Riemannian Channel Selection for BCI With Between-Session Non-Stationarity Reduction Capabilities

Khadijeh Sadatnejad and Fabien Lotte

***Abstract*— Objective: Between-session non-stationarity is a major challenge of current Brain-Computer Interfaces (BCIs) that affects system performance. In this paper, we investigate the use of channel selection for reducing between-session non-stationarity with Riemannian BCI classifiers. We use the Riemannian geometry framework of covariance matrices due to its robustness and promising performances. Current Riemannian channel selection methods do not consider between-session non-stationarity and are usually tested on a single session. Here, we propose a new channel selection approach that specifically considers non-stationarity effects and is assessed on multi-session BCI data sets. Methods: We remove the least significant channels using a sequential floating backward selection search strategy. Our contributions include: 1) quantifying the non-stationarity effects on brain activity in multi-class problems by different criteria in a Riemannian framework and 2) a method to predict whether BCI performance can improve using channel selection. Results: We evaluate the proposed approaches on three multi-session and multi-class mental tasks (MT)-based BCI datasets. They could lead to significant improvements in performance as compared to using all channels for datasets affected by between-session non-stationarity and to significant superiority to the state-of-the-art Riemannian channel selection methods over all datasets, notably when selecting small channel set sizes. Conclusion: Reducing non-stationarity by channel selection could significantly improve Riemannian BCI classification accuracy. Significance: Our proposed channel selection approach contributes to make Riemannian BCI classifiers more robust to between-session non-stationarities.**

***Index Terms*— Brain-computer interfaces, EEG, Riemannian manifold, channel selection, non-stationarity.**

Manuscript received February 19, 2021; revised September 28, 2021 and December 23, 2021; accepted January 15, 2022. Date of publication April 14, 2022; date of current version May 6, 2022. This work was supported by the European Research Council through the Project BrainConquest under Grant ERC-2016-STG-714567. *(Corresponding author: Khadijeh Sadatnejad.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Inria Ethical Committee, the COERLE under Approval Nos. 2019-12 (Cybathlon dataset) and 2019-04 (MIE dataset) and the Legal Authorities of INRIA Bordeaux Sud-Ouest, the COERLE, under Approval No. 2015-004 (MT dataset).

Khadijeh Sadatnejad is with the Inria Bordeaux Sud-Ouest, 33405 Talence, France (e-mail: sadatnejad.kh@gmail.com).

Fabien Lotte is with the Inria Bordeaux Sud-Ouest, 33405 Talence, France, and also with the LaBRI, CNRS, Bordeaux INP, University of Bordeaux, 33405 Talence, France (e-mail: fabien.lotte@inria.fr).

Digital Object Identifier 10.1109/TNSRE.2022.3167262

I. Introduction

A BRAIN-COMPUTER Interface (BCI) is a system that translates users' cognitive states and intentions, measured through neurophysiological signals, into commands for an interactive application [1]. Whereas BCIs have been mostly developed for interaction and communication of severely paralyzed people with the environment, they were also investigated for other purposes, for instance for driving and entertainment [2], [3]. Most BCI applications use neurophysiological signals recorded non-invasively, typically using Electroencephalography (EEG) [1]. A standard BCI system includes preprocessing (including temporal and/or spatial filtering), then represents EEG signals in a compact and discriminative way, then train a model for predicting the class label of trials or generating feedback [1], [8]. Different sources may lead to variations in EEG signals [4] and thus to change in the statistical distribution of description of the data within or between sessions. This includes switching the metastable state of the neural assemblies, changes in subjects' states such as fatigue or attention, changes in electrodes placement between different sessions, and different environmental noises [5]. These could lead to data distribution variations between or within sessions.

Data distribution variation over time (e.g., between training and test set) could strongly affect the BCI system performance, especially for real-life applications, where non-stationarity of the system is a common issue [6], [7].

In this study, we focus on reducing such non-stationarities at the machine learning level, by proposing a new preprocessing step dedicated to Riemannian geometry classifiers [9], for Mental Tasks (MT)-based BCI. We focused here on Riemannian classifiers of EEG covariance matrices due to: 1) the intrinsic properties of the spatial covariance matrix leading to smoothing noise, 2) the robustness of this representation for analyzing data contaminated by artifacts [10] (both of these items being possible sources of non-stationarity [8]), 3) their good performances even with small training datasets [9], 4) its success in multiple BCI challenges [9], and benchmark comparisons [26]. In particular, Riemannian classifiers demonstrated their superiority to other classifiers, notably Common Spatial Pattern (CSP) methods and/or Deep Learning classifiers, in the vast majority of comparative BCI studies [26], [36], [37], [40], [41] or in recent BCI competitions [9],

[38], [39], [42]. They are thus considered by many to be the current state-of-the-art EEG classifiers for most BCI paradigms. We propose here to improve them further by reducing their sensitivity to between-session non-stationarities.

Existing attempts for overcoming between-session non-stationarity in BCI can be categorized into two main approaches: 1) matching the distribution of data, e.g., by applying geometrical transformations such as translation or rotation [11], [12]; 2) modeling variability, for example using classifiers ensembles or adaptive approaches [8], [9], [14]. Modeling non-stationarity either by ensemble-based techniques or by a gradual update of the model in an adaptive approach [13], [27] could lead to a complex or variable model over time, respectively, which may hamper BCI user learning [28].

In this paper, we aim at dealing with the between-session non-stationarity problem at the preprocessing step, by excluding non-stationary sources from the Riemannian classifier input. More precisely, we investigate the possibility of reducing between-session distribution variations by removing channels that are mostly affected by non-stationarity.

A. Existing Channel Selection Methods for BCI

Due to the sensitivity of CSP spatial filters to noise [33] and even to a small number of outliers [30], channel selection approaches were considered to robustify it. Robustifying CSP by selecting channels in the first session, and testing it on other sessions led to comparable BCI performance as compared with selecting channels in each session [30], [33]. In that research, channels were selected based on EEG recorded from a single session, i.e., without between-session variability [30], [33] and processed by a classifier which was calibrated in each session. These works focused on possible data distortion variations within sessions. Besides, they did not address how to find the optimal number of channels to select.

A sparse CSP which selects the smallest number of channels within a constraint of classification accuracy was evaluated on datasets with a single training or test session [6]. Selecting channels on a single session prevents from capturing between-session non-stationarity effects. Experimental evidence showed increasing classification accuracies when removing channels as compared with all channels when the test and training sets were recorded in the same session. However, accuracy decreased when the training and test sets were recorded in two different sessions. This could be attributed to possible between-session non-stationarity effects [6], that were not addressed in that research.

Besides, transferring channel rankings based on CSP weights from a group of subjects to a new subject was also explored for BCI [31]. Despite promising results reported for motor imagery tasks, cross-subject channel selection does not quite achieve the performance of within-subject channel selection [31].

Selecting channels based on neurophysiological knowledge was also explored, although it could not necessarily lead to an improvement in BCI performance as compared with using all channels [16]. Arvaneh et. al., experimentally confirmed that channel selection based on sparse CSP [6] could lead to about 10% improvement in classification accuracy of motor imagery BCI as compared to using only C3, C4, and CZ channels.

