Top right: Average number of iterations until convergence (100 repetitions). Bottom right: Estimation error, i.e., Frobenius norm between estimated and true covariance matrix, at convergence (100 repetitions).

outlier trial (crosses) are not downweighted by $\mathcal{G}$-MDE at convergence point‖, i.e., still have gray color. In our experiment the $\mathcal{G}$-MDE with $\beta = 0.1$ requires 25 iterations on average (100 repetition) to reach this point.

4.2. Group-level Estimators

In structured data sets it may be advantageous to downweight groups of samples, e.g., outliers trials, rather than individual samples. To this end, we first compute summary statistics for each group, and treat them as second-level samples. Summary statistics should be chosen based on the assumption on the distribution of each trial. Since we assume that non-outliers are i.i.d. Gaussian, a natural choice is the sufficient statistics for Gaussian, i.e., sample average and

‖ We assume that the estimator *converges* if $\frac{||\Sigma^{(k+1)}-\Sigma^{(k)}||}{||\Sigma^{(k)}||} < 10^{-8}$.

sample covariance. To obtain group-level robust estimators, we apply Eq. (3) to each of the summary statistics. The sample mean again follows a Gaussian distribution, and therefore, we can directly apply Eq. (6) to estimate the mean parameter from the second-level samples

$$\boldsymbol{\mu}^{(k+1)} \;=\; \frac{\frac{1}{m}\sum_{j=1}^{m}\psi_\beta\big(\ell(\mathbf{z}_j;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)})\big)\mathbf{z}_j}{\frac{1}{m}\sum_{j=1}^{m}\psi_\beta\big(\ell(\mathbf{z}_j;\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)})\big)} \tag{9}$$

where m is the number of groups and $\mathbf{z}_j$ is the average of all samples in group j.

On the other hand, we cannot use Eq. (7) to estimate the covariance parameter from the second-level sample covariances. In the following we derive an update rule for the group-level estimator for the covariance parameter. It is known that the sample covariance follows the Wishart distribution [88], defined as

$$q(\mathbf{S};\boldsymbol{\Sigma},\nu) = \frac{1}{2^{\frac{\nu D}{2}}|\boldsymbol{\Sigma}|^{\frac{\nu}{2}}\Gamma_D\left(\frac{\nu}{2}\right)}|\mathbf{S}|^{\frac{\nu-D-1}{2}}e^{-\mathrm{tr}\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)} \tag{10}$$

where

$$\mathbf{S} \;=\; \sum_{t=1}^{N}(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})^\top \tag{11}$$

is the scatter matrix¶ and Γ_D is the multivariate gamma function defined as

$$\Gamma_D\left(\frac{\nu}{2}\right) \;=\; \pi^{\frac{D(D-1)}{4}}\prod_{j=1}^{D}\Gamma\left[\frac{\nu}{2}\;+\;\frac{(1-j)}{2}\right] \tag{12}$$

with $\Gamma[t] \;=\; \int_0^\infty y^{t-1}e^{-y}dy$. Thus, in order to robustly (wrt group-level outliers) estimate a covariance matrix we compute the scatter matrices $\left\{\mathbf{S}_j\in\mathbb{R}^{D\times D}: j=1\ldots m\right\}$ for each group and treat them as samples of an unknown Wishart distribution with parameters $\boldsymbol{\Sigma}$ and ν. Note that $\boldsymbol{\Sigma}$ denotes the true covariance matrix which we want to estimate and ν (under the assumptions that the samples are i.i.d.) equals the number N (or $N-1$ if mean is subtracted) of samples within a group (which we assume to be fix). The maximum likelihood estimator for the Wishart distribution is

$$\hat{\boldsymbol{\Sigma}} \;=\; \frac{1}{\nu m}\sum_{j=1}^{m}\mathbf{S}_j, \tag{13}$$

or equivalently it is the average covariance matrix. We robustly estimate a covariance matrix $\hat{\boldsymbol{\Sigma}}$ from the scatter matrices $\mathbf{S}_j$ of trials $j=1\ldots m$ by minimizing β-divergence using the following iterative formula

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{j=1}^{m}\psi_\beta\left(\ell\left(\mathbf{S}_j^{(k)};\,\boldsymbol{\Sigma}^{(k)},\nu\right)\right)\mathbf{S}_j^{(k+1)}}{\nu\sum_{j=1}^{m}\psi_\beta\left(\ell\left(\mathbf{S}_j^{(k)};\,\boldsymbol{\Sigma}^{(k)},\nu\right)\right)\;-\;\gamma|\boldsymbol{\Sigma}^{(k)}|^{\frac{(\nu-D-1)\beta}{2}}} \tag{14}$$

where

$$\psi_\beta\left(\ell(\mathbf{S}_j^{(k)};\,\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)},\nu)\right) = |\mathbf{S}_j^{(k)}|^{\frac{(\nu-D-1)\beta}{2}}e^{-\mathrm{tr}\left(\frac{\beta}{2}(\boldsymbol{\Sigma}^{(k)})^{-1}\mathbf{S}_j^{(k)}\right)} \tag{15}$$

¶ Up to a constant the scatter and covariance matrices are equivalent.

is a factor downweighting the influence of outlier trials and $\gamma = \frac{n\beta(D+1)\Gamma_D(\gamma_0)}{2^{\frac{\nu D}{2}}\Gamma_D\left(\frac{\nu}{2}\right)(\beta+1)}\left(\frac{2}{\beta+1}\right)^{D\gamma_0}$ and $\gamma_0 = \frac{\nu(\beta+1)}{2} - \frac{(D+1)\beta}{2}$ are constants. Note that we write $\mathbf{S}_j^{(k)}$ and $\mathbf{S}_j^{(k+1)}$ to indicate that these scatter matrices depend on $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\mu}^{(k+1)}$, respectively. For $\beta \to 0$ this estimator gives the maximum likelihood solution in Eq. (13). As before we may simultaneously estimate the mean and covariance matrix parameter by alternating Eq. (9) (which depends on $\boldsymbol{\Sigma}^{(k)}$) and Eq. (14) (which depends on $\boldsymbol{\mu}^{(k+1)}$ through the estimation of the scatter matrices $\mathbf{S}_j^{(k+1)}$). We refer to this novel minimum divergence estimator as Wishart-MDE or $\mathcal{W}$-MDE. The derivation of the estimator can be found in the appendix.

The lower panel of Figure 3 visualizes the downweighting effect of $\mathcal{W}$-MDE and color encodes the value of ψ_β from Eq. (15). As before our algorithm arrives at a good estimate of the true covariance matrix by downweighting the outlier trial. In contrast to $\mathcal{G}$-MDE the proposed estimator downweights all samples from the outlier trial (crosses). Furthermore, it needs only 9 iterations on average with $\beta = 0.001$ ($\mathcal{G}$-MDE requires 25 iterations) to reach the convergence point, i.e., the point where the relative change in Frobenius norm between the two consecutively estimated covariance matrices is less than 10^{-8}. At the point of convergence the covariance matrix estimated by $\mathcal{W}$-MDE is significantly closer to the true covariance matrix (in terms of Frobenius norm) than the covariance estimate provided by $\mathcal{G}$-MDE (see bottom right panel in Figure 3). We recomputed the results for a range of β values, but did not observe any qualitative change; in all cases $\mathcal{W}$-MDE was the more efficient and more precise estimator.

Finally, we would like to note that other approaches to robust group level parameter estimation exist. For instance, instead of the formula in Eq. 9 one could also rely on trimming and use the median in order to compute a robust average. Also the covariance matrices could be naturally fitted into a Karcher averaging if a covariance matrix is estimated for each epoch and the Riemannian geometry is concerned. An empirical evaluation of such averaging has been performed in [89].

4.3. Combined Estimator

The $\mathcal{W}$-MDE uses scatter matrices $\mathbf{S}$ which are computed for each trial (see Eq. 11). Of course outlier samples may negatively affect the estimation of these matrices. A robust estimator which considers both, sample and trial outliers, can be obtained by applying $\mathcal{G}$-MDE for computing of the scatter matrices for each trial and $\mathcal{W}$-MDE for computing the final covariance matrix. We will refer to such estimator as $\mathcal{W}\mathcal{G}$-MDE. For simplicity we use the same β parameter for both estimation steps ($\mathcal{G}$-MDE and $\mathcal{W}$-MDE), however, a separate parametrization of both steps may be beneficial in practice.

