Information Geometry meets BCI

Spatial filtering using divergences

Wojciech Samek[†], *Member, IEEE,*
[†]Department of Computer Science
Berlin Institute of Technology (TU Berlin)
Berlin, Germany
wojciech.samek@tu-berlin.de

Klaus-Robert Müller[†‡], *Member, IEEE*
[‡]Department of Brain and Cognitive Engineering
Korea University
Seoul, Republic of Korea
klaus-robert.mueller@tu-berlin.de

***Abstract*—Algorithms using concepts from information geometry have recently become very popular in machine learning and signal processing. These methods not only have a solid mathematical foundation but they also allow to interpret the optimization process and the solution from a geometric perspective. In this paper we apply information geometry to Brain-Computer Interfacing (BCI). More precisely, we show that the spatial filter computation in BCI can be cast into an information geometric framework based on divergence maximization. This formulation not only allows to integrate many of the recently proposed CSP algorithms in a principled manner, but also enables us to easily develop novel CSP variants with different properties. We evaluate the potentials of our information geometric framework on a data set containing recordings from 80 subjects.**

Keywords: Information Geometry, Divergences, Brain-Computer Interfacing, Common Spatial Patterns

I. Introduction

Brain-Computer Interface (BCI) systems [1] [2] translate recorded brain signals, e.g. EEG, into control commands for a computer. A crucial step in this translation process is the extraction of relevant brain activity from high-dimensional electroencephalographic recordings [3] [4] [5]. In motor imagery based BCIs a very popular algorithm for this feature extraction step is Common Spatial Patterns (CSP) (e.g. [6] [7]). Spatial filters computed with CSP allow to discriminate between different mental states induced by motor imagery as they focus on the ERS/ERD effect. Many extensions of CSP have been proposed robustifying the solution against artifacts (e.g. [8] [9]), aiming for stationarity of the features (e.g. [10] [11]) or incorporating data from other sessions/subjects (e.g. [12] [13]).

Recently, the authors of [14] [15] showed that Common Spatial Patterns can be cast into an information geometric framework. Information geometry [16] is a branch of mathematics that studies questions of probability theory by means of differential geometry. A key concept are so-called *divergence functions*. A divergence function [17] measures the discrepancy between two points in some parameter space, e.g. the points may represent probability distributions lying on a statistical manifold. Note that divergences are always positive (or zero), but do not have to be symmetric. Various machine learning (ML) algorithms can be formulated in terms of divergences [18] [19] [20] and optimized using methods from differential geometry

In this paper we review the recently proposed information geometric spatial filtering framework and comment on the potentials of information geometry for BCI. An implementation of our framework is available at `http://www.divergence-methods.org`.

II. Information Geometric Spatial Filtering

Spatial filtering is a crucial step in motor imagery BCI as it reduces dimensionality of the data while increasing its signal-to-noise ratio. Spatial filters $\mathbf{w}$ computed by CSP are well-suited to decode imagined movements as they maximize/minimize the variance ratio between the two motor imagery classes (ERD/ERS effect). The filters $\mathbf{w}_1 \ldots \mathbf{w}_d$ that best discriminate between the classes can be computed as

$$\mathbf{w}_i = \underset{\mathbf{w}}{\operatorname{argmax}} \left(\max \left\{ \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}}, \frac{\mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}} \right\} \right) \quad (1)$$

$$\text{s.t. } \mathbf{w}_i (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \mathbf{w}_j = 0 \text{ for } j \neq i \quad (2)$$

where $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2 \in \mathrm{R}^{D \times D}$ are the average covariance matrices of condition 1 and 2, respectively. With this definition the CSP filters are ordered according to their ability to capture the differences between both classes.

One can prove (see [15]) that the spatial filters computed by CSP have a special property, namely they span a subspace with maximum symmetric Kullback-Leibler (KL) divergence between the distributions of both classes (estimated as zero mean multivariate Gaussians). More formally, we can say

$$\operatorname{span}(\mathbf{W}) \quad = \quad \operatorname{span}(\mathbf{V}^*), \quad (3)$$

with span($\mathbf{W}$) being the subspace spanned by the top d CSP filters $\mathbf{W} \in \mathrm{R}^{D \times d}$ and $\mathbf{V}^* \in \mathrm{R}^{D \times d}$ (decomposable into whitening and orthogonal projection) maximizing the symmetric KL divergence between Gaussians

$$\tilde{D}_{kl} \left(\mathcal{N} \left(\mathbf{0}, \mathbf{V}^\top \mathbf{\Sigma}_1 \mathbf{V} \right) \;\; || \;\; \mathcal{N} \left(\mathbf{0}, \mathbf{V}^\top \mathbf{\Sigma}_2 \mathbf{V} \right) \right). \quad (4)$$