In Riemannian geometry-based BCI approaches, Barachant et. al. [18] proposed a channel selection method without between-session nor within-session variation consideration, evaluated on a single session dataset. The subsets of channels were selected by considering the average distance between the means (M-M criterion) of each class pair:

$$\mathrm{M-M} = \sum_{i=1}^{N_c} \sum_{j=i+1}^{N_c} d_R(\bar{C}^{(c_i)}, \bar{C}^{(c_j)})/N_P \tag{1}$$

where $\bar{C}^{(c_i)}$ is the Riemannian mean of class c_i (see Eq. 3), N_P is the number of class pairs, d_R the Riemannian distance (see Eq. 2), and N_c the number of classes. This approach suffers from some limitations: 1) Data distribution between sessions might be affected by different geometrical transformations resulting from different non-stationarity sources. Preserving the channels that reduce between-class mean distances the least may not capture all necessary information for removing non-stationarity effects (e.g. shift between data distribution). Whereas using the distance between means is consistent with the Minimum Distance to Mean (MDM) classifier [25], considering within-class variations and choosing channels leading to a more compact representation around each class mean may also positively affect MDM performance. 2) For choosing the optimal number of channels, this approach considered the evolution of the criterion against the number of selected sensors and kept the minimum number of channels which keeps a constant fraction of distance between class means, using a threshold, without considering the subject and dataset dependency. In other words, only keeping a fraction of distance between means, without considering how much it is affected by undesirable sources that should be removed, is likely to lead to suboptimal channel selection by removing discriminative information or keeping non-stationarity effects. Thus, besides the sensitivity to the threshold value, using the M-M criterion may not lead to optimal performance due to a lack of variability information.

Most current BCI channel selection methods, notably in Riemannian frameworks, were proposed and tested without considering between-session variations [8], [18], [29], [30], [31], [33]. There is thus a need to study methods to reduce between-session variability using channel selection for Riemannian geometry-based approach due to: 1) substantial session-to-session variability that is a major issue, preventing the design of robust BCI, i.e, a system calibrated on multiple sessions and tested on subsequent sessions [7], 2) lack of a Riemannian channel selection approaches considering non-stationarity.

In this paper, we thus propose several (offline) channel selection methods for Riemannian BCI classifiers that specifically aim at reducing between-session non-stationarities. It should be noted that we presented preliminary research on two class problems, with evaluations on two data sets, as a short (4 pages) conference paper in [22]. Contrary to this previous work, here we explore multiclass problems, on three

data sets, we study the optimality of the channel subset search strategy, propose a new search stopping strategy, propose a measure to predict the benefits of the proposed channel selection method, and investigate differences between different criteria, including two new ones that were not presented in this preliminary study.

B. Our Contribution

In this study, we choose covariance matrices as EEG descriptors and a Riemannian framework for analysis. We consider subject-specific between-session non-stationarity effects in choosing the optimal channel subset. This requires quantifying between-session non-stationarity effects and automatically selecting the number and location of the most useful channels. This should lead to a BCI with better generalization abilities across sessions, that is also less complex, and thus with an easier-to-train classifier.

For quantifying non-stationarity effects, we propose different criteria over the Riemannian manifold of covariance matrices. Such criterias consider unbalanced class distribution that may arise in multi-class problems, as well as non-stationarity effects that can be either shift in data distribution between sessions or distortion of the class boundaries. We choose Sequential Floating Backward Selection (SFBS) [21] for optimal channel subset selection based on our proposed criteria. We also adapt the stopping criterion of SFBS to consider possible between-session variability of the data. We aim at continuing to remove channels while between-session non-stationarity effects can still be measured in the remaining channels. We also propose a method to predict in advance whether channel selection would be useful for a given subject, from training EEG data only.

The remainder of this paper is organized as follows: Section II reviews the principles of Riemannian geometry of covariance matrices. Then it describes our proposed criteria, the search strategy used for selecting channels, and our proposed approach for selecting the optimal subset size. In section III we evaluate what we proposed experimentally and discuss the results and finally, we conclude and summarize findings in section IV.

II. Method

A. Riemannian Manifold of Covariance Matrices

This framework relies on representing trials using their spatial covariance matrix, i.e., $C = \frac{1}{T-1}(SS^T)$, where S is a band-pass filtered EEG signal (an$n \times T$ dimensional matrix, where T is the number of observations in each window or trial and n is the number of channels) and C is the emprical estimate of the covariance matrix (an $n \times n$ matrix). It performs analyses based on their non-linear geometry in a Riemannian framework. We use the optimal linear shrinkage of the spatial covariance matrix of each trial [32] for describing it, denoted by $C_i^{(c)}$, where i denotes the trial number and c denotes its corresponding class. $C_i^{(c)}$ is thus an $n \times n$ Symmetric Positive Definite (SPD) matrix. With this representation, 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:

$$d_R\left(C_i^{(c_1)}, C_j^{(c_2)}\right) = \left\|\log\left(C_i^{(c_1)^{-1/2}} C_j^{(c_2)} C_i^{(c_1)^{-1/2}}\right)\right\|_F \quad (2)$$

where $C_i^{(c_1)}$ and $C_j^{(c_2)}$ are the covariance matrices of the i^{th} and j^{th} trial of class c_1 and c_2 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 $\bar{C}^{(c)}$:

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

This point has the minimum squared distance from all points of the class. There is no closed-form solution for this equation, but it can be computed using an iterative algorithm [25]. The standard deviation (std) $\sigma^{(c)}$ within each class c can be defined as follows [16]:

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

B. Channel Selection

We first introduce the different criteria we propose to select or reject channels. Then, we describe the search strategy we use for selecting channel subsets offline, our proposed method for stopping the search algorithm, and finally our efficiency predictor, predicting the usefulness of channel selection.

1) Selection Criteria:

a) Average of intra-class variation: Non-stationary could lead to shifts in data distribution between sessions (Fig.1) [22]. This would lead to large intra-class variation between different sessions as compared to within-session variations. We thus propose the Average of Intra-class Variation (AIV) to measure how much the data is affected by non-stationarity. AIV should thus be minimized.

$$AIV = \frac{\sum_{c_i} \sigma^{(c_i)}}{N_C} \quad (5)$$

b) Mean to mean distance/intra-class variation across pairs of classes: Using the Mean-to-Mean distance divided by the intra-class Variation across Pairs of classes (M-M/VP) criteria, we consider both discriminative information (which may be affected by geometrical transformation like scaling and rotation that can change the boundaries between classes by overlapping) and the intra-class variation (which may be affected by shifts between sessions). This metric is a Riemannian extension of the Fisher criterion [15] over the manifold of SPD matrices. To extend the criterion to multi-class problems, we compute the average over all possible pairs of classes. Our objective is to maximize the criterion, formulated as follows:

$$\text{M} - \text{M/VP} = \sum_{i=1}^{N_c} \sum_{j=i+1}^{N_c} \frac{1}{N_P} d_R\left(\bar{C}^{(c_i)}, \bar{C}^{(c_j)}\right) / (\sigma^{(c_i)} + \sigma^{(c_j)}) \quad (6)$$

where N_P is the number of class pairs, i.e., $N_C(N_C - 1)/2$.

c) Mean to global mean/ intra-class variation: For the Mean-to-Global Mean divided by the intra-class Variation (M-GM/V) criterion, we extended the basic criterion (i.e., the criterion for two classes) for the multi-class problems by computing the relation of all the classes to a specific point (e.g., the geometric mean of all training samples $\bar{C}$) as follows:

$$\mathrm{M-GM/V} = \sum_{c_i} d_R\left(\bar{C}^{(c_i)}, \bar{C}\right) / \sum_{c_i} \sigma^{(c_i)} \tag{7}$$

This criterion, measuring class separation in a multi-class problem, is quite similar to the discrimination score *classDis* proposed in [17]. In this paper, we use M-GM/V both as a selection criterion and as a BCI performance evaluation metric.

