CHANNEL SELECTION OVER RIEMANNIAN MANIFOLD WITH NON-STATIONARITY CONSIDERATION FOR BRAIN-COMPUTER INTERFACE APPLICATIONS

Khadijeh Sadatnejad[1], *Aline Roc*[1,2], *Léa Pillette*[1,2], *Aurélien appriou*[1,2], *Thibaut Monseigne*[1], *Fabien Lotte*[1,2]
[1]Potioc team, Inria Bordeaux Sud-Ouest, Talence, France, [2]LaBRI (CNRS / Univ. Bordeaux / Bordeaux INP), Talence, France

ABSTRACT

In this paper, we propose and compare multiple criteria for selecting ElectroEncephaloGraphic (EEG) channels over the Riemannian manifold, for EEG classification in Brain-Computer Interfaces (BCI). These criteria aim to promote EEG covariance matrix classifiers to generalize well by considering EEG data non-stationarity. Our approach consists of both increasing the discriminative information between classes over the manifold and reducing the dispersion within classes. We also reduce the influence of outliers in both discriminative and dispersion measures. Using the proposed criteria, channel selection is done automatically in a backward elimination process. The criteria are evaluated on EEG signals recorded from a tetraplegic subject and dataset IVa from BCI competition III. Experimental evidences confirm that considering the dispersion within each class as a measure for quantifying the effects of non-stationarity and removing the most affected channels can improve the performance of BCI by 5% on the tetraplegic subject and by 12 % on dataset IVa.

Index Terms— Channel selection, BCI, EEG, covariance matrix, Riemannian manifold

1. INTRODUCTION

A Brain-Computer Interface (BCI) processes a user's brain signals, typically measured using Electroencephalography (EEG), and translates them into commands for an interactive application [1]. EEG signals are essentially non-stationary [2], due to various neurophysiological and extra-physiological causes [3], leading to variations in BCI users' performance [4]. This highlights the necessity of considering the non-stationarity of the data to improve how well an EEG classifier can generalize to unseen EEG data in BCI [5, 7].

In the literature, such non-stationarities have been addressed using a wide range of techniques, at different BCI levels [6]. At the machine learning level, one approach consists in matching the statistical distributions of different datasets - corresponding to different sessions or subjects - using geometrical transformations such as translation, scaling and rotation [8, 9, 10]. Additionally, classifiers ensembles [3] are another approach for handing non-stationarity. With this approach, the information from multiple sources (i.e., subjects or sessions) is handled by multiple classifiers. These classifiers are combined into a "global" classifier. As a general categorization, these methods can be divided into two approaches: the first approach's main goal is to match the distribution of the data by applying some geometrical transformation, while the second approach tries to include all the statistical variabilities and model them. In none of these two approaches, the non-stationarity in the data is removed or reduced: their main aim is to handle its existence. See [11] for a review of all such approaches.

Despite the variety of variables that may lead to data non-stationarity, including psychological characteristics and neuroanatomic properties, the neuro-physiological causes of all these variations are not well nor fully known [9,12]. Thus, modeling these sources and removing/reducing their effect precisely had not been possible, at least so far.

Our contribution to tackling data non-stationarity, in order to make BCI more robust, is to quantify and reduce the undesirable effects of these non-stationary sources by removing EEG channels which are mostly affected by them.

For representing EEG patterns, we use EEG spatial covariance matrices, which have been shown to be efficient descriptors when analyzing them using Riemannian geometry [13, 14]. Indeed, covariance matrices are symmetric positive definite (SPD). Due to their positive definiteness, their feature space is not a vector space. Considering the non-linear geometry of data points in analysis and reformulating the space as a connected Riemannian manifold, by equipping each tangent space with a Riemannian metric, leads to superior results in comparison with analyses in Euclidean space [13]. Thus, in this study, we focus on covariance matrices Riemannian geometry.

Barachant and Bonnet [15] proposed a channel selection method, which is adapted to the Riemannian geometry of covariance matrices. They used the distance between the means of different classes as a criterion for removing the less informative channels [15], as follows:

$$Crit0 = \sum_{c1}\sum_{c2} d_R(\bar{C}^{(c1)}, \bar{C}^{(c2)}) \qquad (1)$$

where $\bar{C}^{(ci)}$is the mean of class *i*. Although their method led to good results, they have not considered changes in data variance. Hence, even with increasing data non-stationarity, this criterion only considers the distance between means, despite changes in variance and possible overlap between classes. Thus, there is a need for channel selection algorithms, adapted to Riemannian classifiers, that can deal with EEG non-stationarities. This is what we propose here.