4.4. Use in Practice

In the discussion so far we have assumed that the parameter of the Wishart distribution ν equals the number of samples used to estimate the scatter matrix (sample size). However, this assumption only holds if the samples are i.i.d. which is often not the case in real-world data

sets, e.g., in EEG recordings. If samples are correlated then the parameter ν should be set to the *effective sample size* [90, 91] which is smaller than the number of samples within a group.

Another practical recommendation concerns the computation of the ratio of Gamma functions in γ. A naive computation of this ratio leads to numerical problems because of the very high values of the Gamma function. A common trick which is applied to stabilize the computation is to log-transform the ratio, compute the difference of the log terms and transform back via an exponential map.

Depending on how the scatter matrices and mean parameter are computed, there exist multiple variants of the $\mathcal{W}$-MDE. If data is sampled from a zero mean distribution, then there is no need to perform mean estimation at all. In this case $\boldsymbol{\mu}$ should be set to 0 and Eq. (14) should be applied iteratively. If the mean parameter is assumed to be same across all groups, then we recommend to compute the scatter matrices of groups j as describe above, namely $\mathbf{S}_j = \sum_{t=1}^{N}(\mathbf{x}_t^j - \boldsymbol{\mu})(\mathbf{x}_t^j - \boldsymbol{\mu})^\top$. If different groups have different means, then one should compute the scatter matrices of group j as $\mathbf{S}_j = \sum_{t=1}^{N}(\mathbf{x}_t^j - \mathbf{z}_j)(\mathbf{x}_t^j - \mathbf{z}_j)^\top$ where $\mathbf{z}_j$ denotes the group average. In both cases we may alternate between Eq. (9) and Eq. (14) or only apply the latter formula iteratively. The choice of the right $\mathcal{W}$-MDE variant largely depends on the problem.

5. Experimental Evaluation

In the following we compare the performance of four parameter estimators, namely sample estimator (SE), minimum covariance determinant estimator (MCDE)$^+$ [2], $\mathcal{G}$-MDE and $\mathcal{W}$-MDE, using simulations and evaluate these estimators on two motor imaginary BCI datasets. Our results show that state-of-the-art BCI algorithms largely benefit from robustly estimated parameters and that, in some cases, group-level estimation is clearly advantageous.

5.1. Simulations

5.1.1. Single Outlier Simulation In the first experiment we evaluate the downweighting effect of the three robust estimators, MCDE, $\mathcal{G}$-MDE and $\mathcal{W}$-MDE, when adding one outlier trial to a data set consisting of 20 clean trials with 100 samples each. Note that the clean data is generated from the same distribution as the circles in Figure 3. After application of the three robust estimators we display the ratio

$$\rho(\alpha, \sigma) = \frac{\sum \psi_\beta(\ell(\mathbf{x}_{out}; \hat{\boldsymbol{\Sigma}}_{\alpha,\sigma}))}{\sum \psi_\beta(\ell(\mathbf{x}_{clean}; \hat{\boldsymbol{\Sigma}}_{\alpha,\sigma}))} \tag{16}$$

i.e., the weights assigned to the samples $\mathbf{x}_{out}$ of the outlier trial relative to the weights assigned to the samples of the clean trials $\mathbf{x}_{clean}$. Note that for MCDE these weights are either 1 (if the sample is among the selected h samples) or 0 (otherwise). For $\mathcal{W}$-MDE we show the corresponding trial weights. Note that a small ρ value means that the outlier trial has been effectively discarded. In the experiment we add outliers by (i) rotating and (ii) uniformly

$^+$ MCDE finds $h \leq n$ samples which have a covariance matrix with the lowest possible determinant, thus MCDE resists $(n - h)$ outliers. We refer the reader to [92] for a critical discussion on MCDE.

scaling the true covariance matrix. Figure 4 visualizes $\rho(\alpha, \sigma)$ for a range of rotation α and scaling σ parameters. Note that black color represents small and white color large ρ values. The optimal robust estimator is shown in the left panel. It assigns $\rho = 0$ (discards outlier trial) to all scale and angle combinations except the true distribution (scale 1 and angle 0). One can see that all three estimators downweight the outlier trial if its covariance matrix is much larger than the true covariance matrix or if it is rotated. However, only the estimator using the Wishart model identifies the trial with significantly smaller variance (scale < 1) as outlier. Both MCDE and $\mathcal{G}$-MDE do not identify this trial (or more precisely its samples) as outliers even when the covariance matrix is largely rotated. Since the variance of samples coming from this outlier trial is significantly smaller than the variance of the clean data, the samples are within the range of the clean data (even when the outlier trial has a rotated covariance matrix), thus they are not identified as outliers on the sample-level. Only the trial information makes these samples distinguishable from the clean data.

This effect of *lying within the region of clean data* is also responsible for the fluctuations (smearing effect) in the maps produced with MCDE and $\mathcal{G}$-MDE. Since some proportion of samples (coming from the outlier trial) will always lie in the range of clean data, their ψ_β values will be relatively high as they are no outliers according to the sample-level view. Therefore the corresponding ρ value will be significantly larger than zero. This effect can be also seen in Figure 3 where the samples which come from the outlier trial (crosses) but lie within the range of the clean data stay gray, i.e., are not downweighted by the estimator. On the other hand when applying $\mathcal{W}$-MDE the outlier trial will have very small ψ_β value irrespectively of whether its samples lie within the range of clean data or not. Thus, the corresponding ρ value will be close to zero. This results can be observed for a range of β parameters.

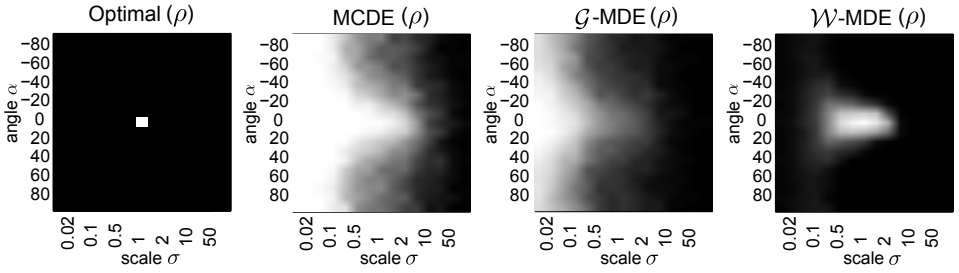

Figure 4: Comparison of three robust estimators with respect to their ability to discard outlier trials. The clean data consists of 20 trials with 100 samples per trial coming from a normal distribution with (almost) fix covariance matrix. We add one outlier trial to these data. The outlier samples come from a distribution with a covariance matrix which is the scaled and/or rotated version of the true underlying covariance matrix. The maps represent the ρ value defined in Eq. (16). Black color stands for small ρ values, white color represents large ρ values. One can see that all estimators downweight the outliers for large scales and angles, however, only the Wishart model identifies the outlier trial when small scales are applied. Furthermore, the Wishart estimator provides much cleaner ρ values.

5.1.2. Multiple Outliers Simulation In a second experiment we sample 50 trials with 100 samples per trial from a 10 dimensional zero-mean normal distribution. The samples come either from a clean distribution or from an outlier distribution which is a scaled and rotated version of the former. We investigate two ways of adding outliers (i) we add outliers sample by sample (sample-level), i.e., for each sample we throw the dice whether it comes from the clean or outlier distribution, or (ii) we sample a whole trial from the outlier distribution (group-level), i.e., the decision whether a trial comes from the clean or outlier distribution affects all samples of the trial. Furthermore, we use two different scales for the outlier distribution, namely scale $\sigma = 0.01$ and scale $\sigma = 100$. The clean covariance matrix is $\Sigma_{clean} = V_{clean} D_{clean} V_{clean}^\top$ with V being a random rotation matrix and D being a diagonal matrix sampled from the uniform distribution. The covariance matrix of the outlier distribution $\Sigma_{outlier} = \sigma V_{outlier} D_{outlier} V_{outlier}^\top$ has the same form but it is scaled by σ. Note that $\Sigma_{outlier}$ is sampled independently of Σ_{clean} and is not fix across trials, i.e., outliers from different trials may come from different outlier distributions.