2) Search Strategy: The Sequential Floating Backward Selection (SFBS) (Algorithm 1) is a top-down search procedure that excludes channels based on sequential backward selection (SBS) [21]. It starts from the current channel set and is followed by a series of successive conditional inclusions of the most significant excluded channels if an improvement can be made to the previous sets. Here, the most significant channels mean the channels maximizing or minimizing (depending on the criterion) the criterion value when adding or removing them. The algorithm starts with all channels, then SBS [21] runs until the two least significant channels are excluded [21]. Then, the algorithm continues with Step 1. This method could correct wrong decisions made in the previous steps to approximate the optimal solution as much as possible. This possibility of correction is the main difference between SBS (used for existing Riemannian channel selection methods) and SFBS (that we propose here to apply to this problem). SFBS achieves "more optimal" subset selection at the expense of computational time, and it is affected by the selection and stopping criteria. Excessive channel removing could lead to different problems like under-fitting, unbalanced class discrimination, and possible effects of noise, depending on the criterion, for the resulting classifier. Thus, improving performances by removing non-stationarity sources also requires a suitable stopping criterion.

3) Stopping Criterion: Since we aim at removing non-stationarity effects between sessions, we adapt the search algorithm stopping condition for this aim. For channel selection, due to variation between sessions/runs, resulting from different non-stationarity sources, stopping based on accuracy on the training/validation set may not meet optimality on the test set. Besides, such accuracy would be classifier-dependent. Moreover, classifier-specific channel subset selection is computationally demanding.

Instead of using a threshold on training/validation classification accuracy [21] (e.g., adapting a threshold based on performance on a previously selected subset with a termination tolerance [20]), we consider removing the variations resulting from non-stationarity sources independently of a classifier.

For deciding about the search stopping strategy, our idea relies on the assumption that between-session variations are longer-term and different than within-run variations. For instance, change in cap positioning or long-term user learning should lead to between-session non-stationarity but not within-run (in a single session) non-stationarity. Besides, we should observe between-session non-stationarity effects in a training set composed of multiple sessions. Therefore, we use the difference between the whole, multi-run and multi-session training set AIV with the selected channels and run-level AIV as an indicator of the existence of between-session non-stationarity effects in the selected channel set. We continue excluding channels until AIV for the training set of the selected subset is greater than an automatically determined threshold, which represents run-level AIV. For each subject, we use $mean_{Run_{i,j} \subset Tr}(AIV\left(Run_{i,j}, \bar{X}_0\right))$ as run-level AIV:

Since the threshold value strongly affects the selected subset and AIV is not perfect for representing all possible variations, our proposed options may not always reach the real optimal

Algorithm 1 SFBS [21]:

Input :
$k = 0$ // Iteration number
$\bar{X}_k = X_0$ // $\bar{X}_k$:remaining channel subset at kth iteration, initialized by X_0
//X_0 : *the set of all channels*
$J\left(\bar{X}_k\right)$ // Selection criterion (M-M, 1/AIV, M-M/VP, M-GM/V) value of $\bar{X}_k$ channel set over a training set
Output:
$\bar{X}_k$ //selected channel *set* at kth iteration

Step 1: //Exclusion step: remove less significant channel if stopping condition is not satisfied
If Checkstop(k, $\bar{X}_k$)
return //$\bar{X}_k$ would be the selected channel set
end
$x_{k+1} = argmin_{x_i \in \bar{X}_k} J(\bar{X}_k - \{x_i\})$ //x_{k+1}: less significant channel (J: 1/AIV, M-M, M-GM/V, M-M/VP)
$\bar{X}_{k+1} = \bar{X}_k$ - $\{x_{k+1}$ //remove less significant channel, x_{k+1}, from selected channel subset at kth iteration, $\bar{X}_k$.

Step 2: //Conditional inclusion: add most significant previously removed channel
//x_r: The most significant previously excluded channel
$x_r = argmax_{x_i \in \bar{X}_0 - \bar{X}_k} J(\bar{X}_k \cup \{x_i\})$
If $x_r == x_{k+1}$ //if most significant previously excluded channel is the last removed channel, go to next iteration
  $k = k + 1$
  Go to step 1
else //otherwise add it to current selected set
    $\bar{X}'_k = \bar{X}_k \cup \{x_r\}$ $\bar{X}'_k = \bar{X}_k \cup \{x_r\}$
    If $k == 2$ //If we have only one previously removed channel, continue with removing new channels
      $\bar{X}'_k == \bar{X}_k$
      Go to step 1
    else // otherwise check other excluded channels for inclusion
      Go to step 3
    end
end

Step 3: //Continuation of conditional inclusion
//x_s: The most significant excluded channel
$x_s = argmax_{x_i \in \bar{X}_0 - \bar{X}'_k} J(\bar{X}'_k \cup \{x_i\})$
If $J\left(\bar{X}'_k \cup \{x_s\}\right) \leq J\left(\bar{X}_{k-1}\right)$
  $\bar{X}'_k = \bar{X}_k$
  Go to step 1
else
  $\bar{X}'_{k-1} = \bar{X}'_k \cup \{x_s$
  $k = k - 1$
  If $k == 2$
    $\bar{X}_k = \bar{X}'_k$
    Go to step 1
  else
    Repeat step 3
  end
end

Algorithm 2 Check Stop

```
//Finding an optimal/near-optimal number of channels
Input: k, X̄_k        //X̄_k: Selected channel subset at kth iteration
Tr                   //The training set
X̄_0                  //All channels
Run_{i,j}            //jth run of ith session of training set
Output: Stop or continue condition
If AIV(Tr, X̄_k) > mean_{Run_{i,j} ⊂ Tr}(AIV(Run_{i,j}, X̄_0))
return false         //Continue channel selection
else                 //X̄_k and its size are near-optimal channel subset and
near-optimal number of channels
return true          // Stop channel selection
end
```

classification accuracy. This is why we call it a near-optimal number of channels in the following.

4) Efficiency Predictor: To predict possible performance improvement when using channel selection, we quantify the existence of between-session non-stationarity effects in the dataset by comparing the *AIV* at the run level to the *AIV* at the whole training set level, where the training set includes possible between-session variations. The size of the gap between run-level *AIV* and training set level *AIV* could be used as an efficiency predictor of channel selection. Similar *AIV* in run-level and training set level could reflect low between-session variability, and thus a low chance of improvement by using channel selection, whereas a large *AIV* gap could reflect an opportunity for improvement by removing the channels that are mostly affected by between-session non-stationarity. Since we want to reveal the difference between within-session and between-session *AIV*, we use $max_{Run_{i,j} \subset Tr}\left(AIV\left(Run_{i,j}, \bar{X}_0\right)\right)$ as an indicator of run-level non-stationarity effects. Therefore, we define $AIV(Tr, \bar{X}_0)$-$max_{Run_{i,j} \subset Tr}(AIV(Run_{i,j}, \bar{X}_0))$ as a potential efficiency predictor to be tested. $Tr = \bigcup_{i=1}^{N_t} \bigcup_{j=1}^{N_r} Run_{i,j}$ denotes a training set composed of $N_t * N_r$ runs recorded in N_t sessions and N_r runs in each session and $\bar{X}_0$ denotes the set of all channels. We will correlate this efficiency predictor with test set classification accuracy gain when using channel selection.