In this study, we first assume that non-stationarity mostly manifests as changes in the variance of data. Thus, we try to remove the channels which convey most of the variance while conveying less discriminative information.

As the second criterion, we quantify the effects of non-stationarity (i.e., change in data distribution) as multi-modal distributions for each class, with overlap between classes. Our second proposed approach consists in considering the clusters of samples from one class surrounded by samples from other classes as an indicator for dispersion, and thus in removing channels leading to this dispersion.

The remainder of this paper is organized as follows: first, we provide notations and some basic definitions in section 2.1. In section 2.2 we describe our proposed approaches. The experimental set-up and results are reported and discussed in section 3. Finally, the paper is concluded in section 4.

2. METHOD

2.1. Riemannian manifold of covariance matrices

The representation of data points in this paper is denoted by a covariance matrix $C_i^{(c)}$, where i denotes the trial number and c denotes its corresponding class. $C_i^{(c)}$ is an $n \times n$ positive definite matrix, where n denotes the number of EEG channels. It is computed by the optimal linear shrinkage of the spatial covariance matrix of each trial [16]. Each class, i.e., each mental task, is represented using a pool of covariance matrices. The Riemannian distance between two covariance matrices along the manifold is computed as:

$$d_R\left(C_i^{(c1)}, C_j^{(c2)}\right) = \left\| \log\left({C_i^{(c1)}}^{-1} C_j^{(c2)}\right) \right\|_F \qquad (2)$$

where $C_i^{(c1)}$ and $C_j^{(c2)}$ are the covariance matrices of the i^{th} and j^{th} trial of class $c1$ and $c2$ respectively, $\|.\|_F$ denotes the Frobenius norm, and $\log(.)$ is the log-matrix operator. To compute the mean of SPD matrices in a Riemannian framework, we use the Riemannian center of mass $\bar{C}^{(c)}$:

$$\bar{C}^{(c)} = argmin_C \sum_i {d_R}^2(C, C_i^{(c)}) \qquad (3)$$

This point has the minimum squared distance from all points of the class c. There is no closed-form solution for this equation, but it can be computed using an iterative algorithm [17].

The variance within each class is defined based on [20]:

$$\sigma^{(c)} = 1/N_c \sum_i {d_R}^2(\bar{C}^{(c)}, C_i^{(c)}) \qquad (4)$$

where $\sigma^{(c)}$ is an empirical estimation of the variance of class c and N_c is the number of samples.

2.2. Channel selection

From the classification point of view, having a discriminative representation of different classes is desirable. Changes in the distribution of data, as the result of sources of non-stationarity, can affect class discrimination. In the following, we propose two criteria that measure both the data dispersion and how much each channel carries discriminative information. The selection of the channels using such criteria is done based on algorithm 1 [15]. It is written for the first criterion. To use it with the second criterion, lines 5 and 7 should be changed to compute Eq. (9) and Eq. (10) (described hereafter), line 9 to compute Eq. (11), and line 13 should be removed.

Algorithm 1. Channel subset selection

```
Input: $\bar{C}^{(c1)}, \bar{C}^{(c2)}, C_i^{(c)}$        % $c, c1, c2$: class type
Input: $N^o, N_a, N_C, N_c$        %$N^o$: No of selected channels, $N_a$: No of all channels,
                                   % $N_C$: No. of classes, $N_c$: No. of samples in each class
Output: $Subset$                   %$Subset$: Subset of selected channels

1: $Subset = [1 \dots N_a]$
2: $For\ k = 1 : N_a - N^o$
3:   $For\ i = 1 : N_a - k + 1$
4:     $For\ c = 1 : N_C$
5:       $\bar{C}_{tmp}^{(c)} = R(\bar{C}^{(c)}, i)$    %$R(C,i)$: Remove $i^{\text{th}}$ row and column from matrix $C$
6:       $\acute{C}_i^{(c)} = R(C_i^{(c1)}, i)$
7:       $\sigma^{(c)} = \sum_{i=1}^{N_c} d_R^2(\bar{C}_{tmp}^{(c)}, C_i^{(c)}) / (N_a - i)$
8:     $End$
9:     $Crt(i) = \sum_{c1} \sum_{c2} d_R(\bar{C}_{tmp}^{(c1)}, \bar{C}_{tmp}^{(c2)}) / (\sum_c \sigma^{(c)})^2$
10: $End$
11: $i^* = argmax_i(Crt(i))$
12: $Subset = R(Subset, i^*)$    % Remove $i^{*\text{th}}$ element of $Subset$
13: $\bar{C}^{(c)} = R(\bar{C}^{(c)}, i^*)$
14: $C_i^{(c)} = R(C_i^{(c)}, i^*)$
15: $End$
```