The symmetric KL divergence between distribution p(x) and q(x) is defined as

$$\int p(x) \log \frac{p(x)}{q(x)} dx \quad + \quad \int q(x) \log \frac{q(x)}{p(x)} dx. \quad (5)$$

A geometrical interpretation of the theorem is given in Fig. 1. The left plot depicts the manifold of D×D covariance matrices

(symmetric, positive definite matrices) and the red and blue dots represent the class-covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. The

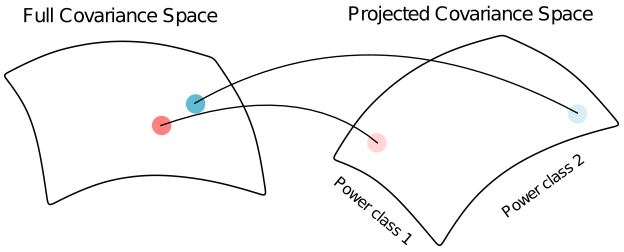

Fig. 1: The left plot shows the class-covariance matrices, represented as two points, lying on the manifold of D×D covariance matrices. The spatial filtering step projects the points to the submanifold of d×d covariance matrices while maximizing the discrepancy (symmetric KL divergence) between them.

"distance" (measured as symmetric KL divergence) between these points is relatively small in this space as the covariance structures of both classes are largely affected by common, motor imagery unrelated activity. The spatial filters computed by CSP project the data to a subspace. The projected covariance matrices now lie on the manifold of d×d symmetric, positive definite matrices. The CSP criterion maximizes the symmetric Kullback-Leibler divergence between the projected points, thus the projected class-covariance matrices are as far apart as possible. In other words the spatial filtering step (in the optimal case) focuses on BCI related activity in the data that is expected to be different for the two motor imagery classes.

Using this information geometric formulation of CSP one can easily regularize the solution by introducing a penalty term Δ. Note that regularizing the solution is very useful in practice as the extracted subspace with maximum divergence between both classes may be the result of overfitting, contain artifacts or BCI unrelated discriminative activity (e.g. eye movements) or be very subject-specific. The penalty term can be easily incorporated into our framework when measured as (sum of) divergence(s). The objective function of the proposed regularized information geometric CSP method is

$$\mathcal{L}(\mathbf{V}) = \underbrace{(1-\lambda)\tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}_1\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}_2\mathbf{V}\right)}_{\text{CSP Term}} - \underbrace{\lambda\Delta}_{\text{Reg. Term}} \quad (6)$$

Thus the goal is to find a subspace where the projected covariance matrices are as far apart as possible while minimizing the penalty. The regularization parameter λ trades off the influence of the CSP objective function and the regularization term. Table I shows different penalty terms.

The first term, called Δ-WS (Within-Session), regularizes the solution towards stationarity (c.f. [21] [10] [11]) by minimizing the deviations of the trial-wise covariance matrices $\mathbf{\Sigma}^i_c$ from the class average $\mathbf{\Sigma}_c$. Another regularization term, called Δ-BS (Between-Session) is shown in the second row and minimizes the difference between projected training $\mathbf{\Sigma}^k_{tr,c}$ and test covariance matrices $\mathbf{\Sigma}^k_{te,c}$ by using data from other subjects *k* (as done in [22]). The third penalty scheme, called Δ-AS (Across-Subject), regularizes the projected covariance matrices of the subject of interest $\mathbf{\Sigma}^\ell_{tr,c}$ towards the covariance matrices of other subjects $\mathbf{\Sigma}^k_{tr,c}$, whereas the last regularization matrix, called Δ-MS (Multi-Subject), solves the CSP problem jointly for all subjects i.e. it aims to make the extracted spatial filters more subject-independent. Many other regularization schemes are possible and can be easily incorporated into the framework.

