On robust spatial filtering of EEG in nonstationary environments

Wojciech Samek

Abstract: Brain-Computer Interfacing (BCI) is a promising technology for patients that are severely motor-disabled, because it enables them to communicate and interact with the environment. A BCI system decodes user's intentions from brain signals, typically recorded with electroencephalography (EEG), and transmits them to a computer application that, e.g., controls a wheelchair. The efficiency of the system largely depends upon a reliable extraction of informative features from the high-dimensional EEG signal. Spatial filtering is a crucial step in this protocol, however, current approaches are prone to errors when data is contaminated by artifacts or is nonstationary. This article provides an overview of a dissertation, which has addressed the problem of robust spatial filtering in BCI. The contributions of the thesis range from the development of regularization schemes and a robust parameter estimator for spatial filtering, to the formulation of an information geometric view on the spatial filtering problem and the proposal of a new family of algorithms based on robust divergences. The developed methods and concepts are applicable to a variety of problems in machine learning and signal processing.

ACM CCS: Computing methodologies → Machine learning; Mathematics of computing → Probability and statistics; Applied computing → Life and medical sciences.

Keywords: Brain-Computer Interfacing, Robust Signal Processing, Generalized Eigenvalue Problems, Robust Estimation, Classification, Nonstationarity.

1 Introduction

Brain Computer Interfaces (BCIs) [1] provide a novel communication channel between a human subject and a computer that does not rely on peripheral nerves and muscles. Since BCI communication solely depends on the user's measured brain signals, it holds a high promise for patients that are severely motor-disabled. In the early days of the technology, BCI's were based on neurofeedback and required weeks of training for the user. The development of novel machine learning algorithms which extract relevant, user-specific information from the recorded data shifted the workload from the user to the machine and drastically reduced training times [2]. Despite significant technological and methodological advances, current BCIs are still scarcely used outside laboratories. One factor limiting the large scale applicability of BCI in clinical practice and its usage as assistive technology for disabled people is the lack of robustness and reliability. Especially in uncontrolled environments, e.g., at patients' homes, a BCI system is confronted with various external and internal noise sources that either alter brain activity directly or affect the sensors (i.e., electrodes) that are used to register brain activity. For instance, changes in task involvement, additional sensory input or movement related artifacts contribute to a large part to the observed nonstationarity in the recorded signal and thus affect feature extraction and classification. Since this additional noise can not be avoided in real-life environments, a new generation of *robust* machine learning algorithms is required before the application range of BCI technology can be extended.

Motor imagery is a popular paradigm for controlling a BCI. Here the user mentally simulates a given action, e.g., a movement with the hands, feet or tongue, and this induces a change of the sensorimotor rhythms (SMRs) over specific spatial locations in the sensorimotor cortex which is then detected by the BCI system. Recognizing the exact spatial localizations of SMR modulations and computing the corresponding spatial filters is a crucial step, because focusing on this part of the signal drastically increases signal-to-noise ratio. Common Spatial Patterns (CSP) [3] is a popular algorithm for enhancing the differences in the modulation of the SMRs between two motor imagery conditions. Since CSP is prone to errors when data is contaminated with artifacts or is nonstationary, the next section presents novel approaches to robust spatial filtering in BCI.

2 Spatial Filtering in BCI

The relation between neural sources $s \in \mathbb{R}^D$ and the EEG signal recorded at the scalp $x \in \mathbb{R}^D$ can be modelled as a noisy linear mixture

$$x(t) = As(t) + n(t), \tag{1}$$

where $A \in \mathbb{R}^{D \times D}$ is the matrix mapping the activity of each source to the electrode space, and n is a noise term. Spatial filtering algorithms [4] aim to estimate the time-courses of particular sources $\hat{s}(t) \in \mathbb{R}^d$ by projecting the signal $x(t)$ linearly onto a set of $d < D$ spatial filters $W \in \mathbb{R}^{D \times d}$. In the case of motor imagery based BCIs spatial filtering increases the signal-to-noise ratio and thus simplifies the classification problem, because it shifts the focus towards sources modulating the SMRs.