III. Evaluations

A. Datasets

For evaluation, we used three different datasets, all of them recorded on different days/sessions. All three studies were approved by the Inria ethical committee, the COERLE (approval numbers 2015-004, 2019-04, and 2019-12).

1) Data Set 1: Mental Tasks (MT) Data Set: Sixteen BCI-naive participants including eight females; aged 21.5 ± 1.2 with subject ID S12: S19 and eight males with subject ID S21, S23: S29 took part in this study, published in [34]. The EEG signals were recorded at 512Hz from two g.USBamp amplifiers (g.tec, Austria), using 30 scalp electrodes (F3, Fz, F4, FT7, FC5, FC3, FCz, FC4, FC6, FT8, C5, C3, C1, Cz, C2, C4, C6, CP3, CPz, CP4, P5, P3, P1, Pz, P2, P4, P6, PO7, PO8, Oz, 10–20 system), referenced to the left ear and grounded to AFz. Each subject participated in 6 sessions; each session was on a separate day with irregular gaps between them. Each session was composed of five 7-minutes runs in which users had to learn to perform mental tasks including Mental Calculation (MC), Mental Rotation (MR), and Left Hand movement Imagination (LHI) tasks. Each run was composed of 15 trials per mental task, presented in a random order.

2) Data Set 2: Motor Imagery-Execution Tasks (MIE) Data Set: Sixteen MT-BCI naïve participants were included in this study (aged 22.31 ± 2.33) [23]. EEG signals were recorded using 11 scalp electrodes (FC4, C4, CP4, C6, FC3, C3, CP3, C5, C1, Cz, C2, 10–20 system). The participants completed 10 sessions; with different sessions being recorded on different days. Each session was composed of seven runs. Each run included 10 trials per task. The tasks were left or right hand motor imagination (LHI/RHI), right or left hand motor execution (RHE/LHE - with the other hand than the imagination), and rest. Two types of feedbacks, either visual feedback (5 sessions) or visual and vibrotactile feedback (5 sessions) were used. Here, we used the first 6 sessions of each subject: 4 sessions for training (all with the same feedback), and 2 sessions for testing (one session with each type of feedback, thus with possible non-stationarities), as for the MT data set.

3) Data Set 3: CYBATHLON Dataset: This dataset was recorded from a tetraplegic user during multiple sessions [35]. The sessions took place on several days at the user'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 electrodes (FC6, C4, FC4, C6, CP6, T8, FC5, C3, C2, Cz, C1, C5, T7, CP5, FC3, PO4, P2, CP4, P4, CP1, CPz, Pz, CP2, O2, Oz, POz, O1, PO3, P1, P3, CP3, F1, F2, AFz, Fz, FC2, FCz, FC1, F8, F4, AF4, FP2, F7, F3, FP1, AF3, 10–20 system). Our experiments used data from 9 sessions (i.e, all multiclass and labelled sessions), called Se10, Se11, Se12, Se13, Se14, Se15, Se16, Se17, and Se18. Each run of this dataset comprised 10 trials per mental task (MC, LHI, RHI, and REst (RE)).

B. Data Pre-Processing

For all three datasets EEG signals were band-pass filtered in 8-30 Hz using a 4^{th} order Butterworth filter. The location and length of the extracted epochs varies depending on the experimental setup of each dataset. We extracted windows from the feedback interval, i.e., 3.75 second and 4-second window from 1250ms after task cue onset for CYBATHLON and MT dataset respectively. For MIE dataset, 2-seconds windows were extracted for rest trials from 500 ms after the rest onset (an interval without feedback) and for LH/RH Execution/ Imagination from 2250 ms from task stimulus onset (selected from feedback interval). In all datasets, to prevent bias, we extracted the same number of epochs for each class. Besides, we rejected all epochs for which band pass-filtered EEG had amplitude higher/lower than $\pm 70\ \mu$V.

C. Experiments

We performed experiments to 1) study possible relationships between channel number and non-stationarity, 2) compare the behavior of different selection criteria across channel numbers, 3) evaluate BCI performance using channel selection, and 4) evaluate our proposed efficiency predictor. As the evaluation

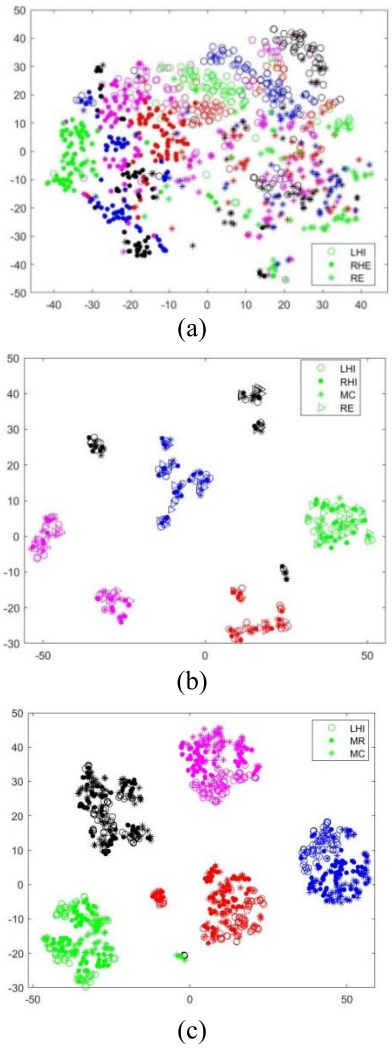


Fig. 1. Visualization of covariance matrices using t-SNE. Different colors represent different sessions (Session 1 to 5 in MT and MIE and session 14 to session 18 are in red, blue, magenta, black, green colors respectively), different shapes represent different classes. (a), (c) All runs of 5 sessions of a randomly selected subject from the MIE and MT datasets, respectively (b) 5 sessions from the CYBATHLON dataset. Two dimensional representations after channel selection are shown in appendix (D).

criteria, we used the classification accuracy of a Riemannian MDM classifier [9] hereafter. In all our experiments, we used the same number of sessions, i.e, first 4 sessions as training set and last 2 sessions as test set, for all subjects of MIE and MT datasets. For CYBATHLON dataset we used sessions Se10-Se13 as training set and sessions Se14-Se18 as test set.

1) Between Session Non-Stationarity Effects: First, we investigated the effects of between-session non-stationarity on data distribution. Fig. 1. shows a two-dimensional representation of data for randomly selected subjects of (a) MIE dataset, (c) MT dataset, and (b) sessions Se14-Se18 of the CYBATHLON dataset. Different colors show different sessions and different shapes show different classes/tasks. We used t-SNE [24] to project the data points (EEG covariance) in two-dimensional space. In t-SNE, to consider the non-linear geometry of the space, we used as custom distance, the Riemannian distance (Eq. 2). We set the Perplexity parameter, which determines the effective number of local neighbors of each point [24], to the number of epochs extracted from each task in each run (i.e., 10 in the case of MIE and CYBATHLON datasets and 15 for MT dataset). Fig. 1. (a) shows considerable overlapping between data recorded in different sessions, suggesting less non-stationarity effects as compared with (b) and (c), which show considerable shifts between different sessions, suggesting high non-stationarities. Different reasons could explain these differences in between-session data distribution variations across datasets, e.g., differences between subjects' ability to control the BCI or differences in the mental tasks used. However, similar patterns between datasets (i.e., less shift in most subjects of the MIE dataset and more between-session variation in the MT and CYBATHLON dataset) may suggest larger non-stationarity effects with more channels. Indeed, the MIE data set only has a small number of channels (11), only located over sensorimotor areas, and little non-stationarity, contrary to the other two which have many more channels (30 and 46), all over the scalp, and display much larger non-stationarities.

