31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Robust Common Spatial filters with a Maxmin Approach

Motoaki Kawanabe, Carmen Vidaurre, Simon Scholler and Klaus-Robert Müller

Abstract— **Electroencephalographic signals are known to be non-stationary and easily affected by artifacts, therefore their analysis requires methods that can deal with noise. In this work we present two ways of calculating robust common spatial patterns under a maxmin approach. The worst-case objective function is optimized within prefixed sets of the covariance matrices that are defined either very simply as identity matrices or in a data driven way using PCA. We test common spatial filters derived with these two approaches with real world brain-computer interface (BCI) data sets in which we expect substantial "day-to-day" fluctuations (session transfer problem). We compare our results with the classical common spatial filters and show that both can improve the performance of the latter.**

I. Introduction

Brain-computer interfaces (BCI) are systems that translate the users intent, coded by a small set of mental tasks, into control actions such as computer applications or prostheses [1]–[4]. In order to translate the brain activity into commands, it is necessary to extract meaningful features from the acquired signals. One of the most popular tools to extract information from the brain signals is the calculation of common spatial filters (CSP) [5]. This data driven approach optimizes spatial filters for each subject individually. CSP analysis is embedded in machine learning methods. Over the last years, machine learning has led to significant advances in the analysis and modeling of neural signals. In EEG-BCI experimentation, the time needed for user's neurofeedback training has been reduced from several days to just a couple of sessions [6]. Typically, collecting examples of EEG signals during which the user is cued to perform repeatedly a small number of e.g. motor imagery tasks [7] is sufficient to adapt the system to the subject and start the feedback. In this step the users can actually transfer information through their brain activity and control applications. However, there are several aspects in which BCI research can profit from improvement, see the 'Challenges' section of [8]. One of them is to gain robustness against non task-related fluctuations and/or non-stationarity of the measured EEG signals.

This work profits from the recent paper [9], in which a maxmin approach to Fisher discriminant analysis (FDA) was applied for robust classification. From their maxmin theorem, the maxmin FDA is guaranteed to have higher discriminative power for *any fluctuations within a prefixed tolerance set*. Inspired by that work, the present paper contributes by investigating such a maxmin approach to common spatial patterns (CSP) [10]. However, in contrast to the FDA case [9], we can obtain the worst case covariance matrices analytically, and define a modified generalized eigenvalue problem.

Motoaki Kawanabe and Simon Scholler are with the IDA Group at FIRST.Fraunhofer, Kekulestr. 7, Berlin 12489, Germany

Carmen Vidaurre, Klaus-Robert Müller and Motoaki Kawanabe are with the Machine Learning Dp, TU-Berlin, Franklinstr. 28/29, Berlin 10587, Germany

simon.scholler@first.fraunhofer.de

II. The original Common Spatial Pattern Algorithm

Many EEG-BCIs, are based on motor imagery. Commonly, subjects using these systems are asked to perform the imagination of hands, feet or mouth. Motor imagery alters the rhythmic activity that can be measured in the EEG over the sensorimotor cortex. Many EEG rhythms are called idle rhythms because they are generated by large populations of cortical neurons that fire in rhythmical synchrony when they are not engaged in a specific task. Oscillations with a fundamental frequency between 9 and 13 Hz can be observed over motor and sensorimotor areas in most subjects (the μ-rhythm). These sensorimotor rhythms (SMRs) are attenuated in the corresponding cortical area when a motor task (e.g. movement or motor imagery) takes place. As this effect is due to loss of synchrony in the neural populations, it is termed event-related desynchronization (ERD), see [11]. The increase of oscillatory EEG (i.e., the reestablishment of neuronal synchrony) is called event-related synchronization (ERS). To distinguish motor imagery tasks of different body parts it is necessary to recognize different spatial localization of SMR modulations. The locations over the sensorimotor cortex are related to corresponding parts of the body. For example, left and right hand are localized in the contralateral hemisphere, i.e., right and left motor cortex, respectively. Thus, spatial filters are an essential step for a meaningful feature extraction and posterior classification of motor intentions. One of the most popular and successful algorithms for calculating spatial filters is CSP. Given two distributions in a high-dimensional space (corresponding in our case to two different mental tasks), the CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and simultaneously minimize variance for the other class. Since band-power is equivalent to the variance of band-pass filtered signals, this criterion corresponds to ERD/ERS effects.