The CSP algorithm computes spatial filters w by maximizing or minimizing the Rayleigh quotient

$$R(w) = \frac{w^\top \Sigma_1 w}{w^\top \Sigma_2 w}, \tag{2}$$

where Σ_1 and Σ_2 are the average covariance matrices of two motor imagery classes. One major source of error in the computation of the filters results from the difficulty in proper estimating the class covariance matrices, especially when data is scarce, high-dimensional and contaminated with artifacts. Another problem is that CSP only considers the average variances, while ignoring the within-class variability and nonstationarity of the signal.

2.1 Regularization Towards Stationarity

The CSP method can be regularized by adding a penalty term $P(w)$ to the denominator of the Rayleigh quotient. This leads to an objective function which maximizes the variance ratio between classes and at the same time aims to minimize the penalty term. If $P(w)$ is a quadratic form, i.e., $P(w) = w^\top \Delta w$ with positive definite matrix Δ, then the resulting optimization problem can be solved very efficiently and has an unique solution. The within-class variability of features (i.e., variances) can be measured in terms of absolute differences between the feature in ith trial and the class average

$$P(w) \;=\; \frac{1}{2n} \sum_{c=1}^{2} \sum_{i=1}^{n} \left| w^\top \Sigma_c^i w \;-\; w^\top \Sigma_c w \right|, \tag{3}$$

Unfortunately, this measure is not a quadratic form, however, with a slight modification [5] one can approximate each term in the above penalty term as

$$\left| w^\top \Sigma_c^i w \;-\; w^\top \Sigma_c w \right| \;\approx\; w^\top \mathcal{F}\left(\Sigma_c^i \;-\; \Sigma_c\right) w, \tag{4}$$

where $\mathcal{F}$ is an operator to make symmetric matrices be positive definite by flipping the sign of all negative eigenvalues. With this modified penalty term, the CSP algorithm is forced to extract more stationary features.

2.2 Divergence Framework

The following theorem states that spatial filtering can be formulated as divergence optimization problem [6].

Theorem: *Let $W \in \mathbb{R}^{D \times d}$ be CSP filters and Σ_c the covariance matrix of class c. Let $V^\top = \tilde{R}P \in \mathbb{R}^{d \times D}$ be decomposable into a whitening projection $P \in \mathbb{R}^{D \times D}$ and a truncated orthogonal projection $\tilde{R} \in \mathbb{R}^{d \times D}$. Then*

$$\mathrm{span}(W) = \mathrm{span}(V^*) \tag{5}$$

$$\text{with } V^* = \operatorname*{argmax}_V \tilde{D}_{kl}\left(V^\top \Sigma_1 V \;\;||\;\; V^\top \Sigma_2 V\right) \tag{6}$$

where $\tilde{D}_{kl}(A \,||\, B)$ denotes the symmetric Kullback-Leibler (KL) divergence between zero mean Gaussians with covariance matrices A and B, and span(M) *stands for the subspace spanned by the columns of M.*

The theorem provides an information geometric view [7] on the CSP algorithm which opens the door for a whole family of novel spatial filtering algorithms. By a slight relaxation of the above objective and the use of alternative divergences, one can construct spatial filtering algorithms with new desirable properties. For instance, using beta divergence $\tilde{D}_\beta$ instead of KL divergence results in a spatial filtering algorithm that is robust to artifacts in the data, because this particular divergence has the property to downweight outlier terms (i.e., artifactual trials) when maximizing the following objective

$$\mathcal{L}_\beta(V) = \sum_i \tilde{\mathrm{D}}_\beta \left(\mathrm{V}^\mathrm{T} \Sigma_1^\mathrm{i} \mathrm{V} \;\;||\;\; \mathrm{V}^\mathrm{T} \Sigma_2^\mathrm{i} \mathrm{V}\right) \tag{7}$$

Several other divergence-based spatial filtering algorithms have been proposed and evaluated in [8]. See also [9] for a related max-min based approach.