TABLE I. INFORMATION GEOMETRIC SPATIAL FILTERING METHOD

Regularization term $\mathbf{\Delta}$
Within-Session (WS) $\mathbf{\Delta} = \frac{1}{2N}\sum_{c=1}^{2}\sum_{i=1}^{N} D_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}^i_c\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}_c\mathbf{V}\right)$
Between-Session (BS) $\mathbf{\Delta} = \frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K} \tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}^k_{tr,c}\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}^k_{te,c}\mathbf{V}\right)$
Across-Subject (AS) $\mathbf{\Delta} = \frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K} \tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}^\ell_{tr,c}\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}^k_{tr,c}\mathbf{V}\right)$
Multi-Subject (MS) $\mathbf{\Delta} = -\frac{1}{K}\sum_{k=1}^{K} \tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}^k_1\mathbf{V} \;||\; \mathbf{V}^\top\mathbf{\Sigma}^k_2\mathbf{V}\right)$

III. Robustness Through Beta Divergence

An advantage of the information geometric formulation of the CSP algorithm is the unified view on the regularization. Since all terms in the objective function have a common interpretation, namely as divergences between two zero mean Gaussian distributions, it is easy to compare and combine different divergence-based CSP variants.

Another advantage that is even more important is the generic formulation of the spatial filtering problem in terms of divergence maximization. Similar to the "kernel-trick" that is often applied to perform classification in other spaces [23], [24], we can change the geometry and the properties of the solution (without changing the mathematical formulation) by considering different divergences. Other divergences induce a geometry that may be advantageous for BCI application e.g. because it is robust to outliers. A divergence that has been used in machine learning, e.g. in Independent Component Analysis [20], is the so-called beta divergence [25] [26]. Beta divergence D_β (p(x) || q(x)) between distributions p(x) and q(x) has been proposed in [25] [26] and is defined (for β > 0) as

$$\frac{1}{\beta}\int (q^\beta(x) - p^\beta(x))q(x)dx \quad (7)$$
$$- \frac{1}{\beta+1}\int (q^{\beta+1}(x) - p^{\beta+1}(x))dx.$$

Beta divergence is known to be robust to outliers as it is related to the Ψ-likelihood principle in statistics [25]. By applying beta divergence to the CSP problem we can decrease the influence of outliers on the spatial filter computation (see [14]) or robustly compute the regularization terms, e.g. in the Δ-MS case we rather prefer subspaces with relatively high divergences for many subjects over subspaces with very high divergences for few subjects; a simple averaging without downweighting (β = 0) may give preference to the latter case.

IV. Experimental results

In this section we evaluate the across-subject (AS) regularization scheme. The idea of finding a subspace where the class-distributions are similar across subjects is a novel regularization strategy. We compare the performance results to standard CSP and to the method proposed in [27]. We denote the method of [27] as covCSP because it applies CSP with

regularized covariance matrices, i.e. the covariance matrix of subject ℓ and class c is estimated as

$$\hat{\mathbf{\Sigma}}^{\ell}_{tr,c} = (1-\alpha)\mathbf{\Sigma}^{\ell}_{tr,c} + \alpha\frac{1}{K}\sum_{k=1}^{K}\mathbf{\Sigma}^{k}_{tr,c} \qquad (8)$$

We use a data set [28] containing EEG recordings from 80 subjects performing motor imagery tasks with the left and right hand or with the feet. The data contain a calibration session and a (1D visual) feedback session. We select the two best motor imagery classes for each subject, resulting in 150 calibration and 300 feedback trials. We manually select 62 electrodes, apply band pass filtering (8-30 Hz) with a 5-order Butterworth filter and extract a time segment from 750ms to 3500ms after the trial start. We extract six spatial filters and apply the LDA classifier. The parameters λ and β are selected from the sets {0, 0.1, 0.2, . . ., 1} and {0, 0.5, 1} by 5-fold cross-validation on the calibration data using minimal error rate as selection criterion. We use the deflation algorithm (see [15]) for optimizing the objective function in Eq. (6), i.e. we extract the spatial filters sequentially (we do not extract the subspace in one run). The parameter α is selected from the set {0, 10^{-5} , . . . , 10^{-1} , 0.2, . . . , 0.9, 1} by using 5-fold cross-validation with same selection criterion.

The scatter plots in Fig. 2 compare the error rates of CSP and covCSP to our divergence-based Δ-AS algorithm. Each square represents the error rate of one of the 80 subjects and the p-value of the one-sided Wilcoxon sign rank test is shown in the lower right corner. The error rates of our method are shown on the y-axis, i.e. if a square is below the dashed line then our method outperforms the baseline. Note that $p < 0.05$ indicates significant performance gain for our method.

