TACKLING NOISE, ARTIFACTS AND NONSTATIONARITY IN BCI WITH ROBUST DIVERGENCES

Wojciech Samek, Member, IEEE

Fraunhofer Heinrich Hertz Institute
Machine Learning Group
Einsteinufer 37, 10587 Berlin, Germany
wojciech.samek@hhi.fraunhofer.de

Klaus-Robert Müller, Member, IEEE

Berlin Institute of Technology
Machine Learning Group
Marchstr. 23, 10587 Berlin, Germany
klaus-robert.mueller@tu-berlin.de

ABSTRACT

Although the field of Brain-Computer Interfacing (BCI) has made incredible advances in the last decade, current BCIs are still scarcely used outside laboratories. One reason is the lack of robustness to noise, artifacts and nonstationarity which are intrinsic parts of the recorded brain signal. Furthermore out-of-lab environments imply the presence of external variables that are largely beyond the control of the user, but can severely corrupt signal quality. This paper presents a new generation of robust EEG signal processing approaches based on the information geometric notion of *divergence*. We show that these divergence-based methods can be used for robust spatial filtering and thus increase the systems' reliability when confronted to, e.g., environmental noise, users' motions or electrode artifacts. Furthermore we extend the divergence-based framework to heavy-tail distributions and investigate the advantages of a joint optimization for robustness and stationarity.

***Index Terms*—** Brain-Computer Interfacing, Common Spatial Patterns, Nonstationarity, Robustness

1. INTRODUCTION

Since the advent of the technology, Brain-Computer Interfacing (BCI) [1] aims to reliably translate recorded brain signals, e.g., EEG, into control commands for a computer. Despite improvements over the last decades, current BCIs are still mostly used inside laboratories. One major limitation preventing the prevalence of this promising technology into everyday life of, e.g., patients, is the lack of robustness and reliability. However, tackling noise, artifacts and nonstationarity still poses a large challenge in practice.

In this work we mainly focus on motor-imaginary BCIs [1]. Motor imagery, i.e., the imagination of hands, feet or tongue movements, is a popular paradigm to voluntarily induce a set of mental states which can be distinguished by a computer. Neurophysiologically, motor imagery tasks alter the sensorimotor rhythms (SMRs) over specific spatial locations in the sensorimotor cortex. Recognizing the exact spatial localitions of SMR modulations is extremely important for reliable BCI communication. A popular method for this task is Common Spatial Patterns (CSP) (e.g., [2] [3]). Mathematically, CSP solves a generalized eigenvalue problem which can be computed efficiently. Since the original version of CSP is sensitive to artifacts and nonstationarity, more robust CSP variants have been proposed, e.g., [4] [5] [6] [7] [8] [9] among others. Recently, a generic divergence-based spatial filtering framework [10] has been developed containing many of these CSP variants as special case. Following this line of research the current work discusses and evaluates the use of divergence-based methods as generic signal processing tools for BCI.

2. NOISE, ARTIFACTS & NONSTATIONARITY

2.1. The Challenge of EEG Analysis

The human brain is a dynamical system with many neuronal processes (e.g., controlling body functions, processing external input, responsible for cognition) running in parallel at each moment. Motor imagery affects only a tiny fraction of the neuronal activity, but changes in each of these processes may alter the overall EEG signal. Due to the presence of large amounts of task-unrelated activity, EEG signals are intrinsically noisy and nonstationary which makes BCI communication be a challenging task.

Noise sources can be internal (i.e., mental states) or external, and they introduce nonstationarities by either altering brain activity directly or affecting the sensors (i.e., electrodes) that are used to register brain activity. Internal factors such as

*This work was supported by the by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115) and by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education.

fatigue, changes in attention, task involvement or the strategy to perform motor imagery are reflected in the recorded signal and may largely affect the features used for classification. In addition, external factors such as changes in impedance (e.g., electrode gel dries out), sensory input, experimental conditions etc. may also contribute to a large part to the observed nonstationarity. In out-of-lab environments the number of external sources distracting the subject are much larger, e.g., the user is constantly faced with stimuli that are not relevant to the task he is involved in, such as radio or TV, traffic noise, and background conversations. Even though not task-relevant, the sensory input is processed by the brain and it may affect mental states, e.g., by distracting attention from the task at hand. Note that changes can have various time scales.