To investigate differences in between-session non-stationarity across datasets, we compared the AIV for a whole training dataset versus the median value of AIV across the runs included in this set. For the MIE and MT datasets, the median of the increase in AIV for the training set vs. the AIV for the runs included in the training set is 1.29% and 24.96% respectively. For the CYBATHLON dataset the increase is 20.18%. When keeping the same dimensionality across data sets, e.g., No. of channels $=$ 11 (when channel are selected based on the AIV criterion) the AIV increase is 20.18% for the MT dataset and is 16.64% for the CYBATHLON dataset, i.e., a smaller non-stationarity than with more channels. The nonetheless larger non-stationarity of these two data sets may also suggest that the channel location spread also matters.

These comparisons could thus suggest that there are more non-stationarity effects for datasets with more channels and/or with channels that are more spread across the scalp.

2) Selection Criteria Specification: To investigate the properties of selection criteria across channel set sizes, we considered the changes in BCI performance in terms classification accuracy across all possible channel set sizes. Here, we thus did not use a stopping criterion.

Fig. 2 illustrates the test classification accuracy of subjects from the MT dataset across channel numbers and their average over subjects, where channels were selected based on AIV and M-GM/V criteria. To reduce the complexity of each figure, we plotted only 8 subjects (see Fig. 3. for average of 16 subjects and Table III for all subjects' performance). Due to the optimal channel subset size being highly subject specific, the average accuracy curve, which smooths the subjects' accuracy peaks, does not illustrate the actual efficiency of channel selection. The mean and standard deviation of the optimal number of channels for the MT dataset (marked in Fig. 2) for AIV and M-GM/V-based channel selection are 20.66$\pm$5.94 and 15.66$\pm$7.30 respectively. The dispersion in optimal channel subset size between subjects and based on different criteria confirms the necessity of tuning this

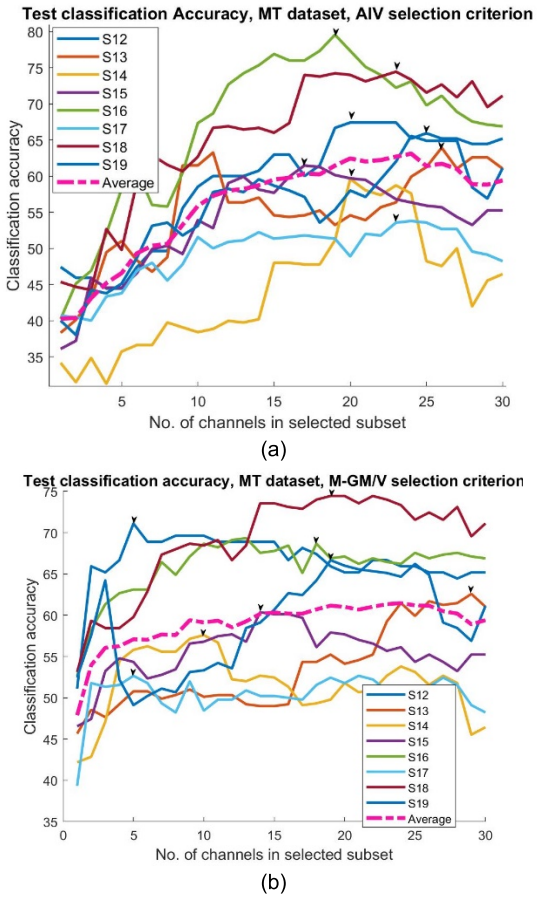


Fig. 2. Test classification accuracy of 8 subjects from MT dataset and their average across channels set size, where channels selected based on a) AIV, b) M-GM/V selection criteria.

parameter for each subject and criterion independently. For AIV criterion, the fast increase in performance (i.e., on average for the 10 first removed channels of the MT dataset over all 16 subjects) should be due to removing the channels mostly affected by non-stationarity sources. Using AIV as a selection criterion for small channel set sizes shows a drop in classification accuracy, e.g., from 66% at peak to 37.92% for only one channel, averaged over all subjects of MT. This drop in performance is probably due to considering only the variance in the channel selection, which finally keeps a channel subset with the lowest variance. This subset is not necessarily the most discriminative one. Considering the distance between means for M-M, M-M/VP, and M-GM/V criteria could lead to more stability in lower dimensions as compared with AIV (see Fig. 3. (a) (b)), as confirmed by statistical tests (see below).Comparing M-M/VP and M-GM/V criteria shows that there is a clear distinguishing pattern. Selection based on the M-GM/V criterion on the MT dataset, on average leads to lower classification accuracy in the last iterations as compared with the M-M/VP criterion (Fig. 3. (a)). Readers interested in how the M-GM/V criteria evolves with the number of selected channels or the channel selection criterion, globally or between pairs of classes, can refer to Appendices A. and B.

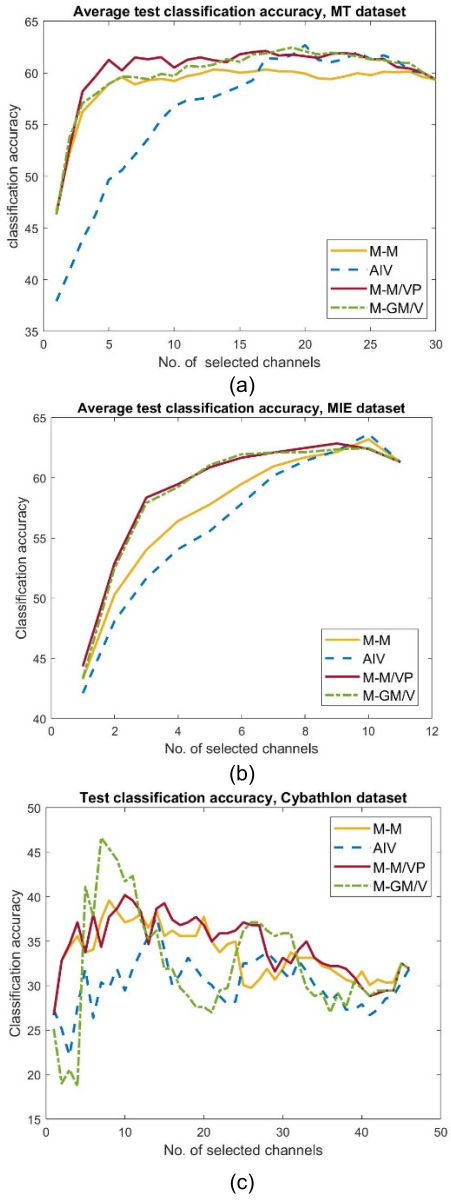


Fig. 3. Average test classification accuracy over subjects across channel subset size (a) MT and (b) MIE datasets; (c) CYBATHLON dataset.