2.3 Parameter Estimation in Structured Data

Artefacts in the EEG may heavily bias the estimation of the class covariance matrices, and thus negatively affect the spatial filter computation. Many robust alternatives have been proposed to the sample covariance matrix estimator, however, none of them is specifically tailored to BCI data which have a particular structure. More precisely, BCI data consist of individual EEG *samples* which are grouped into larger units representing whole motor imagery *trials*. In many BCI systems covariance matrices are first estimated for each trial and then in the second step class averages are computed and provided as input to the CSP algorithm.

For this type of structured data robustness can be defined in two ways, namely with respect to the individual EEG samples or the sample groups (i.e., trials). Covariance matrix estimators proposed in the literature do not distinguish between both types of robustness, but take the sample-level view. In structured data, however, it may be advantageous to downweight outlier trials, rather than individual samples [10].

Parameter estimation can be regarded as a divergence minimization problem between the empirical distribution p and a given model q_θ with parameter θ, i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, D(p \,||\, q_\theta). \tag{8}$$

Using beta divergence for measuring the deviation of p from q_θ results in a robust estimator that downweights the influence of outliers [11]. Sample-level robustness can be achieved in this case when using a zero mean Gaussian distribution model $q_\theta = \mathcal{N}(0, \theta)$ with covariance matrix θ. In order to obtain robustness on trial-level, one has to treat the trial-wise covariance matrices (more precisely, the scatter matrices S_i) as samples of a Wishart distribution model $q_\theta = \mathcal{W}(\nu, \theta)$. The following iterative formula minimizes eq. 8 in this case

$$\Sigma^{(k+1)} = \frac{\sum_{i=1}^{n} \psi_\beta\left(\ell\left(S_i;\, \Sigma^{(k)}, \nu\right)\right) S_i}{\nu \sum_{i=1}^{n} \psi_\beta\left(\ell\left(S_i;\, \Sigma^{(k)}, \nu\right)\right) \;-\; \gamma |\Sigma^{(k)}|^{\alpha}} \tag{9}$$

where ℓ is the log-likelihood function and

$$\psi_\beta\left(\ell(S;\, \Sigma, \nu)\right) = |S|^{\frac{(\nu-D-1)\beta}{2}} e^{-\mathrm{tr}\left(\frac{\beta}{2}\Sigma^{-1}S\right)} \tag{10}$$

is a factor downweighting the influence of outlier trials and α and γ are constants. The robustly estimated covariance matrices can be used as input to CSP for computing spatial filters.

3 Experimental Evaluation

The proposed spatial filtering algorithms are evaluated on the Vital BCI [12] data set which contains EEG recordings from 80 healthy subjects performing motor imagery with the left hand, right hand or feet. The data set consists of a training (75 trials per class) and a test (150 trials per class) session. The signals are filtered in the frequency range 8-30 Hz with a 5th order Butterworth filter and the time segment from 750 to 3500 ms after the trial start is used for feature extraction. Free parameters are selected using cross-validation.

The scatter plots in Figure 1 compare the error rates of CSP (x-axis) and the proposed methods (y-axis). Each circle represents a subject. The plots show an overall reduction in error rate for the proposed spatial filtering methods. The performance gains are significant according to the one-sided Wilcoxon signed-rank test. Although not all subjects benefit from the additional robustness, some users show a remarkable error rate reduction. For instance, the decoding error of user *VPtbo* and *VPkl* drops from 48.6% to 22% and from 40.0% to 17.7%, respectively, when applying stationary CSP. Similar improvements can be observed for the other two methods. In total, up to ten subjects gain BCI control, i.e., the error rate becomes smaller than 30%, when using one of the proposed spatial filtering algorithms.