Figure 5 displays the results for the different estimators. The y-axis shows the log scaled error measure which is the distance (Frobenius norm) to the true covariance matrix Σ_{clean}. The different lines represent the median error over 50 repetitions when selecting the best parameter (among several parameters which have been tested) for each method, repetition and experimental setting. The first row shows the results for the small scale experiment. One can see that MCDE performs slightly worse than $\mathcal{G}$-MDE and the sample estimator. Since MCDE favours covariance matrices with small determinant it naturally focuses on the outlier samples (coming from trials with small variance), therefore the estimate is worse than when applying the other estimators. The heuristic used by MCDE fails as the small variance samples are the outlier samples in this example. $\mathcal{G}$-MDE slightly outperforms the $\mathcal{W}$-MDE in the sample-level experiment but the difference to SE is not large because the small variance samples do not affect the sample covariance estimator very much. For the group-level experiment $\mathcal{W}$-MDE demonstrates its advantages. It gives much better estimates than the other three estimators even when the probability of outlier is very high. Note that the solid line stands for the results when initializing the algorithm with the sample covariance matrix, whereas the dashed line shows the results for random initialization (with scale $\sigma = 1$). Since the β-divergence model depends on the initialization in the sense that all samples/trials are downweighted which are outliers (wrt the model used for the kth iteration), we can improve the performance when applying random initialization. In the case of 90% outliers random initialization of the algorithm performs significantly better than initialization with the sample covariance matrix because in the former case the initialized matrix has the same scale as the clean data (thus the outlier trials are being discarded), whereas in the latter case, the initialized matrix has the scale of the outlier trials (thus the clean data is discarded as outlier). Note that our estimator resists the presence of 90% outliers (i) because it is model-driven, i.e., penalizes the outliers based on their likelihood of being generated by the model and not by using heuristics, and thus is very robust if initialized with a parameter close to the true solution, and (ii) because in this simulation each outlier trial was sampled from a distribution with a different covariance parameter, thus no common outlier model exists. Note that other initialization strategies can

be applied in practice and may positively affect the results.

The bottom row shows the results for the large scale artifacts. Here MCDE shows its advantages (preference of covariance matrices with small determinant) until the probability of outlier exceeds 50% (breaking point). After that its error largely increases. For $\mathcal{G}$-MDE the performance is quite stable until 80% of outliers. Note that this method performs so well because it iteratively discards the extreme outlier samples even when initialized with the average covariance matrix. In other words even when the probability of outlier is very high many samples will lie in the range of the clean data and extreme outlier samples will be rapidly downweighted because the sample covariance matrix, which is used for initialization, does not provide enough support for them. Thus, iteratively these samples will have less and less influence and the final estimate will be better than the sample covariance matrix. Note that the positive effect of random initialization is limited in the Gaussian model as the extreme outliers are downweighted anyway. In the group-level scenario $\mathcal{W}$-MDE performs very well until the point where more than 20% of the trials become outliers*. Beyond this point the error largely increases as the initialization with the sample covariance matrix prefers the outlier trials over the clean trials (which are treated as outliers). However, when using random initialization the initial covariance matrix has the same scale as the clean data and $\mathcal{W}$-MDE downweights the influence of outlier trials until the probability of outlier exceeds 90%. We would like to stress that in practice it is often impossible to correctly estimate parameters in a 90% outlier setting, because the outlier model is much more complex and prior information about the scale of the correct parameter is not available.

5.2. Motor Imagery BCI

This section investigates the impact of robust parameter estimation on spatial filtering algorithms in BCI, in particular CSP and its variants. These methods are well suited to discriminate between two motor imagery classes because they enhance the differences in band power (ERD/ERS♯ [93, 5]) between the conditions. Mathematically, a CSP spatial filter $\mathbf{w} \in \mathbb{R}^D$ maximizes / minimizes the Rayleigh quotient

$$R(\mathbf{w}) \;=\; \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}}, \tag{17}$$

where $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{D\times D}$ are the estimated (average) covariance matrices of the two conditions, e.g., left hand and right hand motor imagery. If these covariance matrices are not well estimated, e.g., due to artifacts in the EEG, then the spatial filters will not be physiologically meaningful, thus will not extract the BCI related neural activity. Robust covariance matrix estimators downweight artifactual samples / trials and thus result in better spatial filters.

5.2.1. Data Sets and Setup We use two data sets, namely the Vital BCI data set [94] consisting of EEG recordings from 80 subjects and the BCI Competition III dataset IVa [95]

* This point may vary in different data sets.

♯ Event-related Desynchronization / Event-related Synchronization occurs in specific locations and frequency bands after motor imagery.

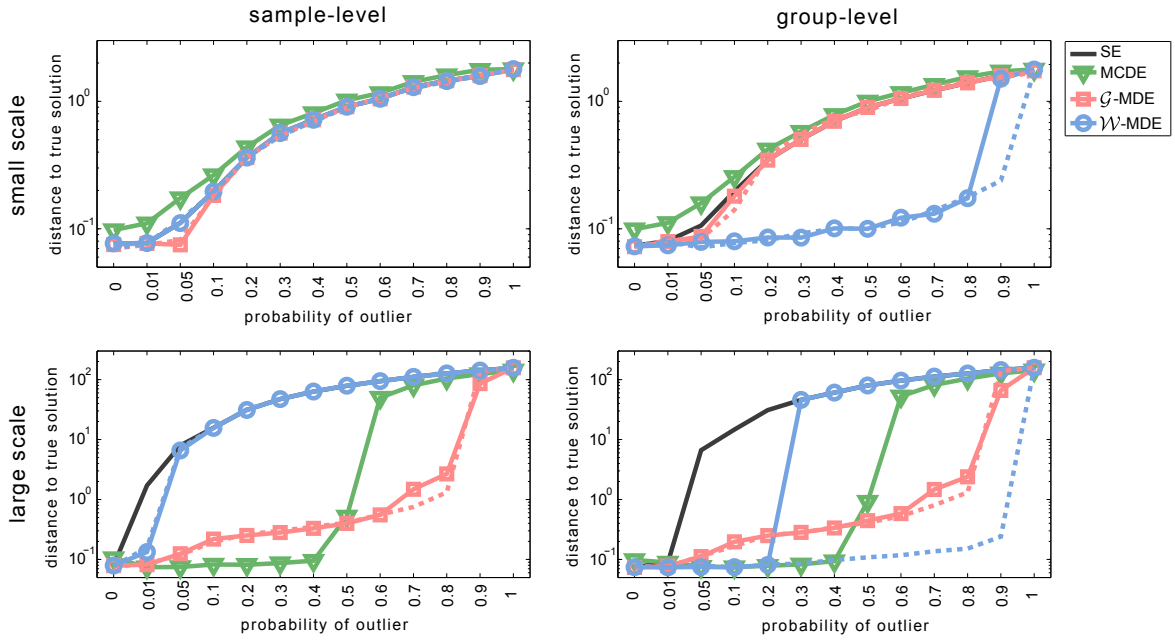

Figure 5: Comparison of different robust covariance matrix estimators in four different outlier scenarios. The proposed $\mathcal{W}$-MDE outperforms all other estimators when whole trials are sampled from an outlier distribution (right column). When outliers have much higher variance than clean data and the probability of outlier exceeds a certain value, the performance of the estimator may largely depend on the initialization. The solid lines stand for initialization with the sample covariance matrix, whereas the dashed lines represent random matrix initialization.

consisting of 5 users, for the experimental evaluation.

In the Vital BCI data set the experiment started with a calibration session in which participants were asked to perform motor imagery tasks with the left and right hand or with the feet. After recording 75 trials for each condition, the best binary combination of motor imagery tasks was selected and the BCI system was trained. Subsequently, the feedback session started in which the system decoded the imagined movement. Visual feedback was provided to the user. The decoding efficiency of the system is measured in terms of classification accuracy. The signals were recorded from 118 Ag/AgCl electrodes, from which we manually select 62 electrodes densely covering the motor cortex. We downsample the signal to 100 Hz and apply a 5th order Butterworth filter with pass-band 8-30 Hz. We use a fix time segment from 750 to 3500 ms after the trial start for feature extraction.

The BCI Competition III dataset IVa contains EEG signals from five healthy subjects performing right hand and foot motor imagery without feedback. Two types of visual cues, a letter appearing behind a fixation cross and a randomly moving object, shown for 3.5 s were used to indicate the target class. The presentation of target cues were sandwiched between periods of random length, 1.75 to 2.25 s, in which the subject could relax. The EEG signal was recorded from 118 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz, so that 280 trials are available for each subject. We manually

select 68 electrodes densely covering the motor cortex and band-pass filter the signal in the frequency range 8-30 Hz using a 5th order Butterworth filter.

For both data sets we compute CSP (3 filters per class) with the class-covariance matrices estimated with SE, MCDE (with parameters $h = 0.5{:}0.05{:}1$), $\mathcal{G}$-MDE (with parameters $\beta = 2^{-15:1:0}$) and $\mathcal{W}$-MDE and $\mathcal{WG}$-MDE (both with parameters $\beta = 2^{-15:0.5:-8}$). We select the parameters by minimizing cross-validation error on the training data. Following spatial filtering log-variance features are computed and the LDA classifier is applied [39].