To evaluate the different criteria statistically, we compared the average classification accuracy they obtained over all subjects of the three datasets, when selecting either 20%, 40%, 60% or 80% of all channels in each dataset, using a Two-way repeated measure ANOVA with Tukey correction for Post-Hoc analysis. The ANOVA confirmed the significant difference between criteria ($p<0.001$) and the significant interaction between criteria and channel subset size ($p < 0.001$).

Post-hoc tests showed that M-M/VP and M-GM/V lead to significantly better accuracy than M-M with $p < 0.001$ and $p < 0.01$ respectively. M-M/VP and M-GM/V are also both significantly superior to AIV ($p<0.001$). There is no significant difference between AIV and M-M ($p > 0.05$).

The difference between M-M/VP and M-GM/V is not significant either ($p>0.5$).

TABLE I
CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON THE CYBATHLON DATASET FOR NEAR-OPTIMAL CHANNEL SUBSET

	M-M	AIV	M-M/VP	M-GM/V	All channels
MDM	37.73	31.60	36.20	36.20	31.90
Channel No.	20	30	24	28	46

TABLE II
AVERAGE CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON MT DATASET AT NEAR-OPTIMAL CHANNEL SUBSET

	M-M	AIV	M-M/VP	M-GM/V	All channels
Mean Accuracy	61.77	**62.42**	61.84	61.55	59.36
Avg. Channel No.	16.1±4.2	22.7±2.6	22.3±2.9	21.5±3.4	30

TABLE III
AVERAGE CLASSIFICATION ACCURACY AND NO. OF CHANNELS ON THE MIE DATASET AT A NEAR-OPTIMAL CHANNEL SELECTED BASED ON DIFFERENT CRITERIA AND FOR ALL CHANNELS

	M-M	AIV	M-M/VP	M-GM/V	All channels
Mean Accuracy	59.98	60.00	61.06	60.99	**61.31**
Avg. Channel No.	7.5±1.4	8.9±0.5	8.7± 0.6	8.7±0.6	11

TABLE IV
AVERAGE CLASSIFICATION ACCURACY WITH CHANNEL SUBSET SELECTED USING LEAVE-ONE SESSION-OUT CROSS-VALIDATION

	M-M	AIV	M-M/VP	M-GM/V
MT	58.63	**61.48**	60.46	60.34
MIE	59.80	59.97	**60.60**	60.11
CYBATHLON	33.74	28.22	39.36	**45.4**

TABLE V
CORRELATION COEFFICIENT AND P-VALUE RESULTING FROM THE PEARSON CORRELATION BETWEEN CLASSIFICATION ACCURACY IMPROVEMENT WITH CHANNEL SELECTION AND OUR PROPOSED EFFICIENCY PREDICTOR

	M-M	AIV	M-M/VP	M-GM/V
Correlation Coefficient	0.47	0.47	0.43	0.48
p-value	0.01	0.01	0.01	0.005

Regarding the interaction between channel set size and criteria, considering only variance in AIV led to significantly lower performance as compared to M-M/VP ($p<0.001$) and M-GM/V($p<0.001$) for 20% and 40% channel subset (i.e., with smaller channel subsets). M-M/VP and M-GM/V are significantly better than M-M for 20% and 40% subsets ($p<0.05$). AIV and M-M do not show significant differences neither in low nor high number of channels.

3) Evaluation of the Proposed Stopping Criterion: In this sub-section, we report the BCI performances obtained when using the proposed stopping criterion to select the near-optimal channel number, based on Algorithm 2.

For the CYBATHLON dataset, applying channel selection based on different selection criteria, using 20 out of 46 channels, could lead to up to 5.83% of improvement in test classification accuracy, as compared with using all channels (Table I). For the MT dataset (Table II – see Appendix E and Table IX for the subject-specific results), applying the repeated measure ANOVA to compare test classification accuracy after channel selection based on different criteria and for all channels confirmed the significant difference between the means of different approaches ($p<0.005$). Applying post-hoc analysis with Tukey correction confirms the significant superiority of AIV (with an average number of channels over subjects of about 22 out of 30) as compared with all channels with $p<0.005$ and the significant superiority of other criteria as compared with all channels with $p<0.05$. The performance difference between the criteria at near-optimal channel subset size is not significant.

Applying repeated measure ANOVA on test classification accuracy after channel selection using different criteria on the MIE dataset (Table III - see Appendix E and Table X for the subject-specific results) led to non-significant loss of performance ($p > 0.05$). Applying post-hoc analysis confirmed that selecting the subsets based on our proposed criteria select a subset without significant loss of performance ($p>0.5$ for M-M/VP and M-GM/V with on average; $p>0.1$ for M-M and AIV criterion). For MIE dataset, the average number of selected channels over subjects for AIV, M-M/VP, and M-GM/V criteria was about 9 and for M-M was 7.5 out of 11 channels.

To further evaluate our proposed stopping criteria, we also compared it to using leave-one-session out cross-validation accuracy (over the training set) for selecting the optimal channel subset size. With this approach, the channel subset size which led to the maximum training set cross-validation accuracy was selected as the optimal subset size for each subject. We report the result of test classification accuracy averaged over all subject of each dataset in Table IV (for more details see Appendix C). Applying one-way repeated measure ANOVA between the accuracies obtained by different selection criteria and all channels over all subjects did not show significant difference between selection by different criteria and all channels neither for the MT dataset ($p > 0.05$) nor the MIE dataset ($p > 0.05$). On the CYBATHLON dataset, a 13.5% improvement was obtained using M-GM/V, whereas this improvement for M-M was only 1.8%. As compared to using Algorithm 2 for the stopping criterion, using cross-validation also led to lower average performances on all data sets, except the CYBATHLON data set.

4) Efficiency Predictor: We evaluated our proposed efficiency predictor on the MIE and MT datasets. We computed the Pearson correlation between the predictor value and the improvement in classification accuracy over all subjects

across both datasets. The results confirmed statistically significant correlations between the predictor and the improvement in classification accuracy, for all channel selection criteria, see Table V.

IV. Conclusion

In this article, we have presented different criteria for channel selection in the Riemannian framework of covariance matrices for BCI application. Our main goal was to consider between session non-stationarity in order to reduce it. We compared the behavior of different criteria for different numbers of channels on three datasets with different levels of non-stationarity.

We obtained about 5% improvement in MDM classification accuracy at a near-optimal number of channels, on the CYBATHLON dataset while keeping 20-28 channels (for different criteria) from 46 channels, and a no-significant loss in the case of the MIE dataset with a small number of channels (11 channels). Channel selection on MT dataset leads to a significant improvement, 3.1% on average, at a subject-specific near-optimal number of channels.

Subject-specific and application-specific channel selection strategies are thus feasible and useful by using the proposed channel selection and stopping criteria. Comparing different criteria revealed that:

- Minimizing AIV, by removing channels mostly affected by non-stationarity (i.e., channels representing most of AIV) could lead to a fast increase in classification accuracy, then a drop in performance. The latter is due to omitting the discriminative information by only focusing on dispersion information, and keeping low variance channels. Our proposed stopping criteria could address that issue, and led to significant improvement with channel selection based on the AIV criterion.