Artifacts due to small-scale (e.g., eye blinks) and large-scale movements (e.g., talking, walking, moving a wheelchair) are also common in natural environments. However, not only do muscles themselves produce electromagnetic fields that project into EEG sensors, the movement preparation itself induces changes in brain activity and bulk movements may involve mechanical artifacts (e.g., movement of the electrodes). Home environments usually contain a large number of electronic devices such as kitchen appliances, TV, and computer, each producing their own electromagnetic field. This may pose an additional challenge to machine learning.

2.2. Impact on Performance

BCI performance usually suffers when the recorded EEG signal is very noisy, affected by artifacts or highly nonstationary. Figure 1 depicts the impact on performance. The left panel shows gray lines representing the amplitude envelope of single motor imagery trials. One can see that the decrease in power (i.e., ERD effect) can be hardly detected in single trials due to the large amount of noise. By averaging (but also by spatial filtering) one significantly increases the signal-to-noise ratio, so that the power decrease becomes evident (black solid line).

The middle panel shows the adverse effect of artifacts on BCI training. Here the spatial filter computation is affected by an artifact. Since the standard method for computing spatial filters, CSP, maximizes the variance ratio between two motor imagery classes in a naive data-driven manner, it is vulnerable to artifacts and overfitting if high power artifacts (e.g., loose electrodes, eye blinks) affect both classes in a slightly different manner. Sometimes single artifactual trials may even lead to degenerated CSP patterns as shown in the figure. Possible remedies are data cleaning or the use of robust algorithms.

The right panel depicts feature nonstationarity, i.e., a time dependency of the feature distribution [4]. As mentioned above EEG signals are intrinsically nonstationary because different processes are active in the brain at different times and the sensory input also changes constantly. This nonstationarity may adversely affect performance as most standard classifiers, e.g., Linear Discriminant Analysis (LDA), assume that data are sampled from a stationary distribution. Possible remedies are the use of (unsupervised) adaptation techniques or extraction of stationary features.

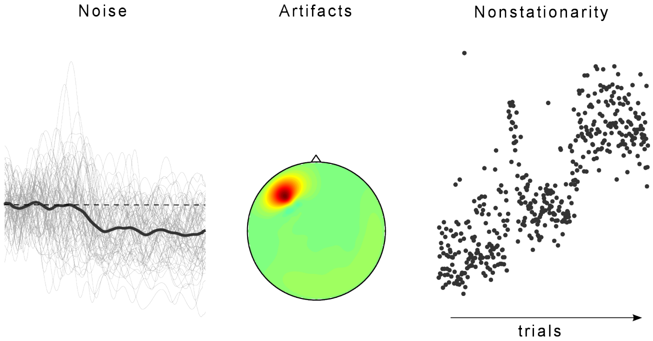

Fig. 1: Left: Gray lines represent amplitude envelopes of single trials; the black solid line depicts the average envelope. Middle: An artifact in electrode FC5 leads to a degenerated CSP pattern and poor classification performance. Right: The nonstationary nature of EEG may lead to a feature distribution which changes with time. Classifiers which assume stationarity will perform poorly in this case.

In practice, the presence of artifacts does not necessarily lead to poor performance. On the contrary, sometimes artifacts such as tiny eye or muscle movements are unconsciously used as basis for controlling a BCI. This is in general undesirable because the aim of a BCI is to rely only on brain activity for control. If a BCI system is controlled by muscle artifacts, it may be of no use in patients who, e.g., are not able to move.

We propose to tackle noise, artifacts and nonstationarity by using a divergence-based spatial filtering approach with a robust divergence and a regularization term which penalizes time dependency of the feature distribution.

3. DIVERGENCE-BASED SPATIAL FILTERING

3.1. Common Spatial Patterns

Spatial filtering is a common way to enhance the signal-to-noise ratio in motor imagery BCIs. A well-suited method for computing spatial filters is Common Spatial Patterns (CSP) [2] [3]. The CSP filters maximize the variance ratio between two motor imagery conditions, thus focus on the synchronization and desynchronization effects (ERS/ERD) induced by motor imagery. Mathematically, spatial filters w_i are computed by solving a generalized eigenvalue problem

$$\Sigma_1 w_i = \lambda_1 \Sigma_2 w_i$$

where Σ_1 and $\Sigma_2 \in \mathbb{R}^{\mathrm{DxD}}$ are the average covariance matrices of motor imagery class 1 and 2. The obtained spatial filters $W = [w_1, w_2, ..., w_D]$ are sorted according to their contributing discriminative qualities [10].

3.2. Spatial Filtering Framework

Recently, the authors of [10, 11] showed that spatial filter computation can be cast into a divergence framework. This formulation has the advantage that it embeds the CSP algorithm into a mathematical framework which allows to propose novel CSP variants by applying the "divergence trick" [10], i.e., by keeping the mathematical formulation of the problem but using other divergences with different properties.