5.2.2. Robust Estimation Improves BCI Performance Table 1 displays the error rates obtained with the different covariance matrix estimators on both datasets. One can see that all robust estimators clearly outperform the SE baseline and that the lowest error rates are obtained for the combined estimator $\mathcal{WG}$-MDE. This result shows that both types of outliers (i.e., sample and trial) occur in the BCI datasets and negatively affect the spatial filter computation. Since the amount of outliers in the data vary from subject to subject, not all users benefit from robust estimation (e.g., subjects A2 and A5).

Figure 6 provides an overview of the Vital BCI results using scatter plots. For individual subjects the improvement over the SE baseline is quite large. For instance, subject 21 has chance-level performance (i.e., error rate 46 %) when computing spatial filters by using SE, but the error rate decreases to 17 % when applying $\mathcal{WG}$-MDE. Similar error rate decreases are obtained for the other robust estimators. The overall performance improvement of $\mathcal{WG}$-MDE over SE, MCDE and $\mathcal{G}$-MDE is significant with $p < 0.05$ and over $\mathcal{W}$-MDE with $p < 0.01$ according to the one-sided Wilcoxon sign-rank test. Also when considering the best parameter for each subject $\mathcal{WG}$-MDE leads to an average error rate of 23.5% and clearly outperforms MCDE, $\mathcal{G}$-MDE and $\mathcal{W}$-MDE with error rates of 27.4%, 25.1% and 25.1%, respectively. This result indicates that some subjects benefit more from the trial-level robustification performed by $\mathcal{W}$-MDE, whereas for others robust sample-level estimation performed by $\mathcal{G}$-MDE (and MCDE) suffices. For instance, subject 74 has an error rate of 50%, 49%, 37% when computing spatial filters by using SE, MCDE and $\mathcal{G}$-MDE, respectively, whereas the error rate decreases to 29% when using $\mathcal{WG}$-MDE. The error rate of this subject can even be lowered to 21% when applying $\mathcal{W}$-MDE with the best parameter, but a comparable performance can not be obtained with the sample-level estimators. Similarly, the error rate of subject 10 can be lowered from 49% to 23% when using $\mathcal{G}$-MDE instead of SE, but the improvement obtained with $\mathcal{W}$-MDE is much smaller (even for the best parameter the error rate stays above 32%).

Since $\mathcal{WG}$-MDE combines the advantages of the trial- and sample-level estimator, it leads to the best overall performance in the Vital BCI dataset. For BCI Competition dataset the advantages of the combined estimator over the group-level one are only marginal, i.e., both estimators lead almost to the same solution. There are two potential explanations for this result. First, subjects A1-A5 may be not affected by sample-level outliers (or at least much less affected than subjects in the Vital BCI dataset), thus applying a robust estimator in the first step of $\mathcal{WG}$-MDE does not have any advantage over applying the standard estimator. This seems not to be the case as $\mathcal{G}$-MDE clearly outperforms the SE baseline. Second, the parameter β used for the robust estimator in the first step of $\mathcal{WG}$-MDE may be too small, so

that the robust sample-level estimator "behaves" like the standard estimator. Note that we use only one parameter for $\mathcal{WG}$-MDE in our experiments, i.e., the same β is used for estimating the trial-wise scatter matrices with $\mathcal{G}$-MDE and for estimating the final covariance matrix with $\mathcal{W}$-MDE. This brings the risk that the parameter may be of optimal scale for the second step of $\mathcal{WG}$-MDE, but too small for the first one; or of optimal scale for the first step, but too large for the second one. The two datasets may be differently affected by this scale effect. By using two separate β parameters (and also adapting ν) the $\mathcal{WG}$-MDE results can be potentially further improved, but we leave the investigation of more complex parameterizations for future work.

5.2.3. Robust Estimation Decreases the Condition Number In the following we discuss why robustly estimated parameters lead to lower error rates in BCI applications. Figure 7 provides an intuitive explanation for the large performance improvement of subject 21. It shows the right hand motor imagery patterns computed from the CSP filters with SE and $\mathcal{W}$-MDE. The pattern obtained by using SE shows activity in the right hemisphere and at the left temporal electrodes. This activity is due to artifacts and does not have neurophysiological origin. The signal recorded at C6 electrode shows strong artifacts which negatively influence the covariance estimation and the spatial filter computation for this subject. The $\mathcal{W}$-MDE (and also the other robust estimators) downweights these artifactual trials (bottom row) and allows to extract the true motor imagery related neural activity.

Mathematically, we can show that the robustly estimated covariance matrices have a significantly smaller condition number compared to the covariance matrices estimated by SE (t-test, $p \ll 0.001$). The condition number of a matrix is defined as the ratio of the largest and smallest eigenvalue. Also we can show that the decrease of the condition number is correlated to the error rate decrease ($r = 0.2927$, $p < 0.01$). This result provides one explanation why robustly estimated parameters lead to lower error rates in BCI applications. The artifacts in the data (if not removed) lead to large condition number of the class-covariance matrix which negatively affects the stability of the CSP solution. Since CSP is a greedy algorithm (i.e., computes a maximum likelihood solution), it is largely affected by the over- and underestimated eigenvalues and thus focuses on the artifacts instead of the true neural activity.

5.2.4. Sample and Trial Robustness Revisited This section analyzes the relation between the proposed sample-level and trial-level estimators in more detail. In particular, we show that the trials considered as outliers by $\mathcal{G}$-MDE and $\mathcal{W}$-MDE only partially overlap. The left upper panel of Figure 8 shows the normalized weights ψ_β assigned to each trial by $\mathcal{G}$-MDE (computed as average over sample weights) and $\mathcal{W}$-MDE for subject 1. Although the correlation between the weights is quite high, $r = 0.7$, some trials are downweighted by $\mathcal{W}$-MDE but not by $\mathcal{G}$-MDE and vice versa. The two trials marked by the red and green circles in Figure 8 are almost completely discarded (weight close to zero) by $\mathcal{W}$-MDE but are not downweighted by $\mathcal{G}$-MDE. The time courses of these trials are shown in the bottom panel of Figure 8. Obviously channel CPz has a problem (recorded values exactly zero) which

Table 1: BCI error rates when using different estimators for spatial filter computation.

Estimator	**BCI Competition**						**Vital BCI**
	A1	A2	A3	A4	A5	Overall	Overall
SE (baseline)	33.9	3.6	41.8	11.2	19.0	21.9	30.4
MCDE	26.8	5.4	40.8	12.1	22.6	21.5	29.6
$\mathcal{G}$-MDE	25.0	3.6	37.2	8.5	13.1	17.5	29.4
$\mathcal{W}$-MDE	19.6	3.6	33.2	9.4	20.2	17.2	29.5
$\mathcal{WG}$-MDE	19.6	3.6	32.7	9.4	20.2	17.1	28.9

negatively affects the spatial filter computation when using the SE (i.e., top CSP filter does not capture BCI related neural activity). Furthermore, $\mathcal{G}$-MDE does not identify these trials as outliers, i.e., it does not downweight them. There exist also the opposite case where trials (right bottom corner of the figure) are downweighted by $\mathcal{G}$-MDE but not penalized by $\mathcal{W}$-MDE. In general it is not only important to downweight the outliers trials correctly, but also to assign high weights to representative, non-outlier trials.

The right panel of Figure 8 demonstrates the superiority of $\mathcal{W}$-MDE in identifying representative trials. The dashed line represents the average (over all subjects) test performance of SE when computing the spatial filters (1 per class) using all 75 calibration trials per class. The three solid lines stand for the average performance of SE when using the 2-20 *best* trials according to $\mathcal{G}$-MDE and $\mathcal{W}$-MDE weighting or random selection. More precisely, we select 2-20 trials per class with the highest $\mathcal{G}$-MDE and $\mathcal{W}$-MDE weights or by random selection (50 repetitions) and compute the spatial filters and the classifier using these trials. Note that the weights ψ_β were computed on all calibration trials. Values above the dashed line stand for an error rate increase relative to the 75 trials baseline. One clearly sees that the weights computed by $\mathcal{W}$-MDE belong to "better", i.e., more informative, trials than the $\mathcal{G}$-MDE weights or random selection. For instance, when selecting the six best trials according to the $\mathcal{W}$-MDE weighting the average error rate only increase by 4% compared to the 75 trials baseline, whereas the performance loss is twice as large when using the $\mathcal{G}$-MDE weights. One intuitive explanation for this result is that it is easier to identify good trials when looking at the data from trial-level than from sample-level perspective because the average sample weight of a trial (sample-level perspective) does not well reflect it's overall quality. The proposed $\mathcal{W}$-MDE provides this trial-level view and performs better on this task.