- Using M-M – the current state-of-the-art for channel selection on the Riemannian manifold - on average led to lower performance as compared with the other proposed criteria that consider both discriminative and dispersion information, both for the same number of channels and with few channels. At the near-optimal number of the channels, the results from M-M are comparable with other criteria. For channel subsets selected based on leave-one session-out cross-validation over training set, M-M showed lower performance as compared with M-M/VP and M-GM/V.

- Considering both the discriminative and dispersion information in M-GM/V and M-M/VP criteria led to a non-significant loss in classification accuracy while using only a small number of channels, both in datasets with a large or a small number of channels. They can thus be used to reduce BCI setup time (electrodes mounting) and computational complexity, and increase user comfort without significant loss of performance.

- Comparing different criteria across different channel subsets over different datasets with different high and low non-stationarity effects confirmed the statistically significant superiority in classification accuracy after channel selection of our proposed criteria over M-M. For channel subsets with small channel numbers, AIV and M-M show statistically significantly lower performance as compared to M-M/VP and M-GM/V. In the subsets with higher number of channels, the difference between criteria is non-significant. The differences when comparing M-M/VP vs M-GM/V and AIV vs M-M are not statistically significant. We thus suggest to use M-M/VP and M-GM/V when reducing the number of channels in our application.

- Selecting a near-optimal number of channels by comparing AIV at run-level and training set-level can reveal significant improvements in accuracy using channel selection in the case of high non-stationarity effects or non-significant loss in the case of lower non-stationarity effects for all criteria. Although using cross-validation over the training set as stopping criterion could lead to superior performances for some subjects, it could not lead to statistically significant improvement when using channel selection when the data suffered from non-stationarity.

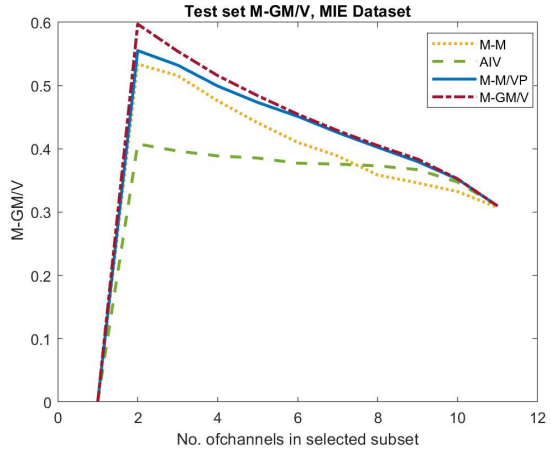
- Our proposed stopping criterion could identify a channel subset size that is close to the subject-specific optimal number of channels. However, it is still not perfect since it does not consider the effects of all geometrical transformation (e.g., rotation) and within-run non-stationarity effects. Defining a stopping criteria that could capture all the potential benefits of channel selection for improving BCI remains an open problem.

It should also be stressed that the benefits of the proposed (offline) channel selection method for Riemannian classifiers go beyond reducing non-stationarity. Indeed, as any channel selection method for BCI, it also 1) reduces model complexity and thus (online) computational demand; 2) reduces BCI setup time, thanks to fewer electrodes to place; and 3) increases users' comfort with fewer electrodes to wear.

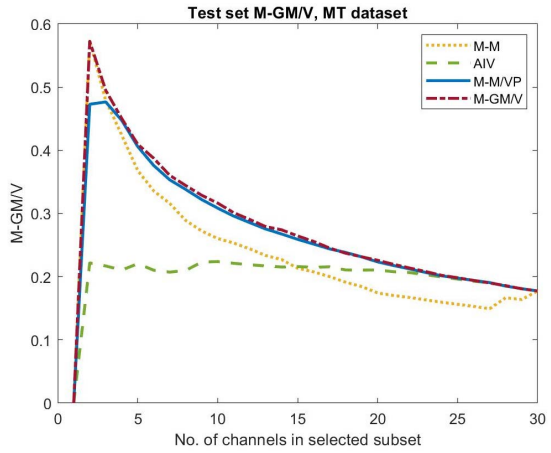
Finally, by reducing the number of parameters in the BCI model, channel selection could lead to a less complex model. Lower complexity could contribute to the development of a BCI with better generalization capabilities. More generalization, beside the better test classification accuracy, could generate more precise feedback, more related to task-related brain activity and less corrupted by unrelated patterns, e.g., physiological or extra-physiological non-stationarity sources. Besides, the accuracy considered in the comparisons are calculated offline on signals modulated by users' strategies based on online feedback. This means subjects could behave differently (better or worse) when training with the proposed algorithms in actual online evaluations. Therefore, investigating channel selection for online BCI user training would be an interesting future work.

Appendix A

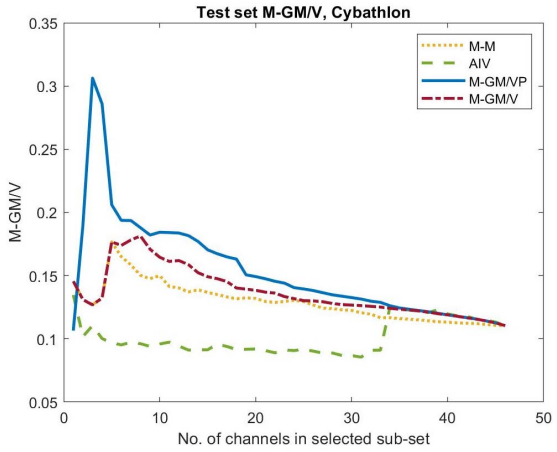
Due to the increasing trend in discrimination score in M-GM/V-based selection, it may seem that this trend may be only a side effect of comparison in different dimensional spaces. In the following, we compute a lower and upper bound for discrimination score (i.e., M-GM/V criterion). Since both upper and lower bounds have a similar relation to the number of channels (N) we could conclude the same behavior for



(a)



(b)

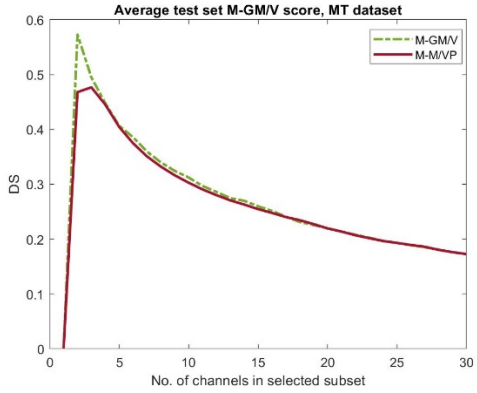


(c)

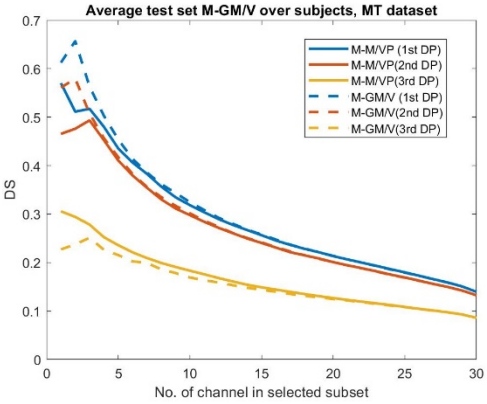
Fig. 4. Average M-GM/V of the test set over all subjects of (a) MIE, (b) MT, and (c) CYBATHLON dataset, channel subset were selected by AIV and M-GM/V criteria.

criterion value. In the next step we show that the increasing and decreasing trend in upper and lower bound of the criterion is not proportional to N, it is controlled by the selection criteria.

$$M-GM/V=\frac{\sum_{c_i} d_R\left(\bar{C}^{(c_i)},\bar{C}\right)}{\sum_{c_i}\sigma^{(c_i)}} \tag{8}$$



(a)



(b)

Fig. 5. (a) M-GM/V score of test set after selection based on M-M/VP vs. M-GM/V criteria. (b) Average test set M-GM/V over subjects of MT dataset within 1st, 2nd, and 3rd most Discriminative Pairs (DP) of the classes at different channel set size.