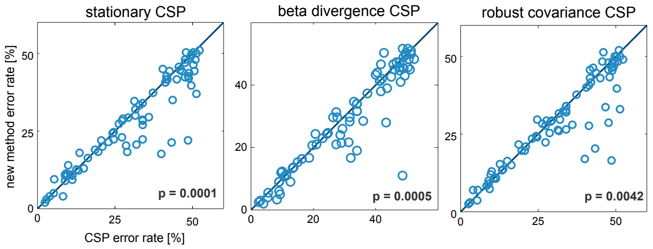

Figure 1: Scatter plots showing error rates of CSP (x-axis) and the proposed methods (y-axis). Each circle represents a subject and the p-value of the Wilcoxon signed rank test is displayed.

Figure 2 displays training (triangles) and test (circles) features extracted by CSP and stationary CSP from a particular subject. The colors represent the two motor imagery classes and the decision boundary is displayed as solid line. The features extracted by CSP show a significant drift between training and test session. Since CSP only considers the average variances, it does not penalize nonstationarities in the training data. Stationary CSP on the other hand takes into account the within-class variability and thus extracts more stable features.

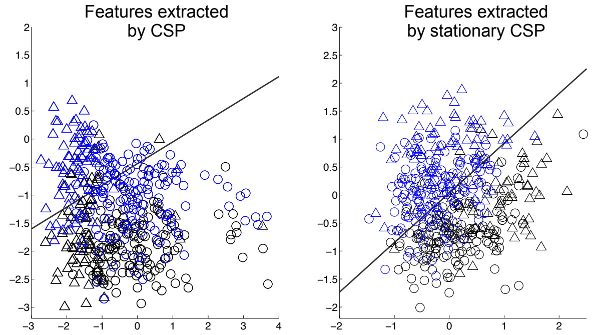

Figure 2: Comparison of training (triangles) and test (circles) features extracted by CSP and stationary CSP. Solid line is the decision boundary. CSP features significantly drift over time.

Figure 3 illustrates the robustness property of beta divergence CSP (a) and the robust covariance matrix estimator (b). The activation pattern in (a) clearly shows that the spatial filter computed with CSP does not focus on motor imagery related activity, but on artifacts in the left frontal electrode. Using beta divergence reduces the influence of these artifacts and allows one to extract a textbook-like right hand motor imagery activation pattern. The trial-wise weights ψ_β of the proposed estimator are displayed in (b). The estimator automatically identifies the artifactual trial as outlier and reduces it's influence by assigning it a weight close to zero.

4 Discussion

This work presented novel concepts and methods to improve BCI robustness and to alleviate the nonstationarity problem which is one of the main challenges preven-

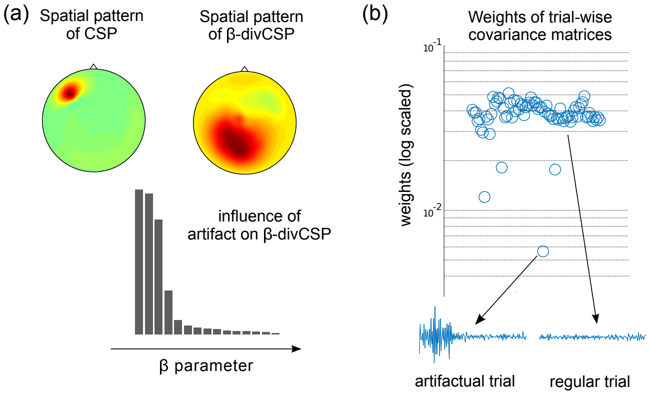

Figure 3: (a) Spatial patterns computed with CSP and beta divergence CSP. The former one shows that CSP is unable to extract relevant activity due to artifacts in the data. Using beta divergence reduces the influence of these artifacts. (b) Trial-wise weights ψ_β show that the proposed covariance matrix estimator strongly penalizes the artifactual trial.

ting current BCI technology to be widely applied in realistic environments. The proposed methods extend the relevant prior art algorithms not only in that they enforce stationarity of extracted features and increase the robustness against outlier samples and outlier trials, but they also provide a framework for transferring information about the expected nonstationarities in the data between subjects [13]. The formulation of spatial filter computation as a divergence maximization problem is a key contribution, because it easily allows one to robustify the algorithm against artifacts, to incorporate data from other subjects into the optimization process and to enforce different types of invariances on the extracted features. Similar to the kernel trick [14] which revolutionized the field of machine learning, because it makes linear algorithms to be non-linear, it is hoped that the here introduced “divergence trick” will have a similarly big impact in the future.