2.2.1 Maximum distance minimum variance criteria

It is assumed that channels that are mostly affected by different sources of non-stationarity could have a higher influence on data dispersion. Such channels may also carry less discriminative information. By considering these two factors simultaneously, we can avoid removing high variance channels that convey high discriminative information. To do so, our first selection criterion is:

$$Crit\ 1 = \sum_{c1} \sum_{c2} d_R(\bar{C}^{(c1)}, \bar{C}^{(c2)}) / (\sum_c \sigma^{(c)})^2 \qquad (5)$$

where the numerator represents the distance between class means over the manifold and the denominator represents the total variance within different classes. For channel selection, we start with all channels. Then, we assess the effects of removing each channel on *Crit1*, and the less informative channel (i.e., the channel whose removal leads to a maximum value for *Crit1*) is removed. In other words, channels that contribute most to the variance and carry less discriminative information are removed first using this criterion. However, using the variance for controlling the non-stationarity may suffer from outliers interference. Thus, we propose a second criterion taking into account outliers.

2.2.2 Maximum margin –minimum inter-class dispersion

To inhibit the effects of outliers, we consider both discriminative information and the dispersion component of the selection criterion. For representing discriminative information, we use the margin size (i.e., $\frac{2}{\|W\|^2}$) of a soft margin Support Vector Machine (SVM) classifier, where

$$\|W\|^2 = \sum_i \sum_j \alpha_i \alpha_j y_i y_j \; K\left(C_i^{(c1)}, C_j^{(c2)}\right) \qquad (6)$$

$C_i^{(c1)}$ and $C_j^{(c2)}$ are support vectors covariance matrices, K a kernel between covariance matrices (see Eq. (8) below) and y_i and y_j are the matrices labels. α_i are the coefficients obtained by optimizing the objective function of SVM. With an appropriate set-up of regularization parameter (not very large value (i.e., overfitting to data including outliers) nor very small (i.e., a lot of misclassification in training process), the margin size is robust to outliers [7]. To represent the dispersion, as an indicator of non-stationarity, instead of considering the within-class variance, we represent it by batches of samples from one class that are surrounded by samples from other classes. These samples can affect the generalization of the classifier. For example, in the case of an SVM classifier, these samples may lead to a larger upper bound for the probability of test error, by increasing the empirical risk (i.e., $\frac{m}{l}$), as follows:

$$P_{error} \le \frac{m}{l} + \frac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4m}{l\varepsilon}}\right)$$
$$\varepsilon = 4\,\frac{h\left(\ln\left(\frac{2l}{h}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)}{l} \qquad (7)$$

where m and l denotes the number of misclassified training samples and the number of all the training samples respectively, and h denotes the Vapnik–Chervonenkis dimension of the SVM-classifier, which is inversely proportional to the margin size [7]. Therefore, the influence of non-stationary sources (i.e., undesirable inter-class dispersion), which leads to lower SVM generalization (i.e., higher upper bound for the probability of test error), can be reduced by decreasing the distance of non-outlier misclassified samples from the boundary between classes. By focusing on non-outlier samples when computing the dispersion, we reduce the effects of outliers.

To capture these samples, as an indicator of within-class dispersion, we use the inconsistency between the prediction of a local classifier ($k - NN$) and a global classifier (SVM) in the same feature space. This inconsistency can represent some of the undesirable effects of non-stationarity. For this purpose, we train an SVM using a Riemannian kernel:

$$K\left(C_i^{(c1)}, C_j^{(c2)}\right) = exp\left(-\frac{d_R^2\left(C_i^{(c1)}, C_j^{(c2)}\right)}{2\sigma^2}\right) \qquad (8)$$

Then a weighted $k - NN$ is trained on the same data:

$$f_{NN}(C) = \sum_{i=1}^{k} K(C, C_i)\, y_i$$
$$y_{NN}(C) = sign\left(f_{NN}(C)\right). \qquad (9)$$

The k parameter should be large enough to avoid capturing the outliers in computing the dispersion indicator. The inconsistency between the prediction of the SVM and $k - NN$ classifiers is used for measuring the undesirable interclass dispersion as follows:

$$Disp \propto 1/\sum_c \sum_i \delta\left(y_{svm}(C_i^{(c)}) - y_{NN}(C_i^{(c)})\right) \qquad (10)$$

where δ is a Dirac delta function and $y_{svm}(.)$ and $y_{NN}(.)$ denotes the prediction of SVM and $k - NN$ classifiers respectively. Both the discriminative information and the dispersion measure are used to evaluate the channel subset:

$$Crit2 = \frac{2}{\|W\|^2} + \lambda\left(\sum_c \sum_i \delta\left(y_{svm}(C_i^{(c)}) - y_{NN}(C_i^{(c)})\right)\right) \qquad (11)$$