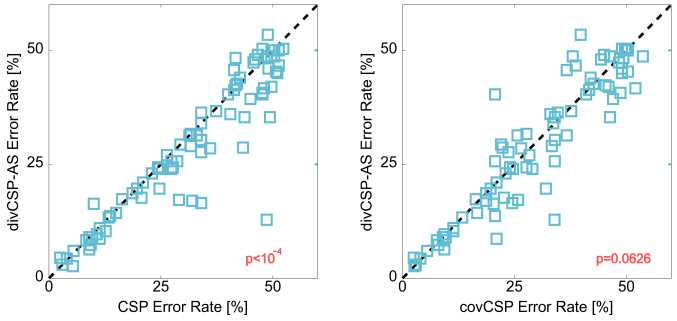

Fig. 2: Scatter plots comparing the error rates of the proposed divergence-based CSP method with AS regularization to CSP and covCSP. Each square represents one subject and if the square is below the solid line then our method outperforms the baseline for this subject. The p-value of the one-sided Wilcoxon signed rank test is shown in the right bottom corner.

One can clearly see that utilizing information from other subjects significantly improves the quality of the spatial filters. Our method significantly outperforms the CSP baseline and (almost significantly) decreases error rates when comparing the results to covCSP. This suggests that regularizing the divergences gives better performance than regularizing the covariance matrices (as done by covCSP) prior to CSP computation. In other words it seems that regularization in the projected space is more effective than regularization in the full covariance space. Note that the covCSP algorithm applies regularization to the covariance matrices inside the divergence function (i.e. prior to CSP computation) whereas our method applies regularization to the divergences (i.e. after the non-linear divergence function was applied).

Let us consider the single filter case. We can write the covCSP objective as

$$\frac{\mathbf{v}^\top((1-\alpha)\mathbf{\Sigma}^{\ell}_1 + \alpha\tilde{\mathbf{\Sigma}}_1)\mathbf{v}}{\mathbf{v}^\top((1-\alpha)\mathbf{\Sigma}^{\ell}_2 + \alpha\tilde{\mathbf{\Sigma}}_2)\mathbf{v}} + \frac{\mathbf{v}^\top((1-\alpha)\mathbf{\Sigma}^{\ell}_2 + \alpha\tilde{\mathbf{\Sigma}}_2)\mathbf{v}}{\mathbf{v}^\top((1-\alpha)\mathbf{\Sigma}^{\ell}_1 + \alpha\tilde{\mathbf{\Sigma}}_1)\mathbf{v}}$$

$$\text{with} \quad \tilde{\mathbf{\Sigma}}_{\mathrm{c}} = \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\mathbf{\Sigma}^{\mathrm{k}}_{\mathrm{c}}$$

being the average class-covariance matrix computed on other subjects' data. On the other hand our method maximizes the following term

$$(1-\lambda)\left[\frac{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_1\mathbf{v}}{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_2\mathbf{v}} + \frac{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_2\mathbf{v}}{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_1\mathbf{v}}\right] - \lambda\frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\left[\frac{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_c\mathbf{v}}{\mathbf{v}^\top\mathbf{\Sigma}^{k}_c\mathbf{v}} + \frac{\mathbf{v}^\top\mathbf{\Sigma}^{k}_c\mathbf{v}}{\mathbf{v}^\top\mathbf{\Sigma}^{\ell}_c\mathbf{v}}\right]$$

We compare the behaviour of both objective functions in a simulation experiment for $\mathbf{v} \in [-1\ \ 1] \times [-1\ \ 1]$ and different α and λ parameters. Let $\ell = 1$ and

$$\mathbf{\Sigma}^1_1 = \begin{bmatrix}1.3 & 0\\ 0 & 1.0\end{bmatrix}, \quad \mathbf{\Sigma}^1_2 = \begin{bmatrix}0.7 & 0\\ 0 & 1.0\end{bmatrix}$$

$$\mathbf{\Sigma}^2_1 = \begin{bmatrix}1.3 & 0\\ 0 & 1.5\end{bmatrix}, \quad \mathbf{\Sigma}^2_2 = \begin{bmatrix}0.7 & 0\\ 0 & 0.5\end{bmatrix}$$

One can see in Fig. 3 (green color stands for higher values) that covCSP prefers the spatial filter $\mathbf{v}^T = [1\ \ 0]$ for small α parameters but if the regularization of the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ increases (large α) it switches its preference to $\mathbf{v}^T = [0\ \ 1]$ because the second source of subject 2 is more discriminative than the first source (ratio 1.5/0.5 compared to 1.3/0.7). The divCSP-AS regularization on the other hand aims to find discriminative sources that are similar between both subjects thus it always prefers the filter $\mathbf{v}^T = [1\ 0]$ that extracts source 1 which is discriminative and common to both users. Depending on the particular application scenario the covCSP regularization or the divCSP-AS regularization scheme may be more advantageous.