Mathematically CSP analysis works as follows. Let Σ_+ and Σ_- be covariance matrices of the band-pass filtered EEG signals of two different motor imagery tasks. These two matrices are simultaneously diagonalized such that the eigenvalues of Σ_+ and Σ_- sum to 1. This can be done by calculating the generalized eigenvectors W:

$$\Sigma_+ W = (\Sigma_+ + \Sigma_-) W D. \quad (1)$$

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

Here, the diagonal matrix D contains the (generalized) eigenvalues of Σ_+ (defined such that they are between 0 and 1) and the column vectors of W are the filters w's for computing the CSP features. The best discrimination is provided by those filters with high eigenvalues (large variance for condition 1 and small variance for condition 2) and by filters with low eigenvalues (vice versa). Therefore, the common practice in a classification setting is to use several eigenvectors from both ends of the eigenvalue spectrum as features for classification. Alternatively, the solution for the eigenvector with the largest eigenvalue can also be obtained by maximizing the Rayleigh quotient:

$$\underset{w \in \mathbb{R}^C}{\text{maximize}} \qquad \frac{w^\top \Sigma_+ w}{w^\top (\Sigma_+ + \Sigma_-) w}. \tag{2}$$

This correspondence is often useful for algorithmic considerations.

III. Common Spatial Pattern under the Maxmin Approach

The class covariance matrices Σ_+ and Σ_- used in CSP can vary substantially because of non task-related fluctuations and/or non-stationarity of the EEG signals. In BCI applications, it is thus important to make the features robust against such changes. In those cases, the maxmin approach [9] applied successfully to FDA could be one of the promising directions to construct robust CSP filters. The key idea is that, instead of just two single matrices, we consider convex sets $\mathcal{S}_+$ and $\mathcal{S}_-$ for the class covariances Σ_+ and Σ_-, respectively. These sets specify the tolerance regions of fluctuations around the class covariances. For simplicity, we assume that the sets $\mathcal{S}_+$ and $\mathcal{S}_-$ are independent of each other. Based on the maxmin framework, robust CSP filters can be constructed by maximizing the worst case (minimum) Rayleigh quotient within all possible covariance matrices in the tolerance regions. That is, the optimization problems for our maxmin CSP can be expressed as

$$\max_{w \neq 0} \quad \min_{\Sigma_+ \in \mathcal{S}_+, \Sigma_- \in \mathcal{S}_-} \frac{w^\top \Sigma_+ w}{w^\top (\Sigma_+ + \Sigma_-) w} \tag{3}$$

$$\max_{w \neq 0} \quad \min_{\Sigma_+ \in \mathcal{S}_+, \Sigma_- \in \mathcal{S}_-} \frac{w^\top \Sigma_- w}{w^\top (\Sigma_+ + \Sigma_-) w} \tag{4}$$

From now, we consider only the first optimization problem (3), because the other (4) can be handled in the same way.

A. Derivation of the Maxmin Filters

We further assume that the sets $\mathcal{S}_+$ and $\mathcal{S}_-$ can be defined by balls in the space of $C \times C$ positive definite matrices (i.e. $\Sigma_\pm \succeq 0$) centered at $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$

$$\begin{aligned} \mathcal{S}_+ &= \left\{ \Sigma_+ \middle| \Sigma_+ \succeq 0, \ \|\Sigma_+ - \overline{\Sigma}_+\| \leq \delta_+ \right\}, \\ \mathcal{S}_- &= \left\{ \Sigma_- \middle| \Sigma_- \succeq 0, \ \|\Sigma_- - \overline{\Sigma}_-\| \leq \delta_- \right\}, \end{aligned} \tag{5}$$

where $\|\cdot\|$ denotes an appropriate norm of the matrix space. We used the average class covariances for the centers $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$. The following two norms are considered in this paper. One is $\|X\|_P^2 := \text{Tr}\left(P^{-1} X P^{-1} X\right)$ for any symmetric matrix X, where P is a $C \times C$ positive definite matrix specifying the shape of the balls. When $P = I$, this boils down to the standard 'Frobenius' norm. The other takes into account variability of covariance matrices over time through PCA, which will be explained in the next section.

For the first choice, the worst-case covariances in the tolerance sets can be determined explicitly, and therefore the maxmin filter can be determined by a generalized eigenvalue problem as the original CSP.

