ROBUST COMMON SPATIAL PATTERNS BY MINIMUM DIVERGENCE COVARIANCE ESTIMATOR

Wojciech Samek[*]

Berlin Institute of Technology
Machine Learning Group
10587 Berlin, Germany

Motoaki Kawanabe

ATR Institute International
Brain Information Communication Research Laboratory
619-0288 Kyoto, Japan

ABSTRACT

Reliable estimation of covariance matrices from high-dimensional electroencephalographic recordings is crucial for a successful application of Brain-Computer Interface (BCI) systems. Artifactual trials and non-stationarity effects may have a large impact on the estimation quality and adversely affect the spatial filter computation and consequently the classification accuracy of the system. In this work we propose a novel robust estimator for covariance matrices that takes into account the trial structure of BCI experiments. Our estimator minimizes beta divergence between the empirical and a model Wishart distribution, thus allows to robustly average the estimated covariance matrices of different trials and downweight the influence of outlier trials. We evaluate this novel estimator on a data set with recordings from 80 subjects.

Index Terms— Brain-Computer Interface, Robust Estimation, Beta Divergence

1. INTRODUCTION

Brain-Computer Interface (BCI) systems [1] [2] translate recorded brain signals, e.g. EEG, into control commands for a computer application by decoding the mental state of a subject (e.g. induced by left or right hand movement imagination). Common Spatial Patterns (CSP) (e.g. [3] [4]) is a popular algorithm for motor imagery based BCIs as it largely reduces the dimensionality of the data and focuses on the relevant part by maximizing the variance ratio between classes (ERS/ERD effect). The performance of CSP (and the whole BCI system) depends on reliable estimation of class covariance matrices from trials performed in the calibration session. However, artifacts such as loose electrodes, eye movements, jaw clenching or muscle activity may largely influence the recorded brain signals and lead to "outlier trials" with very different covariance structure. These trials may have a large impact on the estimated class covariance matrices used for spatial filter computation as neither the standard covariance estimator nor the averaging of trial covariance matrices are robust to outliers.

The development of robust spatial filtering methods have recently gained a lot of attention in the BCI community [5] [6] [7] [8] [9] [10]. Classical ICA-based approaches for artifact identification and removal have also been applied in this context [11]. In this paper we neither directly robustify the CSP algorithm nor do we apply advanced artifact removal. We rather provide a novel tool for robust estimation of class covariance matrices from multiple trials.

Note that the idea to robustly estimate the covariance matrices before applying CSP is not novel. It has been applied in [5] where the authors use shrinkage to improve the estimation. But also the minimum covariance determinant (MCD) estimator has been used for this task [12]. Note that the MCD estimator provides robust estimates with respect to individual samples, i.e. it ignores trial structure and downweights outlier samples rather than whole trials. The authors of [13] introduce the idea to robustly average trial covariance matrices, i.e. downweight outlier trials. This approach relies on robust p-norms but does not have an interpretation in terms of maximizing the likelihood of an underlying data generating model. In this paper we propose an estimator that robustly averages the trial covariance matrices and has a maximum likelihood interpretation. The estimator is based on the concept of Ψ-likelihood [14] and minimizes the beta divergence between the empirical and a model Wishart distribution.

This paper is organized as follows. Section 2 discusses robust estimation on sample and on trial level and introduces our novel estimator. Section 3 applies this estimator to a data set of 80 subjects, compares the results to standard CSP and briefly discusses the advantages and limitations. We conclude this work with a summary and outlook in Section 4.

[*]This work was supported by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115), by the Japan Science and Technology Agency (German-Japanese cooperation program on computational neuroscience), the Ministry of Education, Culture, Sports, Science and Technology (Grant-in-Aid for Scientific Research B, 24300093) and by the Ministry of Internal Affairs and Communications. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein.