V. Discussion

Applying information geometric methods to BCI is a new direction of research with high potentials. But also classical methods of differential geometry have been recently applied to BCI, e.g. the authors of [29] directly classify trials on the manifold of covariance matrices. It would be interesting to connect both research directions, information geometry and Riemannian geometry, in the future. A question that may also be relevant for future research on divergence-based spatial filtering algorithms for BCI is whether it is advantageous to extend the Gaussian approximation and use more complicated (higher moments) distributions to compute the CSP filters. Restricting the analysis to variance makes sense when only considering the ERD/ERS effect, however, higher moments

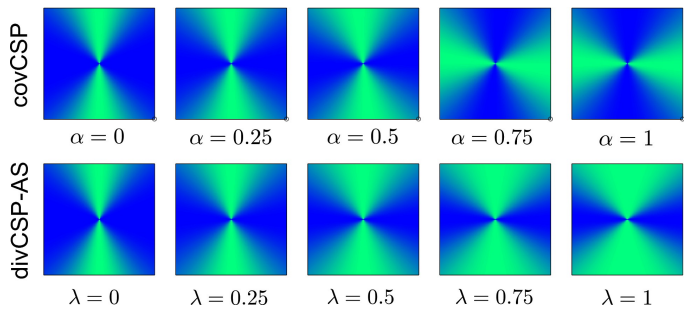

Fig. 3: Values of covCSP and divCSP-AS objective function for $\mathbf{v} \in [-1\ \ 1] \times [-1\ \ 1]$ and different α and λ parameters. Green color represents larger values. The upper left (right) corner represents $\mathbf{v}^T = [-1\ \ -1]$ ($\mathbf{v}^T = [-1\ \ 1]$), the lower left (right) one stands for $\mathbf{v}^T = [1\ \ -1]$ ($\mathbf{v}^T = [1\ \ 1]$).

may contain valuable information e.g. when it comes to minimizing non-stationarity of the extracted sources or considering statistical independence. Our information geometric framework can be used for arbitrary probabilistic distributions as it is based on divergences.

Finally one may also apply other divergences [30] to the CSP problem. Although other divergences may impose different valuable properties on the solution, the optimization process may become more difficult.

Acknowledgment

This work was supported in part by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115), in part by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008 and in part by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education.

References

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

[3] R. Tomioka and K.-R. Müller, "A regularized discriminative frame work for EEG analysis with application to brain-computer interface," NeuroImage, vol. 49, no. 1, pp. 415–432, 2009.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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 Ad. in NIPS 20, pp. 113–120, 2008.

[9] 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.

[10] 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.

[11] 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.

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

[13] H. Kang and S. Choi, "Bayesian multi-task learning for common spatial patterns," in Int. Workshop on Pattern Recognition in NeuroImaging (PRNI), 2011, pp. 61–64.

[14] W. Samek, D. Blythe, K.-R. Müller, M. Kawanabe, "Robust spatial filtering with beta divergence," in Advances in Neural Information Processing Systems 26 (NIPS), 1007-15, 2013.

[15] W. Samek, M. Kawanabe, K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," IEEE Reviews in Biomedical Engineering, 2014, in press.

[16] S. Amari, H. Nagaoka, and D. Harada, Methods of information geometry. American Mathematical Society, 2000.

[17] S. Amari and A. Cichocki, "Information geometry of divergence functions," Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 58, no. 1, pp. 183–195, 2010.

[18] A. Hyvärinen, "Survey on independent component analysis," Neural Computing Surveys, vol. 2, pp. 94–128, 1999.

[19] M. Kawanabe, W. Samek, P. von Bünau, and F. Meinecke, "An information geometrical view of stationary subspace analysis," in Artificial Neural Networks and Machine Learning - ICANN 2011, ser. LNCS. Springer, 2011, vol. 6792, pp. 397–404.

[20] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," Neural Computation, vol. 14, no. 8, pp. 1859–1886, 2002.

[21] 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, pp. 214 101+, 2009.

[22] 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.

[23] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural computation 10 (5), 1299-1319, 1998.

[24] G. Montavon, M. Braun, T. Krüger, and K.-R. Müller, "Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment," IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 62–74, 2013.

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

[26] 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.

[27] F. Lotte and C. Guan, "Learning from other subjects helps reducing Brain-Computer interface calibration time," in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2010, pp. 614-617.

[28] 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.

[29] 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), pp. 472–476, 2010.

[30] 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.