Lemma 1: For the sets $\mathcal{S}_+$ and $\mathcal{S}_-$ defined in Eq. (5), the worst case Rayleigh quotient becomes

$$\frac{w^\top \left(\overline{\Sigma}_+ - \frac{\delta_+}{\sqrt{C}} P_+\right) w}{w^\top \left(\overline{\Sigma}_+ + \overline{\Sigma}_- - \frac{\delta_+}{\sqrt{C}} P_+ + \frac{\delta_-}{\sqrt{C}} P_-\right) w}, \qquad \forall w, \tag{6}$$

if $\overline{\Sigma}_+ - \frac{\delta_+}{\sqrt{C}} P_+ \succeq 0$.

In our experiment, we take Frobenius norm, i.e. $P_+ = P_- = I$ (identity maxminCSP). Although the identity matrix ignores plausible directions of fluctuation in EEG signals, the maxmin CSP with this setting still improved the performance in the "day-to-day" transfer experiment. We conjecture that this is analogous to the fact that Bayesian regularization helps even with non-informative priors. If we have extra (prior) information about possible fluctuations as is the case with the real world BCI data in [12], the covariance of the distortions can be used for the matrices P_+ and P_-. This approach was called invariant CSP (iCSP). Fig. 1 is an illustrative explanation of our method. Although we develop the theory only for the first eigenvectors, in the experiments we will use a few eigenvectors each. Further work should be done to extend Lemma 1 for multiple eigenvectors.

B. Alternative Maxmin Filters Depending on Actual Non-stationarity

Although the maxmin CSP with the identity matrix improved the original CSP in our experiments, the corresponding tolerance sets do not well capture the actual variability of the covariance matrices over time. Therefore, it is natural to analyze their non-stationarity and to take such information into account for calculating robust filters. In this paper, we propose a variant of the maxmin CSP with matrix PCA. The procedure consists of the following three steps: 1) matrix PCA, 2) computing the worst case covariances and 3) derivation of the maxmin filters. We will explain step by step.

Let $\{\Sigma_\pm^{(k)}\}_{k=1}^K$ be sets of locally-averaged covariance matrices in time, where in our experiment we will use trial-wise covariances (without averaging), session-wise averages and local averages within four blocks in each session. We would like to fit the tolerance regions (the ellipsoids in Fig.1.(d)) so that they match variability of the covariances $\{\Sigma_\pm^{(k)}\}_{k=1}^K$. To do so, we need to find directions of large fluctuation, which can be done by PCA. At first, $C \times C$ matrices are transformed to C^2-dimensional vectors. Then, the covariance of the extended vectors is calculated and its eigen decomposition is obtained. Finally, the C^2-dimensional eigenvectors are

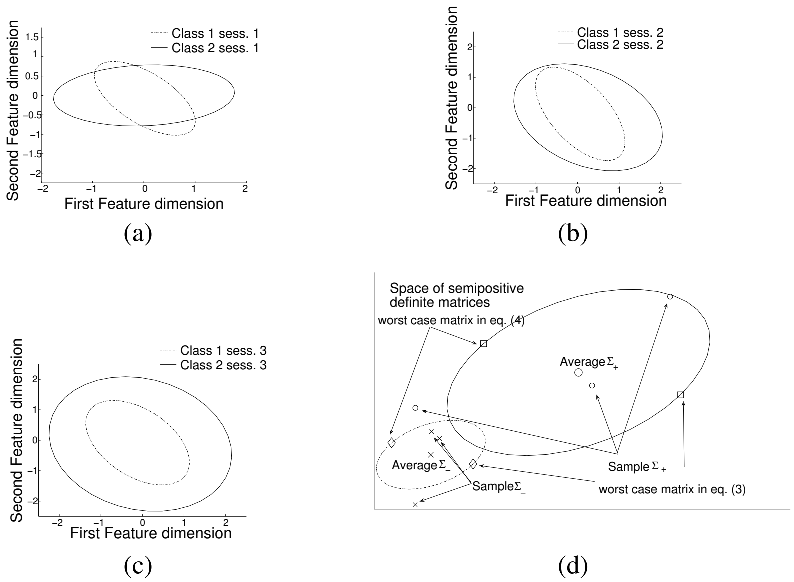

Fig. 1. Figs. (a), (b) and (c) represent Σ_+ and Σ_- at different time points. Mean of the features is 0, as it is bandpass filtered data. Fig. (d) represents the previous matrices as points in the space of positive definite matrices. The ellipsoids in Fig. (d) are the tolerance sets S_+ and S_- centered at the average matrices $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$, respectively. From both ellipsoids, a pair of the worst case covariances is obtained for each optimization problem (3) or (4).