2. ROBUST COVARIANCE ESTIMATION

Reliable computation of the covariance matrix is of crucial importance in motor imagery based BCI. The problem can be formulated as estimation of a parameter θ of a statistical model (e.g. zero-mean Gaussian distributions $f(y, \theta)$) given observations $\mathcal{D} = \{y_i : i = 1 \ldots n\}$. A standard procedure to estimate this parameter is to maximize the log-likelihood $\mathcal{L}(\theta \mid \mathcal{D})$ of the parameter given the observations

$$\mathcal{L}(\theta \mid \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta). \tag{1}$$

This method is not robust (because of the averaging operation) in the sense that single observations y_i may dominate the solution. Eguchi and Kano [14] introduced the concept of Ψ-likelihood to perform robust parameter estimation. Intuitively they apply a sigmoid-like function Ψ to the likelihood $\ell(y_i, \theta)$ limiting the influence of each observation. In their work they show that this principle is equivalent to the minimization of so-called Ψ-divergence between the empirical and the model distribution.

In this work we use a special choice of Ψ, namely

$$\Psi_\beta(z) = \frac{\exp(\beta z) - 1}{\beta}. \tag{2}$$

By using this function the Ψ-divergence reduces to β-divergence

$$\begin{aligned} D_\beta(p(x) \mid\mid q(x)) = \int \Bigg[& \frac{1}{\beta} \left\{ p^\beta(x) - q^\beta(x) \right\} p(x) \qquad (3) \\ & - \frac{1}{\beta + 1} \left\{ p^{\beta+1}(x) - q^{\beta+1}(x) \right\} \Bigg] dx, \end{aligned}$$

where $p(x)$ denotes the empirical data distribution and $q(x)$ stands for the model probability distribution with parameter θ. By minimizing this quantity we obtain a robust estimate of the parameter θ.

2.1. Sample Perspective

One way to apply this estimator to our BCI problem is to pool data from all trials and to estimate the parameter $\theta = \Sigma \in \mathbb{R}^{C \times C}$ by minimizing beta divergence. This estimator has been derived in [14] and can be computed iteratively by

$$\Sigma^{(k+1)} = \frac{\frac{1}{n} \sum_{i=1}^{n} \psi_\beta(y_i; \Sigma^{(k)}) y_i y_i^\top}{\frac{1}{n} \sum_{i=1}^{n} \psi_\beta(y_i; \Sigma^{(k)}) - \beta / (\beta + 1)^{C/2+1}}, \tag{4}$$

where $\Sigma^{(k)}$ denotes the estimate of the parameter in kth step and $\psi_\beta(y_i; \Sigma) = e^{-\frac{1}{2}\beta y_i^\top \Sigma^{-1} y_i}$ is a factor downweighting the influence of outlier samples y_i. Note that for $\beta = 0$ this estimator reduces to the standard maximum likelihood estimator $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\top$.

2.2. Trial Perspective

In this work we propose a novel estimator that does not downweight individual EEG samples y_i, but rather reduces the influence of whole trials. Since trials represent a group of samples, our estimator provides robust estimates on a group level.

Let us assume that we have estimates of the trial covariances $\{\Sigma_i \in \mathbb{R}^{C \times C} : i = 1 \ldots n\}$ maximizing the log-likelihood in each trial and our goal is to estimate the (average) class covariance matrix in a robust manner i.e. by downweighting outlier trial covariance matrices. In other words we regard the trial covariances (or the scatter matrices) as our samples, not the EEG samples as before, and robustly combine them. For that we use the Wishart distribution $q(S; \Sigma, \nu)$

$$\frac{1}{2^{\frac{\nu C}{2}} |\Sigma|^{\frac{\nu}{2}} \Gamma_C\left(\frac{\nu}{2}\right)} |S|^{\frac{\nu - C - 1}{2}} \exp\left\{ -\mathrm{tr}\left(\frac{1}{2} \Sigma^{-1} S \right) \right\}, \tag{5}$$

where $S = \sum_{i=1}^{\nu} y_i y_i$ is the scatter matrix and Γ_C is the multivariate gamma function defined as

$$\Gamma_C\left(\frac{\nu}{2}\right) = \pi^{\frac{C(C-1)}{4}} \prod_{j=1}^{C} \Gamma\left[\frac{\nu}{2} + \frac{(1-j)}{2} \right]. \tag{6}$$