Figure 9 displays the C3 channel signal of the 8 best trials according to the $\mathcal{G}$-MDE and $\mathcal{W}$-MDE weighting of subject 19. A good weighting would select trials with a typical motor imagery effect over the C3 channel, i.e., high variance in the right hand condition and low variance in the left hand condition. One can see that $\mathcal{W}$-MDE selects "better" trials and produces a neurophysiologically more meaningful CSP pattern when maximizing the 'right hand' condition than $\mathcal{G}$-MDE.

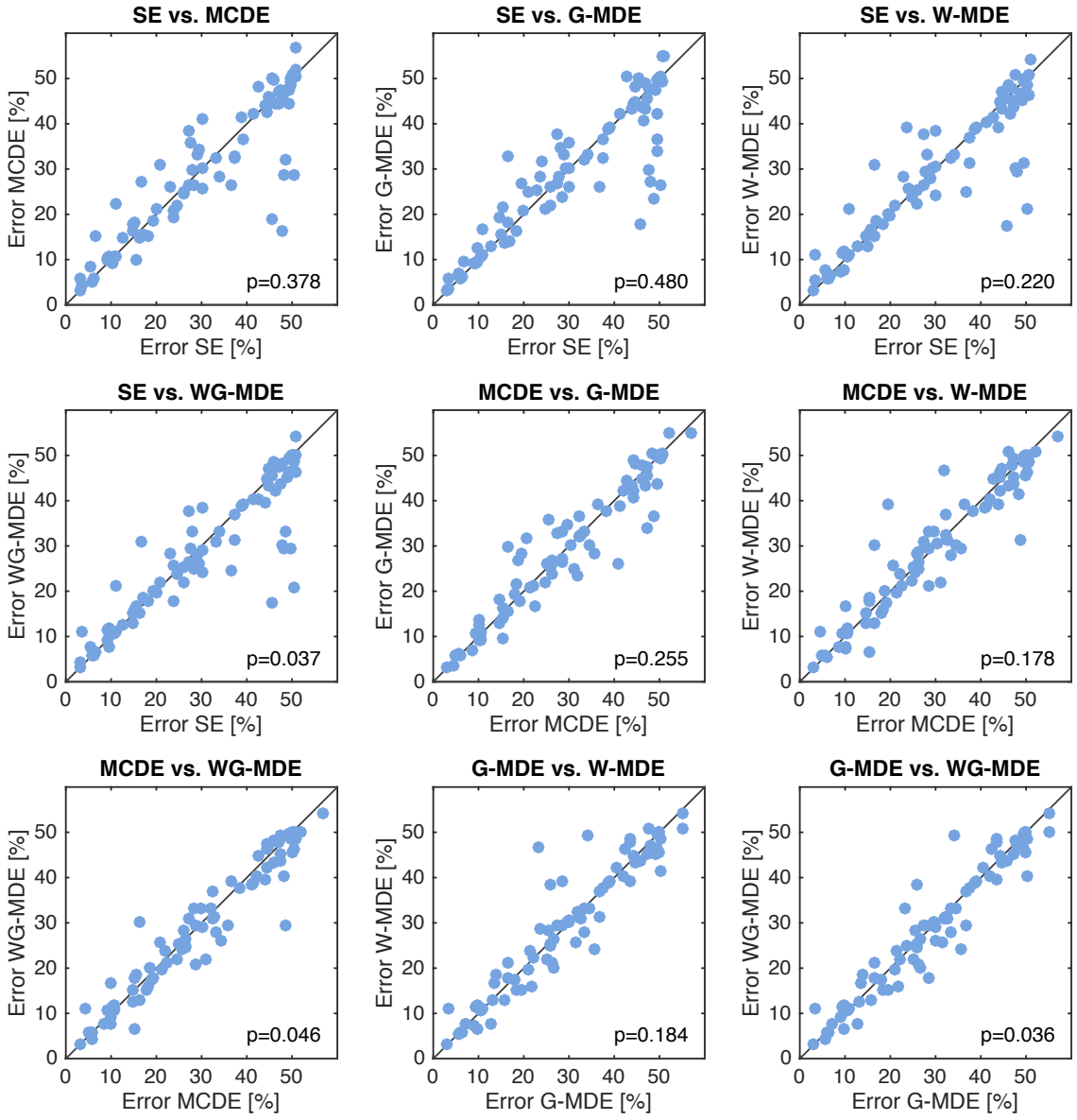

Figure 6: Scatter plots comparing the error rates obtained when different estimators are used for spatial filter computation.

5.2.5. Do CSP Variants also Benefit from Robust Estimation ? In the following analysis we show that robust parameter estimation is also important for more advanced CSP variants. Table 2 displays the average error rate of the 80 Vital BCI subjects (median error rate is shown in the brackets). The asterisks indicate significance (0.05, 0.01 and 0.001 level) when comparing the error rate of the given method to the corresponding (non-robust and robust) CSP baseline using the Wilcoxon sign-rank test. For simplicity, we apply TRCSP [44] and sCSP [50] with a fixed parameter of $\lambda = 0.02$ and β-CSP [51] with $\beta = 0.1$, i.e., we do not select these parameters for each subject separately. The first row shows the error rates when estimating the covariance matrices with SE, whereas the second row shows the results when using the covariance matrices computed by $\mathcal{WG}$-MDE. Also here we rely on the β parameter

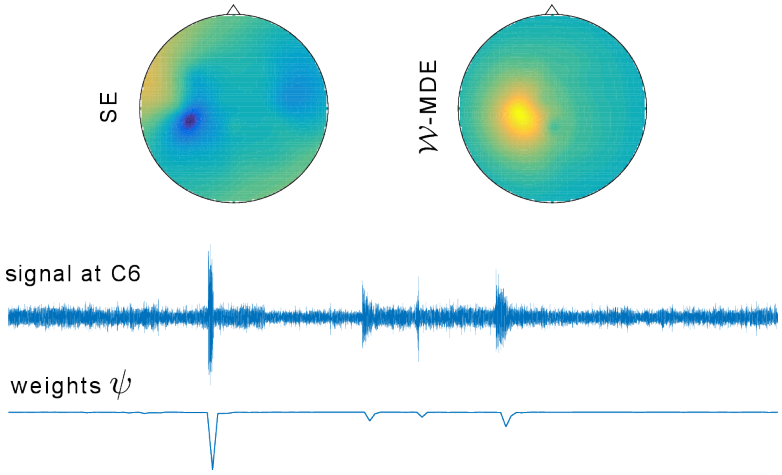

Figure 7: Top: Right hand motor imagery patterns computed from the CSP filters with SE and $\mathcal{W}$-MDE. Bottom: The signal at C6 electrode is affected by artifacts which influence the estimation of the sample covariance matrix and the computation of spatial filters. The SE pattern (top left panel) is clearly affected by these artifacts as it shows activity over the right hemisphere. $\mathcal{W}$-MDE downweights these artifactual trials and shows a much clearer right hand motor imagery pattern.

Table 2: Comparison of BCI error rates when using a standard / robust estimators for spatial filter computation. The asterisks indicate significant improvement over the CSP baseline.

	CSP	**TRCSP [44]**	**sCSP [50]**	**β-CSP [51]**
Standard	30.4 (29.7)	28.8 (26.5)**	29.4 (27.5)**	29.4 (28.9)*
Robust	28.9 (29.2)	27.7 (24.2)***	28.7 (28.1)*	28.9 (25.8)

which was selected in the above analysis, i.e., we refrain from selecting the optimal parameters for each of the CSP variants separately.

The results show a significant error rate decrease relative to the CSP baselines. All methods perform significantly better than CSP in the non-robust as well as in the robust setting (for β-CSP the error rate decrease is only significant in the non-robust setting). Also in all cases the results improve when robustly estimating the covariance matrices (i.e., comparing the first and second row). This result shows that even algorithms such as β-CSP which are robust by design or TRCSP which stabilize the CSP algorithm (e.g., in the case of large condition numbers) by restricting the norm of the filters benefit from robust parameter estimation.