To give an equal contribution to both factors, both are normalized by their maximum value within an iteration (line 3 of Algorithm 1)

3. EVALUATIONS

3.1. Datasets

The proposed criteria were evaluated on two datasets. The first dataset is recorded from a tetraplegic subject during multiple sessions. The sessions took place on several days at the subject’s home, an environment with low control on background luminosity, ambient sounds or electromagnetic interference, or in the lab. EEG signals were recorded with 46 active scalp electrodes. Our experiments used data from 3 sessions, called S9, S11, and S13. They are composed of 3, 3, and 2 runs respectively. Each run comprised 10 trials per mental task (left or right imagined hand movements). For the offline evaluation of this dataset, EEG signals were band-pass filtered in 8-24 Hz using a 5th order Butterworth filter. For each trial, two epochs were extracted from 250ms after the instruction cue. The window length was 2s with 50% overlap between consecutive windows.

The second dataset used is dataset IVa from BCI competition III [17]. EEG signals were filtered in 8-30 Hz

Table I. Evaluations (accuracy Ac.) on dataset IVa

	Maximum Ac.	Ac. Nº = 10	Mean Ac. Across channels	Mean Ac across subjects
Crit0	0.8321	0.7724	0.7563	0.6666
Crit1	0.8731	0.8212	0.7952	
Crit2	0.8704	0.8245	0.8079	
Margin	0.8673	0.8185	0.8036	

using a 5th order Butterworth filter. For each trial, one epoch is extracted from 0.5s to 4s after the cue.

3.2. Results

We first compared the influence of *Crit0* and *Crit1* on dispersion and discriminative information. For these two criteria, the summation of the variance within each class and the distance between class means for different subsets of channels, are illustrated in Fig.1. All runs of the first dataset (S9-S11-S13) are included. With *Crit1*, about 45% of the variance was removed by removing the first 6 channels (i.e., 55% was preserved by the 40 remaining channels) versus only 17% of it using *Crit0* (Fig 1. (b)).

In the second experiment, to illustrate the effects of reducing the variance, we compared the average accuracy of a Fisher geodesic Minimum Distance to the Mean (fgMDM) classifier [19], across different numbers of channels. Sessions S9, S11, S13 alone and S9-S11-S13 together are used as data sets in this experiment (Fig.2). Leave-one-run-out cross-validation was used for evaluation on each dataset.

In our experiments, we used k = 5 and 10 for detecting probable informative badly located samples using *k-NN*. These values were selected experimentally, by considering the number of samples in the training sets. The SVM regularization parameter was selected by examining C = 0.1, 1, 10, 10^2, 10^3, 10^4 using cross-validation on the training set. We used $\lambda = 1$, in Eq. (11), to use an equal weight for dispersion and discriminative information in *Crit2*.

We ran a one-way repeated measure ANOVA to compare the average accuracy achieved for each dataset for selected channels, by each criterion. For statistical analysis, the numbers of channels considered for analysis are selected from an interval by considering the tradeoff between within and between class-dispersion (e.g., for the first dataset, we considered the numbers of channels which preserve at least 50% discriminative information and remove more than 40% within-class dispersion, i.e., from 10 to 40). Considering the variance in *Crit1* led to a significantly higher accuracy than with *Crit0* ($p<0.05$). However, whereas using margin as selection criteria led to significantly higher accuracy than with *Crit0* ($p<0.05$), the difference between Crit1 and Crit2 was not significant. A comparison between *Crit0* and *Crit2* showed significant superiority of *Crit2* ($p<0.05$). For BCI competition III dataset IVa, the average accuracy with all channels, the best accuracy, and the accuracy for a specific number of channels (10) after channel selection using the different criteria, are averaged across the 5 subjects and are reported in Table I. Statistical testing of the average accuracy across channels (No=10 to 40) between different subjects and criteria, confirmed that *Crit0* and *Crit1* are significantly different ($p<0.05$).

4. CONCLUSION

In this paper, we proposed, studied and compared different criteria for channel selection over the manifold of SPD matrices and test it for EEG classification in BCI application. Our experiments confirm that considering the variance as a factor for controlling the data non-stationarity and removing channels affected by the sources of non-stationarity can improve BCI performances, both as compared to using all channels and as compared to selecting channels based on the between class means only. Our works thus contributed new tools for channel selection in EEG-based Riemannian classifiers.