transformed back to $C \times C$ matrices. Suppose that $\lambda_\pm^{(i)}$ and $V_\pm^{(i)}$ are eigenvalues and matrices by class-wise PCA. We define the tolerance sets as

$$\mathcal{S}_\pm := \left\{ \Sigma_\pm = \overline{\Sigma}_\pm + \Delta_\pm \;\middle|\; \Delta_\pm = \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}, \; \|\Delta_\pm\|_{\mathrm{PCA}} \le \delta_\pm \right\}, \quad (7)$$

where

$$\|\Delta_\pm\|_{\mathrm{PCA}}^2 := \sum_i \frac{\left(\alpha_\pm^{(i)}\right)^2}{\lambda_\pm^{(i)}}. \quad (8)$$

This norm allows larger variations in the directions with large eigenvalues. To avoid instability, we ignore small eigenvalues in the following discussion. For a given w, the worst case covariances in (3) can be obtained by optimizing

$$\min_{\Sigma_+ \in \mathcal{S}_+} w^\top \Sigma_+ w = \min_{\boldsymbol{\alpha}_+} \sum_i \alpha_+^{(i)} w^\top V_+^{(i)} w, \quad (9)$$

$$\max_{\Sigma_- \in \mathcal{S}_-} w^\top \Sigma_- w = \max_{\boldsymbol{\alpha}_-} \sum_i \alpha_-^{(i)} w^\top V_-^{(i)} w, \quad (10)$$

under constraints

$$\sum_i \frac{\left(\alpha_\pm^{(i)}\right)^2}{\lambda_\pm^{(i)}} \le \delta_\pm^2, \quad (11)$$

$$\overline{\Sigma}_\pm + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)} \succeq 0. \quad (12)$$

If we ignore the positive definiteness (12), the solutions can be obtained analytically as

$$\alpha_+^{(i)} = \frac{-\delta_+ \lambda_i w^\top V_+^{(i)} w}{\sqrt{\sum_i \lambda_i \left(w^\top V_+^{(i)} w\right)^2}}, \quad (13)$$

$$\alpha_-^{(i)} = \frac{\delta_- \lambda_i w^\top V_-^{(i)} w}{\sqrt{\sum_i \lambda_i \left(w^\top V_-^{(i)} w\right)^2}}, \quad (14)$$

When $\overline{\Sigma}_\pm + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}$ violate (12), we truncate the negative eigenvalues of the worst case covariances to zeros. Once the worst case covariances are obtained, we can update the maxmin filters by CSP of these matrices. In contrast to the previous case (6), we need to iterate the second and third steps, since the worst case covariances depend on the current filter w. However, as we will show in our experiment, only a single update from the original CSP works fine in practice.

IV. Results

In this paper we evaluate the proposed algorithm on offline data in which substantial fluctuations are expected. In particular, we test the algorithm for obtaining robust filters against session-to-session (day-to-day) variability which may be caused by different mental conditions, materials (cap and electrodes) and different preparation of the measurement devices. For the analysis, we use ‘calibration measurements’ in which no BCI feedback is provided to avoid bias toward any method. The trials of the ‘calibration data’ have a fixed length of 3.5 seconds in which an imagery motor task (right or left hand or foot movement) is performed in response to a visual cue (letter L, R, F). For the evaluation, we only consider binary classification. The best pair of three motor imaginary tasks, is selected based on separability for each subject. The data is recorded using 48 Ag/AgCl electrodes

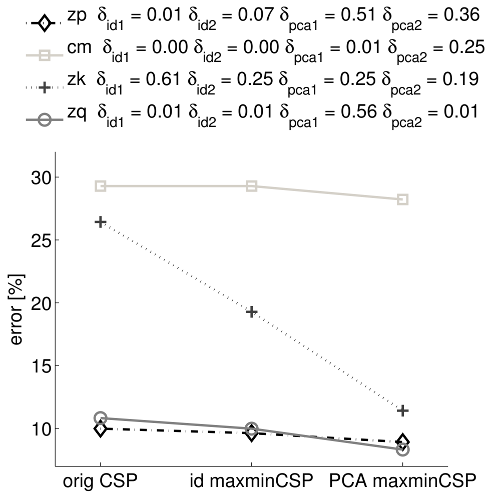

Fig. 2. Test errors of the original CSP algorithm (orig CSP) and the two maxmin versions, identity maxminCSP (id maxminCSP) and PCA maxminCSP. The legend provides subject codes (two letter codes) and the selected parameters in square brackets for the session-to-session transfer.

in an extended 10-20 system sampled at 1000 Hz with a band-pass from 0.05 to 200 Hz. Linear discriminant analysis (LDA) is used for classification and the performance is measured by the error rate.