5.2.6. Robust Estimation and Nonstationarity Nonstationarity in EEG is a critical issue, especially because it aggravates the session-to-session transfer in BCI. The development of techniques which robustify the signal analysis against nonstationarity is therefore of large interest to researchers as well as practicians. Data from BCI experiments can be contaminated

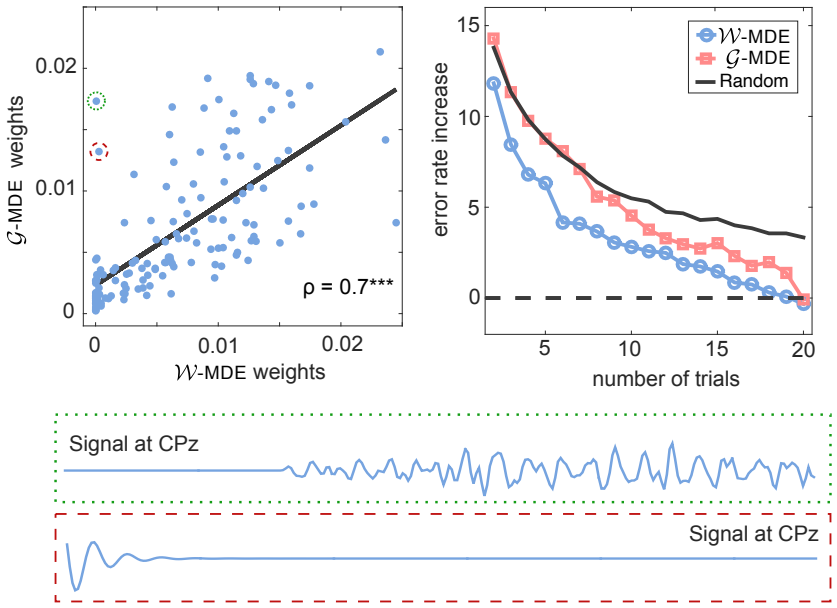

Figure 8: Left: Comparison of the weights of subject 1 computed by $\mathcal{G}$-MDE and $\mathcal{W}$-MDE. Although a strong correlation exists, one can identify trials which are downweighted by $\mathcal{W}$-MDE but not by $\mathcal{G}$-MDE and vice versa. Right: The three solid lines represent the performance (relative to the 75 trials baseline, dashed line) of SE when computed on 2-20 trials per class selected based on the $\mathcal{G}$-MDE and $\mathcal{W}$-MDE weights or random selection. The weights computed by $\mathcal{W}$-MDE select more representative trials, thus lead to smaller error rate increase than $\mathcal{G}$-MDE. Bottom: Signal at CPz channel of the two trials marked by the red and green circles in the top left panel. These artifactual trials are correctly downweighted by $\mathcal{W}$-MDE.

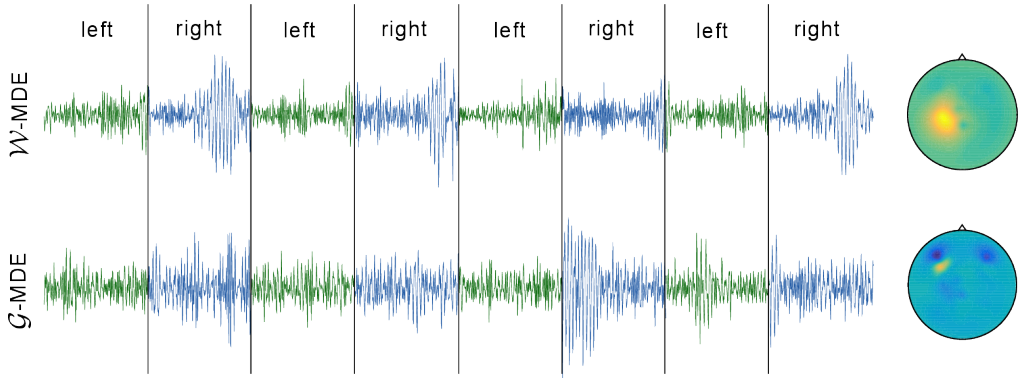

Figure 9: The best 8 trials (C3 channel) selected by $\mathcal{W}$-MDE and $\mathcal{G}$-MDE. The former method selects more representative trials and leads to a neurophysiologically more meaningful CSP pattern.

with artifacts or be free of outliers; it can be stationary or it's distribution can change significantly over time. The robust estimator proposed in this work is designed to tackle the outlier problem, however, the results in Section 5.2.2 suggest that it also has a positive effect (because it clearly outperforms the standard estimator) on datasets, which are known to exhibit nonstationarity between training (calibration data) and test (feedback data) phase. Here are two potential explanations for this observation.

(1) The smaller condition number of the proposed estimator (see Section 5.2.3) acts as a regularizer and thus prevents the estimator to overfit on the training data. Thus, the estimated parameters are less training data specific and better capture the global properties of the signal which we believe are often more stationary. This reduces the vulnerability to nonstationarity.

(2) The robust estimator implicitly focuses more on the stationary part of the data, because it downweights the impact of trials which exhibit the most changes (i.e., outliers) in the training data (see Eq. (14)). Discarding these trials does not completely remove the nonstationarity from the data, but it reduces it by a significant amount.

Thus, robust parameters estimation implicitly robustifies the signal analysis against nonstationarities in the training data, which often also helps in session-to-session transfer. However, it does so only to a certain extend. The results in Table 2 show that the combination of robust estimation and an explicit minimization of nonstationarity (by applying sCSP) gives further improvement over mere robust estimation. Depending on the data, robustness against outliers or the minimization of nonstationarity may have a larger effect on performance.

6. Conclusion

This work introduced a novel robust covariance matrix estimator based on the minimum divergence principle and a Wishart distribution model. We demonstrated the advantages of this estimator for structured data in simulations and on real data sets.

In future work we will consider the use of alternative disparity measures, e.g., optimal transport [96] or γ-divergence [85], and models, e.g., Multimodal distribution or Dirichlet distribution, for robust estimation. Furthermore, we aim to provide a Bayesian interpretation for multi-scale robustness and apply the proposed estimator in the medical domain for clinical multi-site studies. Furthermore, we will investigate the advantages of robust parameter estimation for multi-modal data [97, 98] and multi-subject BCI settings [99]. Finally, we will also study the impact on outliers on BCI performance prediction methods (e.g., [100, 101, 102]).

Acknowledgement

This work was supported by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education. The Institute for Information & Communications Technology Promotion (IITP) grant, funded by the Korea government (No.

2017-0-00451), supported this work. KRM gratefully acknowledges financial support from DFG (DFG SPP 1527, MU 987/14-1) and BMBF (BBDC). This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein. Correspondence to WS and KRM.

References

[1] P. J. Huber, *Robust Statistics*, ser. Wiley Series in Probability and Statistics. Wiley-Interscience, 1981.

[2] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[3] S. Van Aelst and P. Rousseeuw, "Minimum volume ellipsoid," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 71–82, 2009.

[4] A. Basu, H. Shioya, and C. Park, *Statistical inference: the minimum distance approach*. CRC Press, 2011.

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

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

[7] M. M. Moore, "Real-world applications for brain-computer interface technology," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 162–165, 2003.

[8] F. Lotte, J. Fujisawa, H. Touyama, R. Ito, M. Hirose, and A. Lécuyer, "Towards ambulatory brain-computer interfaces: A pilot study with p300 signals," in *Proceedings of the International Conference on Advances in Computer Enterntainment Technology*. ACM, 2009, pp. 336–339.

[9] B. Blankertz, L. Acqualagna, S. Dähne, S. Haufe, M. Schultze-Kraft, I. Sturm, M. Usćumlic, M. A. Wenzel, G. Curio, and K.-R. Müller, "The berlin brain-computer interface: Progress beyond communication and control," *Frontiers in Neuroscience*, vol. 10, p. 530, 2016.

[10] S. Haufe, M. S. Treder, M. F. Gugler, M. Sagebaum, G. Curio, and B. Blankertz, "EEG potentials predict upcoming emergency brakings during simulated driving," *Journal of Neural Engineering*, vol. 8, no. 5, p. 056001, 2011.

[11] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K.-R. Müller, and T. Wiegand, "Toward a direct measure of video quality perception using eeg," *IEEE transactions on Image Processing*, vol. 21, no. 5, pp. 2619–2629, 2012.

[12] M. De Vos, K. Gandras, and S. Debener, "Towards a truly mobile auditory brain–computer interface: exploring the p300 to take away," *International journal of psychophysiology*, vol. 91, no. 1, pp. 46–53, 2014.

[13] A. Ojeda, N. Bigdely-Shamlo, and S. Makeig, "Mobilab: an open source toolbox for analysis and visualization of mobile brain/body imaging data," *Frontiers in Human Neuroscience*, vol. 8, no. 121, 2014.