We aim to derive the covariance matrix Σ from the scatter matrices S_i of trials $i = 1 \ldots n$ by minimizing beta divergence. We can show[1] that beta divergence between Wishart distributions can be minimized by an iterative procedure

$$\Sigma^{(k+1)} = \frac{\sum_{i=1}^{n} \psi_\beta\left(S_i; \Sigma^{(k)}, \nu\right) S_i}{\nu \sum_{i=1}^{n} \psi_\beta\left(S_i; \Sigma^{(k)}, \nu\right) - \gamma |\Sigma^{(k)}|^{\frac{(\nu - C - 1)\beta}{2}}}, \tag{7}$$

where

$$\psi_\beta(S; \Sigma, \nu) = |S|^{\frac{(\nu - C - 1)\beta}{2}} \exp\left\{ -\mathrm{tr}\left(\frac{\beta}{2} \Sigma^{-1} S \right) \right\}. \tag{8}$$

is a factor downweighting the influence of outlier trials and

$$\begin{aligned} \gamma = & \frac{n\beta(C+1)}{2^{\frac{\nu C}{2}} \Gamma_C\left(\frac{\nu}{2}\right) (\beta + 1)} \left(\frac{2}{\beta + 1} \right)^{\frac{\nu C(\beta+1)}{2} - \frac{C(C+1)\beta}{2}} \qquad (9) \\ & \times \Gamma_C\left(\frac{\nu(\beta+1)}{2} - \frac{(C+1)\beta}{2} \right). \end{aligned}$$

Note that for $\beta = 0$ our estimator gives the maximum likelihood solution

$$\Sigma = \frac{1}{n\nu} \sum_{i=1}^{n} S_i. \tag{10}$$

[1] Due to space limitations we can not show the derivation here.

3. EXPERIMENTAL EVALUATION

In this section we evaluate the performance of CSP with class covariances matrices estimated by our novel Wishart Estimator and compare the results to the standard maximum likelihood CSP baseline. We use a data set [15] containing EEG recordings from 80 BCI novices performing motor imagery tasks with the left and right hand or with the feet. The data contain a calibration session and a test session with 1D visual feedback. We select the two best motor imagery classes for each subject, resulting in 150 calibration and 300 test trials. We manually select 62 electrodes densely covering the motor cortex, apply a 5th order Butterworth filter to band pass filter the data in 8-30 Hz and extract a time segment ranging from 750ms to 3500ms after the trial start. We use six spatial filters for feature extraction and perform classification by using Linear Discriminant Analysis.

Figure 1 compares the results of CSP with robustly averaged trial covariance matrices (Wishart Estimator CSP) to the standard CSP baseline ($\beta = 0$). Each circle represents the error rate of one subject and one can clearly see that the application of our novel estimator leads to an improved classification accuracy for most of the participants (circles below solid line). This improvement is statistically significant with $p = 0.0042$ according to the one-sided Wilcoxon sign-rank test. Note that the β parameter has been selected by five-fold cross-validation on the training data from $\{0, 2^{-20}, \ldots, 2^0\}$. The ν parameter has been set to a fraction (1/20) of the number of samples in the trial because EEG recordings are far from being i.i.d.

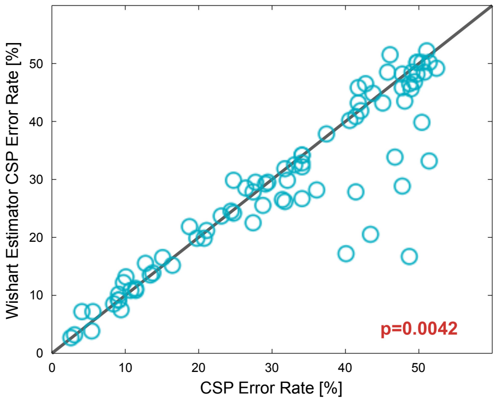

Fig. 1. Comparison of error rates of Wishart Estimator CSP and standard CSP for 80 subjects.

Figure 2 shows the median error rate (over all 80 subjects) for different β parameters. One can see that the error rate curve has a U-shape, i.e. it decreases up to a specific β value and then increases again. Very small β parameters have no robustness effect whereas too large values have a too strong influence on the solution. Finding the right trade-off is key for obtaining good performance in practice.