In the divergence framework, robustness can be achieved by decomposing the divergence between the average class distributions into the sum of trialwise divergences and limiting the influence of single (potentially outlier) terms. This changes the objective function to

$$\mathcal{L}_{rob}(V) = \\ \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{D}_{\beta}\left(\mathcal{N}\left(0, V^{\top}\Sigma_1^i V\right) || \, \mathcal{N}\left(0, V^{\top}\Sigma_2^j V\right)\right)$$

where $\tilde{D}_{\beta}(p\,||\,q) = D_{\beta}(p\,||\,q) \; + D_{\beta}(q\,||\,p)$ stands for the symmetric Beta divergence, $\mathcal{N}(0, \Sigma)$ denotes the Gaussian distribution with mean 0 and covariance Σ, n represents the number of trials per motor imagery condition and Σ_c^i stands for the estimated covariance matrix of condition c and trial i. Note that the divergence formula used in [10, 11] only compares the ith trial of one class with the ith trial of the other class. We refer to the above objective function as i vs. j and to the objective function used in [10, 11] as i vs. i.

In order to tackle the nonstationarity problem, a regularization term can be included into the divergence framework. One way to measure nonstationarity is by using the average divergence between the data distribution of individual trials and the overall data distribution of a class. Since we aim to minimize nonstationarity we need to subtract this regularization term from the objective function, i.e.,

$$\mathcal{L}_{robstat}(V) = (1-\lambda)\mathcal{L}_{rob}(V) \; - \\ \lambda\frac{1}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n}D_{\beta}\left(\mathcal{N}\left(0, V^{\top}\Sigma_c^i V\right) \;\; || \;\, \mathcal{N}\left(0, V^{\top}\Sigma_c V\right)\right)$$

For $\lambda > 0$ the solution is regularized towards stationarity. Note that although the authors of [10] used the same regularization term, they did not jointly tackle the robustness and nonstationarity problem.

3.3. Heavy-Tail Model

An alternative to using Beta divergence for robust spatial filter computation is the usage of heavy-tailed probability models. For instance, the authors of [12] proposed a robust variant of the CSP algorithm based on a Student's t-distribution model. In this paper we investigate the use of Student's t-distribution as an alternative to the Gaussian model in the divergence framework. The natural equivalent to KL divergence for this class of distributions is the so called t-divergence [13] defined as

$$D_t(p\,||\,\tilde{p}) = \int q(x)\log_t p(x) - q(x)\log_t \tilde{p}(x)dx$$

with $\log_t(x) = \begin{cases} \log(x) & \text{if t} = 1 \\ \frac{x^{1-t}-1}{1-t} & \text{otherwise} \end{cases}$
For the zero-mean Student's t-distribution model

$$\mathcal{S}\left(x; 0, \bar{\Sigma}, \nu\right) = \\ \frac{\Gamma((\nu+d)/2)}{(\pi\nu)^{d/2}\Gamma(\nu/2)|\,\bar{\Sigma}\,|^{1/2}}(1 + x^{\top}(\nu\Sigma)^{-1}x)^{-(\nu+d)/2}$$

the objective function to be maximized

$$\mathcal{L}_t(V) = \\ \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{D}_t\left(\mathcal{S}\left(0, V^{\top}\Sigma_1^i V, \nu\right) || \, \mathcal{S}\left(0, V^{\top}\Sigma_2^j V, \nu\right)\right)$$

has explicit form (see [13])

$$\mathcal{L}_t(V) = \frac{\zeta}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(|\,\bar{\Sigma}_1^i\,|^{\frac{t-1}{2}}\left(d - \operatorname{tr}\left((\bar{\Sigma}_1^i)^{-1}\,\bar{\Sigma}_2^j\right)\right)\right. \\ \left. +|\,\bar{\Sigma}_2^j\,|^{\frac{t-1}{2}}\left(d - \operatorname{tr}\left((\bar{\Sigma}_2^j)^{-1}\,\bar{\Sigma}_1^i\right)\right)\right)$$

with constant $\zeta = \frac{1}{d(t-1)-2}\left(\frac{\Gamma\left(\frac{1}{t-1}\right)}{\left(\pi\left(\frac{2}{t-1}-d\right)\right)^{\frac{d}{2}}\Gamma\left(\frac{1}{t-1}-\frac{d}{2}\right)}\right)^{1-t}$ and $\bar{\Sigma} = V^{\top}\Sigma V \in R^{d\times d}$ being the projected covariance matrix, $\Gamma(\cdot)$ representing the Gamma function and $t = 1 + \frac{2}{\nu+d}$ being a free parameter. One can show that for $t \to 1$ this model is equivalent to using KL divergence with Gaussians, i.e.,

$$\lim_{t\to 1}\mathcal{L}_t(V) \;\; = \;\; \lim_{\beta\to 0}\mathcal{L}_{rob}(V)$$

For $t > 1$ the distribution has heavier tails.