[14] S. Brandl, J. Höhne, K.-R. Müller, and W. Samek, "Bringing bci into everyday life: Motor imagery in a pseudo realistic environment," in *Proceedings of the International IEEE/EMBS Neural Engineering Conference (NER)*, 2015, pp. 224–227.

[15] 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.

[16] F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, "Semi-automatic identification of independent components representing eeg artifact," *Clinical Neurophysiology*, vol. 120, no. 5, pp. 868–877, 2009.

[17] B. W. McMenamin, A. J. Shackman, J. S. Maxwell, D. R. Bachhuber, A. M. Koppenhaver, L. L. Greischar, and R. J. Davidson, "Validation of ica-based myogenic artifact correction for scalp and source-localized eeg," *NeuroImage*, vol. 49, no. 3, pp. 2416–2432, 2010.

[18] I. Winkler, S. Haufe, and M. Tangermann, "Automatic classification of artifactual ICA-components for artifact removal in EEG signals," *Behavioral and Brain Functions*, vol. 7, no. 1, p. 30, 2011.

[19] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[20] N. Mammone, F. La Foresta, and F. C. Morabito, "Automatic artifact rejection from multichannel scalp eeg by wavelet ica," *IEEE Sensors Journal*, vol. 12, no. 3, pp. 533–542, 2012.

[21] M. Chaumon, D. V. Bishop, and N. A. Busch, "A practical guide to the selection of independent components of the electroencephalogram for artifact correction," *Journal of neuroscience methods*, vol. 250, pp. 47–63, 2015.

[22] A. Barachant, A. Andreev, and M. Congedo, "The riemannian potato: an automatic and adaptive artifact detection method for online experiments using riemannian geometry," in *TOBI Workshop lV*, 2013, pp. 19–20.

[23] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "Emg and eog artifacts in brain computer interface systems: A survey," *Clinical neurophysiology*, vol. 118, no. 3, pp. 480–494, 2007.

[24] J. A. Urigüen and B. Garcia-Zapirain, "Eeg artifact removal state-of-the-art and guidelines," *Journal of neural engineering*, vol. 12, no. 3, p. 031001, 2015.

[25] L. Frølich, I. Winkler, K.-R. Müller, and W. Samek, "Investigating effects of different artefact types on motor imagery bci," in *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 1942–1945.

[26] S. Brandl, L. Frølich, J. Höhne, K.-R. Müller, and W. Samek, "Brain-computer interfacing under distraction: An evaluation study," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056012, 2016.

[27] N. Tomida, T. Tanaka, S. Ono, M. Yamagishi, and H. Higashi, "Active data selection for motor imagery eeg classification," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 458–467, 2015.

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

[29] 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*, ser. LNCS. Springer, 2009, pp. 279–282.

[30] W. Samek and M. Kawanabe, "Robust common spatial patterns by minimum divergence covariance estimator," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2059–2062.

[31] T. Uehara, T. Tanaka, and S. Fiori, "Robust averaging of covariance matrices by riemannian geometry for motor-imagery brain–computer interfacing," in *Advances in Cognitive Neurodynamics (V)*. Springer, 2016, pp. 347–353.

[32] Y. Wang, S. Gao, and X. Gao, "Common spatial pattern method for channel selection in motor imagery based brain-computer interface," in *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2006, pp. 5392–5395.

[33] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in eeg-based bci," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 6, pp. 1865–1873, 2011.

[34] C. Sannelli, T. Dickhaus, S. Halder, E.-M. Hammer, K.-R. Müller, and B. Blankertz, "On optimal channel configurations for smr-based brain–computer interfaces," *Brain Topography*, vol. 23, no. 2, pp. 186–193, 2010.

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

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

[37] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Transactions on*

Biomedical Engineering, vol. 59, no. 3, pp. 653–662, 2012.

[38] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 1998.

[39] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

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

[41] 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.

[42] 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 Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.

[43] 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, 2010.

[44] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355 –362, 2011.

[45] 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.

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

[47] 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.

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

[49] 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 *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008, pp. 113–120.

[50] 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.

[51] 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.

[52] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[53] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges," *Signal Processing*, vol. 85, no. 11, pp. 2190 – 2212, 2005.

[54] 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, p. 214101, 2009.

[55] D. A. Blythe, F. C. Meinecke, P. von Bünau, and K.-R. Müller, "Explorative data analysis for changes in neural activity," *Journal of neural engineering*, vol. 10, no. 2, p. 026018, 2013.

[56] 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.

[57] 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.

[58] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 4, pp. 610–619, 2013.

[59] A. Balzi, F. Yger, and M. Sugiyama, "Importance-weighted covariance estimation for robust common

spatial pattern," *Pattern Recognition Letters*, vol. 68, pp. 139–145, 2015.

[60] P. von Bünau, F. Meinecke, S. Scholler, and K.-R. Müller, "Finding stationary brain sources in EEG data," in *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 2810–2813.

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

[62] I. Horev, F. Yger, and M. Sugiyama, "Geometry-aware stationary subspace analysis," in *Proceedings of the Asian Conference on Machine Learning (ACML)*, 2016, pp. 430–444.

[63] 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.

[64] 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.

[65] 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.

[66] ——, "Multiclass brain-computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[67] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: a review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2016.

[68] 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.

[69] D. Bartz and K.-R. Müller, "Generalizing analytic shrinkage for arbitrary covariance structures," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 1869–1877.

[70] J. Höhne, D. Bartz, N. Hebart, K.-R. Müller, and B. Blankertz, "Analyzing neuroimaging data with subclasses: A shrinkage approach," *NeuroImage*, vol. 124, pp. 740–51, 2014.

[71] A. Yuksel and T. Olmez, "A neural network-based optimal spatial filter design method for motor imagery classification," *PLOS ONE*, vol. 10, no. 5, p. e0125039, 2015.

[72] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted boltzmann machines," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 6, pp. 566–76, 2017.

[73] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of eeg motor imagery signals," *Journal of Neural Engineering*, vol. 14, no. 1, p. 016003, 2016.

[74] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of Neuroscience Methods*, vol. 274, pp. 141–145, 2016.

[75] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[76] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *arXiv*, no. 1706.07979, 2017. [Online]. Available: https://arxiv.org/abs/1706.07979

[77] 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.

[78] S. Amari and H. Nagaoka, "Methods of information geometry," in *Translations of Mathematical Monographs*. Oxford University Press, 2000, vol. 191.

[79] S. Amari, "Information geometry in optimization, machine learning and statistical inference," *Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 3, pp. 241–260, 2010.

[80] D. L. Donoho and R. C. Liu, "The "automatic" robustness of minimum distance functionals," *The Annals of Statistics*, vol. 16, no. 2, pp. 552–586, 06 1988.

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

[82] N. Murata, T. Takenouchi, and T. Kanamori, "Information geometry of u-boost and bregman divergence,"

Neural Computation, vol. 16, pp. 1437–1481, 2004.

[83] R. Beran, "Minimum hellinger distance estimates for parametric models," *The Annals of Statistics*, vol. 5, no. 3, pp. 445–463, 05 1977.

[84] 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.

[85] A. Notsu, O. Komori, and S. Eguchi, "Spontaneous clustering via minimum gamma-divergence," *Neural Computation*, vol. 26, no. 2, pp. 421–448, 2014.

[86] A. Basu and B. G. Lindsay, "The iteratively reweighted estimating equation in minimum distance problems," *Computational statistics & data analysis*, vol. 45, no. 2, pp. 105–124, 2004.

[87] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053 – 2081, 2008.

[88] J. Wishart, "The generalised product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20A, pp. 32–52, 1928.

[89] F. Yger, F. Lotte, and M. Sugiyama, "Averaging covariance matrices for eeg signal classification based on the csp: An empirical study," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2721–2725.

[90] R. V. Lenth, "Some practical guidelines for effective sample size determination," *The American Statistician*, vol. 55, no. 3, pp. 187–193, 2001.

[91] H. J. Thiébaux and F. W. Zwiers, "The interpretation and estimation of effective sample size," *Journal of Climate and Applied Meteorology*, vol. 23, no. 5, pp. 800–811, 1984.

[92] D. Olive, "Why the rousseeuw yohai paradigm is one of the largest and longest running scientific hoaxes in history," Southern Illinois University, Tech. Rep., 2012.

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

[94] 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.

[95] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 153–159, 2006.

[96] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[97] S. Dähne, F. Bießmann, W. Samek, S. Haufe, D. Goltz, C. Gundlach, A. Villringer, S. Fazli, and K.-R. Müller, "Multivariate machine learning methods for fusing multimodal functional neuroimaging data," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1507–1530, 2015.