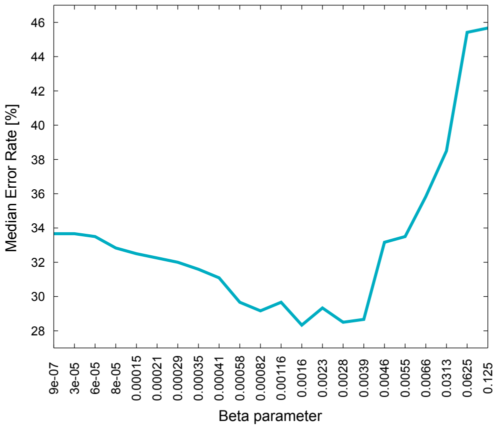

Fig. 2. Median error rate (over 80 subject) for different beta parameters.

The top panel of Fig 3 shows the weights $\psi_\beta\left(S_i; \Sigma^{(k)}, \nu\right)$ applied to the trials when computing the class covariance matrices for the subject with largest error rate decrease in Fig. 1. One can clearly see that some weights are very small, almost zero; this indicates that some trials are real outliers. On the other hand there are only small differences in the weights among the majority of trials. The bottom panel of Fig. 3 shows the signal at electrode FFC5 of the trial with lowest weight. This trial contains large artifactual amplitude activity at the beginning. This activity has an amplitude that is almost one order larger than the standard amplitude, thus this trial is an outlier trial and would have a large impact on the estimated class covariance matrix. Fortunately, it has been downweighted by our method.

4. DISCUSSION

In this work we introduced a novel robust estimator for covariance matrices which takes into account trial structure. It robustly combines trial covariance matrices thus downweights the influence of outlier trials. The estimator can be computed by minimizing beta divergence between the empirical data (samples are trial scatter matrices) and a model Wishart distribution. We derived a fast iterative algorithm to perform the minimization and showed that our estimator significantly improves BCI performance. Note that our estimator naturally takes into account the uncertainty in the estimation of the scatter matrices by using the parameter ν.

In future work we will study the advantages and limitations of the trial perspective over the sample perspective. Furthermore we will apply the idea of robust averaging of covariance matrices to other problems like the combination of covariance matrices of different users / session. Finally we will investigate the relations between robustness on the parameter

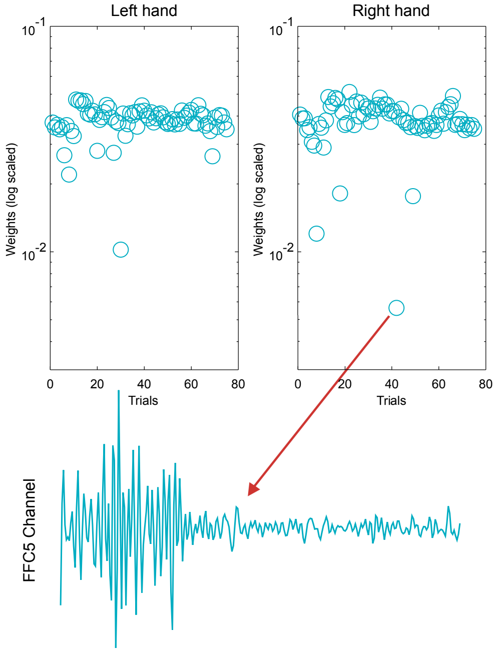

Fig. 3. Trial weights for the subject with largest performance increase (top) and the signal at electrode FFC5 of the trial with lowest weight (bottom).

estimation level (as proposed in this work), analytic shrinkage (e.g. [16]) and robustness on the CSP level (e.g. [8]).

5. REFERENCES

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

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

[3] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, “Optimizing Spatial filters for Robust EEG Single-Trial Analysis,” *IEEE Signal Proc. Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

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

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

[6] 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, pp. 026013, 2012.

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

[8] 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.

[9] W. Samek, M. Kawanabe, and K.-R. Müller, “Divergence-based framework for common spatial patterns algorithms,” *IEEE Reviews in Biomedical Engineering*, 2014, in press.

[10] 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.

[11] I. Winkler, S. Haufe, and M. Tangermann, “Automatic classification of artifactual ICA-components for artifact removal in EEG signals,” *Behavioral and brain functions : BBF*, vol. 7, no. 1, pp. 30, 2011.

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

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

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

[15] 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.

[16] D. Bartz and K.-R. Müller, “Generalizing analytic shrinkage for arbitrary covariance structures,” in *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013, pp. 1869–1877.