A detailed derivation of the heavy-tail model and an implementation of all divergence-based algorithms are available at `www.divergence-methods.org`.

4. EXPERIMENTAL EVALUATION

4.1. Data Set & Setup

We use the Vital BCI data set [14] for experimental evaluation. It contains EEG recordings from 80 healthy subjects performing motor imagery tasks with the left and right hand or with the feet. It consists of one calibration session (75 trials per class) and one test session (150 trials per class) with visual feedback (cursor moving on the screen), both recorded on the same day using a multichannel EEG amplifier (BrainAmp DC by Brain Products GmbH) and 118 Ag/AgCl electrodes. The following preprocessing steps were applied to the data. We manually select 62 electrodes densely covering the motor cortex and filter the data in a subject specific frequency range [2] with a 5th order Butterworth filter. The time segment used for classification is also subject specific [2]. For simplicity and computational efficiency reasons we only use two spatial filters and fixed parameters $\lambda = \beta = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$ and $t = 1.05, 1.1, 1.2, 1.3, 1.4, 1.5$. The divergence-based algorithms are initialized with the CSP solution and max. 100 iterations are performed. We refer the reader to [10] for more algorithmic details on the divergence framework.

4.2. Results

The mean and median error rates of the different spatial filtering methods are summarized in Table 1. The top row displays error rates when using the signal recorded at fixed electrode locations (C3, C4 and Cz) as basis for classification. Since spatial filtering increases the signal-to-noise ratio, the error rates of CSP and the divergence-based methods are much lower than the error rates of this simple "raw signal" approach. This result demonstrates that spatial filtering effectively reduces noise. The third and fourth rows display the error rates of the robust Beta divergence-based spatial filtering method. Note that the asterisks indicate significant improvements over the CSP baseline (number of asterisks corresponds to significance levels 0.05%, 0.01% and 0.001%). The left columns show the significance test results when evaluating the mean differences using a one-sided t-test whereas the right columns display the results of the one-sided Wilcoxon sign-rank test (i.e., median differences).

For β values 0.2 and 0.3 the robust (i vs. i) method significantly outperforms CSP. This result demonstrates that (i) artifacts negatively affect the CSP algorithm, (ii) one can significantly improve the quality of the computed spatial filters by using a robust divergence and (iii) error rates can be decreased even when using a fixed β parameter for all subjects. Our results also indicate that computing the filters in a i vs. i manner is superior to the i vs. j strategy. This is likely due to the nonstationarity of the signal. Considering the divergence between trials from the beginning and end of the training session may negatively affect the results because some of the difference may be due to nonstationarity and be neurophysiologically meaningless. In i vs. i, the divergence is only computed between close (in terms of recording time) trials, thus this problem does not occur. The average computation time (100 iterations) was 28 sec for i vs. i and 1040 sec for i vs. j.

The next two rows of Table 1 show that the joint optimization of robustness and stationarity further improves the results. Note that we neither vary the time scale of potential changes (i.e., chunk size, see [10]) nor do we select the optimal β and λ parameters. We expect further improvement in classification accuracy when carefully adjusting these free parameters. Although the average improvement is rather small, for particular subjects it can be up to 30%. For instance, the error rates of subject 22 decreases from 49.3 % to 26 % . For this subject only the combination of robust divergences and regularization yields an improvement, applying each of these strategies separately does not decrease error rates.

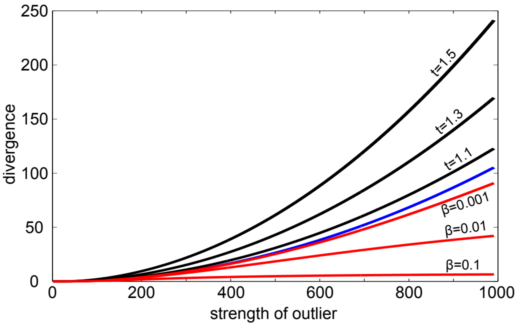

Fig. 2: Effect of a single outlier on the KL, Beta and t-divergence.