[98] A. von Lühmann, H. Wabnitz, T. Sander, and K.-R. Müller, "M3ba: A mobile, modular, multimodal biosignal acquisition architecture for miniaturized eeg-nirs based hybrid bci and monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 6, pp. 1199–1210, 2017.

[99] S. Fazli, S. Dähne, W. Samek, F. Bießmann, and K.-R. Müller, "Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based brain-computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 891–906, 2015.

[100] E. M. Hammer, S. Halder, B. Blankertz, C. Sannelli, T. Dickhaus, S. Kleih, K.-R. Müller, and A. Kübler, "Psychological predictors of smr-bci performance," *Biological psychology*, vol. 89, no. 1, pp. 80–86, 2012.

[101] H.-I. Suk, S. Fazli, J. Mehnert, K.-R. Müller, and S.-W. Lee, "Predicting bci subject performance using probabilistic spatio-temporal filters," *PLOS ONE*, vol. 9, no. 2, p. e87056, 2014.

[102] W. Samek, D. Blythe, G. Curio, K.-R. Müller, B. Blankertz, and V. V. Nikulin, "Multiscale temporal neural dynamics predict performance in a complex sensorimotor task," *NeuroImage*, vol. 141, pp. 291–303, 2016.

Appendix

Suppose that we have a set of scatter matrices $\{\mathbf{S}_1, \ldots, \mathbf{S}_m\}$ where $\mathbf{S}_j = \sum_{t=1}^{N} (\mathbf{x}_t^j - \boldsymbol{\mu})(\mathbf{x}_t^j - \boldsymbol{\mu})^\top$ and $\mathbf{X}_j = [\mathbf{x}_1^j \ldots \mathbf{x}_N^j] \in \mathbb{R}^{D \times N}$ consists of the N original D-dimensional observations in the jth group. We aim to determine $\boldsymbol{\Sigma}$ by minimizing the beta divergence between the empirical distribution of the observed scatter matrices and a model Wishart distribution. The following terms can be expressed explicitly

$$\ell(\mathbf{S};\, \boldsymbol{\Sigma}, \nu) \;=\; \log \underbrace{\frac{1}{2^{\frac{\nu D}{2}} |\boldsymbol{\Sigma}|^{\frac{\nu}{2}} \Gamma_D\left(\frac{\nu}{2}\right)}}_{\alpha} + \frac{\nu - D - 1}{2} \log |\mathbf{S}| \;-\; \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1}\mathbf{S}\right)$$

$$\psi_\beta(\ell(\mathbf{S};\, \boldsymbol{\Sigma}, \nu)) \;=\; \alpha^\beta \;\cdot\; |\mathbf{S}|^{\frac{\beta(\nu - D - 1)}{2}} \;\cdot\; e^{-\frac{1}{2}\beta \mathrm{tr}(\Sigma^{-1}\mathbf{S})}$$

Note that in the following we write $\mathbf{S}^{(k)}$ and $\mathbf{S}^{(k+1)}$ to stress the dependence on $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\mu}^{(k+1)}$, respectively. If we put these definitions into Eq. (3) and set $\kappa = \mathrm{tr}((\Sigma^{(k+1)})^{-1}\mathbf{S})$ we obtain

$$\frac{1}{m}\sum_{j=1}^{m} \psi_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu)) \left(\frac{1}{2}(\boldsymbol{\Sigma}^{(k+1)})^{-1}\mathbf{S}_j^{(k+1)}(\boldsymbol{\Sigma}^{(k+1)})^{-1} \;-\; \frac{1}{2}\nu(\boldsymbol{\Sigma}^{(k+1)})^{-1}\right) \;=$$

$$\int \left(\alpha |\mathbf{S}|^{\frac{(\nu - D - 1)}{2}} e^{-\frac{1}{2}\kappa}\right) \left(\alpha^\beta |\mathbf{S}|^{\frac{\beta(\nu - D - 1)}{2}} e^{-\frac{1}{2}\beta\kappa}\right) \left(\frac{1}{2}(\boldsymbol{\Sigma}^{(k+1)})^{-1}\mathbf{S}(\boldsymbol{\Sigma}^{(k+1)})^{-1} \;-\; \frac{1}{2}\nu(\boldsymbol{\Sigma}^{(k+1)})^{-1}\right) d\mathbf{S},$$

After multiplication of both sides with $\sqrt{2}\boldsymbol{\Sigma}^{(k+1)}$ from the left and from the right we obtain

$$\frac{1}{m}\sum_{j=1}^{m} \psi_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu)) \left(\mathbf{S}_j^{(k+1)} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) \;=\; \int \left(\alpha^{\beta+1} |\mathbf{S}|^{\frac{(\beta+1)(\nu - D - 1)}{2}} e^{-\frac{1}{2}(\beta+1)\kappa}\right) \left(\mathbf{S} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) d\mathbf{S},$$

Let $\tilde{\boldsymbol{\Sigma}}^{(k+1)} = \frac{1}{\beta+1}\boldsymbol{\Sigma}^{(k+1)}$, $\nu' = (\beta+1)\nu - \beta D - \beta$ and $\alpha' = \frac{1}{2^{\frac{\nu' D}{2}} |\tilde{\boldsymbol{\Sigma}}^{(k+1)}|^{\frac{\nu'}{2}} \Gamma_D\left(\frac{\nu'}{2}\right)}$. Then we obtain

$$\frac{1}{m}\sum_{j=1}^{m} \psi_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu)) \left(\mathbf{S}_j^{(k+1)} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) \;=\; \frac{\alpha^{\beta+1}}{\alpha'} \int \left(\alpha' |\mathbf{S}|^{\frac{\nu' - D - 1}{2}} e^{-\frac{1}{2}\mathrm{tr}((\tilde{\Sigma}^{(k+1)})^{-1}\mathbf{S})}\right) \left(\mathbf{S} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) d\mathbf{S},$$

Note that if splitting the integral on the right hand side into two integrals then the first one gives the first moment of the Wishart distribution and the second one is the zeroth moment times $\nu\boldsymbol{\Sigma}^{(k+1)}$. Thus, we obtain

$$\frac{1}{m}\sum_{j=1}^{m} \psi_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu)) \left(\mathbf{S}_j^{(k+1)} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) \;=\; \frac{\alpha^{\beta+1}}{\alpha'} \left(\nu'\tilde{\boldsymbol{\Sigma}}^{(k+1)} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right)$$

Assuming $|\boldsymbol{\Sigma}^{(k+1)}| = |\boldsymbol{\Sigma}^{(k)}|$ at convergence point this is

$$\frac{1}{m}\sum_{j=1}^{m} \underbrace{\left(|\mathbf{S}_j^{(k)}|^{\frac{\beta(\nu - D - 1)}{2}} e^{-\frac{1}{2}\beta \mathrm{tr}((\Sigma^{(k+1)})^{-1}\mathbf{S}_j^{(k)})}\right)}_{\psi'_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k+1)}, \nu))} \left(\mathbf{S}_j^{(k+1)} \;-\; \nu\boldsymbol{\Sigma}^{(k+1)}\right) \;=\; -\frac{\alpha(\beta D + \beta)}{\alpha'(\beta+1)}\boldsymbol{\Sigma}^{(k+1)}$$

This leads to the iterative formula

$$\boldsymbol{\Sigma}^{(k+1)} \;=\; \frac{\frac{1}{m}\sum_{j=1}^{m} \psi'_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu))\mathbf{S}_j^{(k+1)}}{\frac{\nu}{m}\sum_{j=1}^{m} \psi'_\beta(\ell(\mathbf{S}_j^{(k)};\, \boldsymbol{\Sigma}^{(k)}, \nu)) \;-\; \frac{\alpha(\beta D + \beta)}{\alpha'(\beta+1)}}$$

which implicitly depends on $\boldsymbol{\mu}$ parameter through the estimation of the scatter matrices $\mathbf{S}_j$. Note that

$$\gamma = \frac{\alpha(\beta D + \beta)}{\alpha'(\beta+1)} \;=\; \frac{\beta(D+1)\Gamma_D\left(\frac{\nu(\beta+1)}{2} \;-\; \frac{(D+1)\beta}{2}\right)}{2^{\frac{\nu D}{2}}(\beta+1)\Gamma_D\left(\frac{\nu}{2}\right)} \left(\frac{2}{\beta+1}\right)^{\frac{\nu D(\beta+1) - D(D+1)\beta}{2}} |\boldsymbol{\Sigma}^{(k)}|^{\frac{\beta(\nu - D - 1)}{2}}$$