Acknowledgment

The European Research Council with project BrainConquest (grant ERC-2016-STG-714567) supported this work.

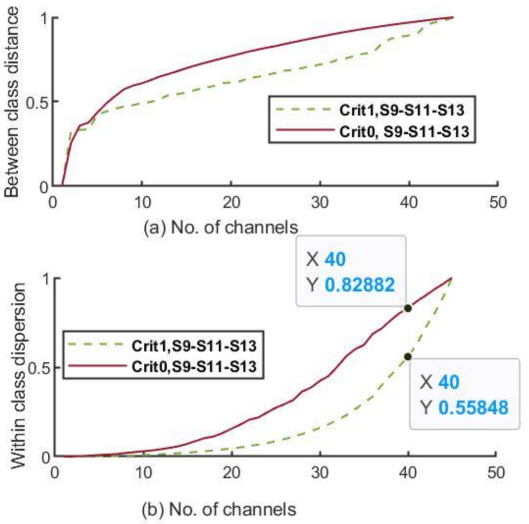

Fig. 1. (a) Distance between means, (b) Variance sum.

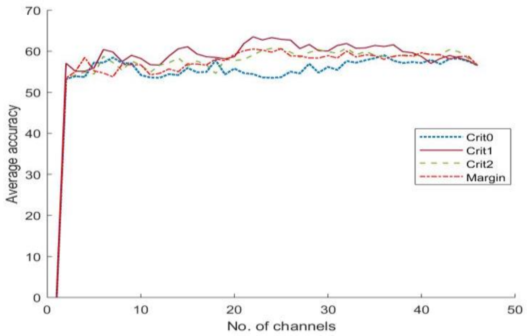

Fig2. Average classification accuracy of fgMDM after channel selection by *Crit0, Crit1*, *Crit2*, and *Margin* size.

5. REFERENCES

[1] M. Clerc, L. Bougrain, and F. Lotte, “Brain-computer Interfaces: Foundations and Methods,” ISTE Limited, 2016.

[2] J. Wolpaw, and E. W. Wolpaw, “Brain-computer interfaces: principles and practice,” OUP USA, 2012.

[3] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces." Journal of neural engineering , 2007.

[4] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges," Signal processing, 2190-2212, 2005.

[5] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Robust EEG channel selection across sessions in brain-computer interface involving stroke patients," The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012.

[6] M. Grosse-Wentrup, and B. Schölkopf, "A review of performance variations in SMR-based Brain− Computer interfaces (BCIs)," Brain-Computer Interface Research. Springer, Berlin, Heidelberg, 39-51, 2013.

[7] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.P.

[8] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, “Transfer learning: a Riemannian geometry framework with applications to braincomputer interfaces,” IEEE Transactions on Biomedical Engineering, pp. 1–1, 2017.

[9] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian procrustes analysis: Transfer learning for brain-computer interfaces," IEEE Transactions on Biomedical Engineering, 2018.

[10] I. Horev, F. Yger, and M. Sugiyama, "Geometry-aware stationary subspace analysis," Asian Conference on Machine Learning, 2016.

[11] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, ”A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update, ” Journal of neural engineering, 2018.

[12] M. Grosse-Wentrup, "What are the causes of performance variation in brain-computer interfacing?," International Journal of Bioelectromagnetism, pp. 115-116, 2011.

[13] F. Yger, M. Berar, and F. Lotte. "Riemannian approaches in brain-computer interfaces: a review." IEEE Transactions on Neural Systems and Rehabilitation Engineering 25.10, pp. 1753-1762, 2016.

[14] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," Brain-Computer Interfaces 4.3 pp. 155-174, 2017.

[15] A. Barachant, and Stephane Bonnet, "Channel selection procedure using Riemannian distance for BCI applications," 5th International IEEE/EMBS Conference on Neural Engineering, IEEE, 2011.

[16] O. Ledoit, and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," Journal of multivariate analysis, 88.2 pp. 365-411, 2004.

[17] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, “OpenViBE: An Open-Source Software Platform to Design, Test and Use Brain-Computer Interfaces in Real and Virtual Environments,” Presence, pp. 35–53, 2010.

[18] G. Dornhege, B. Blankertz, G. Curio, and K. R. Müller, “Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms,” *IEEE Trans. Biomed. Eng.*, pp. 993-1002, 2004.

[19] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Riemannian geometry applied to bci classification,” In Proc LVA-ICA, 2010.

[20] F. Nielsen and B. Rajendra, Matrix information geometry, Heidelberg: Springer, 2013.