The presented concepts and methods are applicable beyond spatial filtering and BCI. Popular projection algorithms such as Principal Component Analysis (PCA) or Canonical Correlation Analysis (CCA) can be formulated in terms of divergences, thus may directly benefit from the divergence trick. Research fields such as natural sciences use these algorithms for different purposes (e.g., visualization, dimension reduction, feature extraction) and may benefit from robust or stationary alternatives. The estimation of parameters from structured data is also an important topic in various fields. For instance, in the medical domain when analyzing data from clinical multi-site studies the definition of robustness becomes important, because it may refer to the individual data sample or to all data from a specific site. Also date is an often used grouping variable which allows for a multi-scale (e.g., day, month, year) definition of robustness which is naturally included in the parameter estimation approach described in this work.

Literature

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

[2] K.-R. Müller, C. W. Anderson, G. E. Birch. *Linear and Non-Linear Methods for Brain-Computer Interfaces.* IEEE Trans. Neural Syst. Rehabil. Eng., 11(2):165-169, 2003.

[3] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.-R. Müller. *Optimizing Spatial filters for Robust EEG Single-Trial Analysis.* IEEE Signal Process. Mag., 25(1):41–56, 2008.

[4] 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.* Proc. IEEE, 103(9):1507-30, 2015.

[5] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe. *Stationary Common Spatial Patterns for Brain-Computer Interfacing.* J. Neural Eng., 9:026013, 2012.

[6] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe. *Robust Spatial Filtering with Beta Divergence.* Adv. in NIPS, 1007-15, 2013.

[7] S. Amari and H. Nagaoka. *Methods of information geometry.* vol. 191 of Transl. of Math. Monogr., American Math. Soc., 2000.

[8] W. Samek, M. Kawanabe, and K.-R. Müller. *Divergence-based Framework for Common Spatial Patterns Algorithms.* IEEE Rev. in Biomed. Eng., 7:50-72, 2014.

[9] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre. *Robust Common Spatial filters with a Maxmin Approach.* Neural Comp., 26(2):1-26, 2014.

[10] W. Samek and M. Kawanabe. *Robust Common Spatial Patterns by Minimum Divergence Covariance Estimator.* Proc. IEEE ICASSP, 2059-62, 2014.

[11] S. Eguchi and Y. Kano. *Robustifying maximum likelihood estimation.* Tokyo Institute of Statistical Mathematics, Technical Report, 2001.

[12] 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, 51(4):1303–1309, 2010.

[13] W. Samek, F. C. Meinecke, and K.-R. Müller. *Transferring Subspaces Between Subjects in Brain-Computer Interfacing.* IEEE Trans. Biomed. Eng., 60(8):2289-98, 2013.

[14] B. E. Boser, I. M. Guyon, and V. N. Vapnik. *A training algorithm for optimal margin classifiers.* Proc. Workshop on COLT, 144-152, 1992.

Dr. Wojciech Samek is head of the Machine Learning group at Fraunhofer Heinrich Hertz Institute and associated researcher at the Berlin Big Data Center. He received the Diploma degree in Computer Science from Humboldt-Universität zu Berlin in 2010 and the Ph.D. degree from the Technische Universität Berlin in 2014. He was scholar of the Studienstiftung des deutschen Volkes and a Ph.D. Fellow at a DFG Research Training Group and the Bernstein Center for Computational Neuroscience Berlin. During his studies he had research stays at the University of Edinburgh, U.K., the NASA Ames Research Center, Mountain View, CA, USA, and ATR International, Kyoto, Japan. His research interests include machine learning, neural networks, signal processing and computer vision.

Address: Fraunhofer Heinrich Hertz Institute, Department of Video Coding & Analytics, D-10587 Berlin, E-Mail: wojciech.samek@hhi.fraunhofer.de