For the analysis, data from four subjects for whom we recorded several sessions in different days (even with more than 12 months difference) is used. For subject *zq* we use 6 available datasets, 5 datasets for *cm* and *zp* and finally 4 for subject *zk*. All files except one are used for training the maxmin filters. The parameters of the model are selected by cross-validating in the training set. The selected delta values (see Sections III-A and III-B) are shown in Fig. 2. Matrices P_+ and P_- are either the identity matrix (id maxminCSP, see Section III-A) or a matrix selected using PCA (see Section III-B). The last file is used to test the performance of each subject. Fig. 2 shows the error rate when using the original CSP, identity maxminCSP and PCA maxminCSP for pre-processing the data. The identity maxminCSP outperforms original CSP in 3 of 4 cases, whereas PCA maxminCSP method outperforms identity maxminCSP and original CSP in all subjects. The large error reduction of subject zk implies that the covariances in the test session have similar characteristics to the worst cases.

V. Discussion and Conclusion

BCI data is contaminated by a variety of noise sources, artifacts, non-stationarities and outliers that make it indispensable to strive for more robust learning methods. In this paper we proposed a novel algorithm for robust spatial filtering that is inspired by [9]. In particular, we analyze the worst case performance among possible class covariance matrices and optimize the respective CSP-like filters based on such a criterion. We take balls or ellipsoids in the matrix space for the sets of the covariances and the algorithm can be elegantly reduced to a generalized eigenvalue problem similar to the original CSP, but with modified covariance matrices. The simulations presented in this paper show that the maxminCSP framework is indeed more robust as it allows transfer BCI classifier knowledge from session to session. This permits to construct a BCI system that is more stable with respect to non-stationarities and non-task related fluctuations. In future studies we will continue working towards more robust BCIs. One of our goals is to use the ideas presented in this paper to define a non-parametric variant of the problem and skip the model selection needed until now.

Acknowledgments

This work was partly supported by the *Bundesministerium fuer Bildung und Forschung* (BMBF), FKZ 01IBE01A/B, by the German Science Foundation (DFG, contract MU 987/3-1), by the EU's IST Programme under the PASCAL2 Network of Excellence, IST-216996, by the EU's Marie Curie grant number MEIF-CT-040666 and by the EU's ICT Programm under TOBI, ICT-224631. This publication only reflects the authors' views. Funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, pp. 767–791, 2002.

[2] J. Millán, *Handbook of Brain Theory and Neural Networks*. Cambridge: MIT Press, 2002.

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

[4] B. Blankertz, K.-R. Müller, D. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer, "The BCI competition III: Validating alternative approachs to actual BCI problems," *IEEE Trans Neural Sys Rehab Eng*, vol. 14, no. 2, pp. 153–159, 2006.

[5] 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. [Online]. Available: http://dx.doi.org/10.1109/MSP.2008.4408441

[6] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces," *IEEE Trans Biomed Eng*, vol. 54, no. 3, pp. 550–556, 2007.

[7] K.-R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring," *J Neurosci Methods*, vol. 167, no. 1, pp. 82–90, 2008.

[8] A. Nijholt, D. Tan, G. Pfurtscheller, C. Brunner, J. Millán, B. Allison, B. Grainmann, F. Popescu, B. Blankertz, and K.-R. Müller, "Brain-computer interfacing for intelligent systems," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 72–79, 2008.

[9] S.-J. Kim, A. Magnani, and S. Boyd, "Robust fisher discriminant analysis," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 659–666.

[10] Z. J. Koles, "The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG," *Electroencephalogr Clin Neurophysiol*, vol. 79, no. 6, pp. 440–447, 1991.

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

[12] B. Blankertz, M. Kawanabe, R. Tomioka, F. 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*. Cambridge, MA: MIT Press, 2008, pp. 113–120.