The heavy-tail distribution model does improve performance over the CSP baseline in our experiments. The reason for this is that we applied a heavy-tail distribution with fixed parameter t to all trials, despite the fact that trials were affected by outliers quite differently. Figure 2 shows the divergence between data x sampled from a zero-mean Gaussian distribution with variance one and the same samples with an additional outlier included $y = [x \;\; \xi]$. If we do not add the outlier ξ to the data (strength 0), then the divergence is zero because $y = x$. However, when increasing the strength of the outlier (i.e., it's value), the divergence between x and y increases, i.e., both distributions become more and more dissimilar. For KL divergence (blue line) this increase is substantially larger than for Beta divergence (red lines), but smaller than for t-divergence (black lines). This is because the heavy-tail distribution (with large t) fits the "clean + outlier" data y much better than a Gaussian distribution, but is a bad model for the "clean" data x which does not show heavy tails. Since we use a fixed t parameter for all trials, i.e., do not distinguish between clean trials which should be modeled by a Gaussian and artifactual trials which benefit from a heavy-tailed distribution model, we are not able to robustify the spatial filtering against artifacts using the Student's t-distribution model. Thus, in practice we recommend to use Beta divergence when individual outliers are present in the data and use the t-divergence model when most trials follow a heavy-tailed distribution.

5. CONCLUSION & FUTURE WORK

In this paper we investigated the use of divergence methods for Brain-Computer Interfacing. Using the recently proposed divergence-based spatial filtering framework we could significantly increase classification accuracy and reduce the influ-

Table 1: Error rates for 80 subjects of the Vital BCI Data Set and different parameters β, λ and t.

Method	Mean [%]						Median [%]					
λ, β	0.05	0.1	0.2	0.3	0.4	0.5	0.05	0.1	0.2	0.3	0.4	0.5
t	1.05	1.1	1.2	1.3	1.4	1.5	1.05	1.1	1.2	1.3	1.4	1.5
raw signal	36.3						37.0					
CSP	29.3						27.3					
robust (i vs. i)	29.5	28.9	28.3*	28.3*	28.5	28.8	28.6	28.7	27.7	26.7*	26.7	26.5
robust (i vs. j)	29.5	28.8	29.1	29.1	29.3	29.7	26.6	26.0	26.3	26.5	26.6	27.5
robust & stationary (i vs. i)	29.3	28.5*	27.9**	27.9**	28.7	29.0	28.6	28.9*	26.2*	26.2*	27.0	26.8
robust & stationary (i vs. j)	29.0	28.5	28.7	28.3*	28.0**	28.4*	26.5	26.0	26.3	26.2	26.2**	26.7*
heavy-tail (i vs. i)	31.0	30.8	30.8	30.8	30.8	31.1	29.8	29.7	29.8	30.7	31.3	32.0
heavy-tail (i vs. j)	30.3	30.1	29.1	29.1	29.4	29.9	31.0	30.8	26.5	27.3	28.2	27.6

ence of artifacts and nonstationarity. The heavy-tail distribution model could not improve performance over the CSP baseline, because this model does not work well when trials are affected by outliers very differently. Since a heavy-tailed distribution is a suboptimal model for clean data, the t parameter should be determined for every trial individually. Beta divergence implicitly applies this kind of individual weighting of trials, thus should be preferred when data is contaminated by artifacts. In future work we will continue investigating the use of robust divergences in the context of out-of-lab BCI. Furthermore we plan to extend the scope of application of divergence-based methods from mere spatial filter computation to direct and adaptive classification. Finally we plan to compare the divergence methods with the Riemannian Distance [15], the natural metric for covariance matrices, in terms of performance and computational efficiency.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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, 2013.

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

[8] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject eeg classification," *Neural Networks*, vol. 57, pp. 39–50, 2014.

[9] 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.

[10] 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.

[11] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Adv. in NIPS*, 2013, pp. 1007–1015.

[12] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A probabilistic framework for learning robust common spatial patterns." *Proc. of the IEEE Eng. Med. Biol. Soc.*, vol. 2009, pp. 4658–61, 2009.

[13] N. Ding, Y. Qi, and S. Vishwanathan, "t-divergence based approximate inference," in *Adv. in NIPS*, 2011, pp. 1494–1502.

[14] B. Blankertz et al., "Neurophysiological predictor of SMR-based BCI performance," *NeuroImage*, vol. 51, no. 4, pp. 1303–1309, 2010.

[15] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by riemannian geometry," *IEEE Trans. on Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, 2012.