We start by expanding the numerator of M-GM/V for a very simple state, 2-dimensional data:

$$\begin{aligned} & d_R\left(\bar{C}^{(c_k)},\bar{C}\right) \\ &= \left\|\log\left(\bar{C}^{(c_k)^{-\frac{1}{2}}}\bar{C}\bar{C}^{(c_k)^{-\frac{1}{2}}}\right)\right\|_F \\ &= \left\|\log(M^{c_k})\right\|_F \\ M^{c_k} &= \begin{bmatrix} \bar{C}^{(c_k)}{}_{11}^{-\frac{1}{2}} & \bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}} \\ \bar{C}^{(c_k)}{}_{21}^{-\frac{1}{2}} & \bar{C}^{(c_k)}{}_{22}^{-\frac{1}{2}} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \bar{C}_{11} & \bar{C}_{12} \\ \bar{C}_{21} & \bar{C}_{22} \end{bmatrix} \times \begin{bmatrix} \bar{C}^{(c_k)}{}_{11}^{-1/2} & \bar{C}^{(c_k)}{}_{12}^{-1/2} \\ \bar{C}^{(c_k)}{}_{21}^{-1/2} & \bar{C}^{(c_k)}{}_{22}^{-1/2} \end{bmatrix} \end{aligned} \tag{9}$$

Each element of M^{c_k} matrix is as follows:

$$\begin{aligned} M_{11}^{c_k} &= (\bar{C}^{(c_k)}{}_{11}^{-\frac{1}{2}}\bar{C}_{11}+\bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}}\bar{C}_{21})\bar{C}^{(c_k)^{-\frac{1}{2}}} \\ &\quad +(\bar{C}^{(c_k)}{}_{11}^{-\frac{1}{2}}\bar{C}_{12}+\bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}}\bar{C}_{22})\bar{C}^{(c_k)}{}_{21}^{-\frac{1}{2}} \\ M_{21}^{c_k} &= \left(\bar{C}^{(c_k)}{}_{21}^{-\frac{1}{2}}\bar{C}_{11}+\bar{C}^{(c_k)}{}_{22}^{-\frac{1}{2}}\bar{C}_{21}\right)\bar{C}^{(c_k)}{}_{11}^{-\frac{1}{2}} \\ &\quad +(\bar{C}_{21}^{(c_k)^{-\frac{1}{2}}}\bar{C}_{12}+\bar{C}^{(c_k)}{}_{22}^{-\frac{1}{2}}\bar{C}_{22})\bar{C}^{(c_k)}{}_{21}^{-\frac{1}{2}} \end{aligned}$$

TABLE VI
AVERAGE CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON MT DATASET CHANNEL SUBSET SELECTED BASED ON LEAVE-ONE SESSION –OUT CROSS-VALIDATION OVER TRAINING SET & DIFFERENT CRITERIA, AND ACCURACIES FOR ALL CHANNELS

	M-M	AIV	M-M/VP	M-GM/V	All channels
S12	55.74	**59.98**	54.74	59.2	58.19
S13	57.49	**62.53**	62.30	58.05	52.01
S14	45.54	**59.60**	52.01	57.59	46.43
S15	53.23	57.68	**62.81**	54.34	55.23
S16	**68.67**	71	66.78	68.11	67.67
S17	**58.67**	57.78	54.67	55.11	**58.67**
S18	56.22	48	**56.89**	56.22	56.11
S19	45.11	**50.67**	51	50.44	**50.67**
S21	77.48	76.13	78.83	77.93	**79.28**
S23	54.67	**58.67**	58	55.89	57.00
S24	51.56	50.22	52.89	**53.33**	51.70
S25	**64.84**	60.83	60.83	60.83	60.83
S26	64.09	65	**66.82**	64.09	60.00
S27	60.44	**69.78**	65.33	67.11	61.78
S28	50.15	**63.88**	51.49	55.22	60.00
S29	**74.22**	72	72	72	**74.22**
Mean	58.63	**61.48**	60.46	60.34	59.36
Avg. No. of channels	14.5± 8.3	22.3±5.9	15.0625±8.7	16.0625±8.7	30

$$M_{12}^{c_k} = \left(\bar{C}^{(c_k)}{}_{11}^{-\frac{1}{2}}\bar{C}_{11} + \bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}}\bar{C}_{21}\right)\bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}} + (\bar{C}^{(c_k)}{}_{11}^{-1/2}\bar{C}_{12} + \bar{C}^{(c_k)}{}_{12}^{-1/2}\bar{C}_{22})\bar{C}^{(c_k)}{}_{22}^{-1/2}$$
$$M_{22}^{c_k} = \left(\bar{C}^{(c_k)}{}_{21}^{-\frac{1}{2}}\bar{C}_{11} + \bar{C}^{(c_k)}{}_{22}^{-\frac{1}{2}}\bar{C}_{21}\right)\bar{C}^{(c_k)}{}_{12}^{-\frac{1}{2}} + (\bar{C}^{(c_k)}{}_{21}^{-1/2}\bar{C}_{12} + \bar{C}^{(c_k)}{}_{22}^{-1/2}\bar{C}_{22})\bar{C}^{(c_k)}{}_{22}^{-1/2} \quad (10)$$

So, each M_{ij} is summation of 2^2 elements, each of them is the multiplication of 3 elements from the matrix $\bar{C}^{(c_k)}$ and $\bar{C}$ matrices. By induction, for data recorded in N channels matrix M^{c_k} would be N×N dimensional matrix each element is the summation of N^2 elements.

To estimate an upper bound for criterion we consider an upper bound for the numerator and a lower band for the denominator of the criterion as follows:

$$\sqrt{\sum_{i,j=1}^{N}\log({M^{c_k}}_{ij})^2} \leq \sqrt{N^2 max_{ij}\left(\log\left(M_{ij}^{c_k}\right)\right)^2} \quad (11)$$

For the denominator, we compute a lower band as follows:

$$\frac{\sum_{i=1}^{N_{c_k}} d_R\left(C_i^{(c_k)}, \bar{C}^{(c_k)}\right)}{N_{c_k}} \geq \frac{N_{c_k}}{N_{c_k}} \times N \times min_{ij}\left(\log\left(X_{ij}^{c_k}\right)\right),$$
$$X^{c_k} = {C_i^{(c_k)}}^{-\frac{1}{2}}\bar{C}^{(c_k)}{C_i^{(c_k)}}^{-\frac{1}{2}} \quad (12)$$

TABLE VII
AVERAGE CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON MIE DATASET CHANNEL SUBSET SELECTED BASED ON LEAVE-ONE SESSION –OUT CROSS-VALIDATION OVER TRAINING SET & DIFFERENT CRITERIA, AND ACCURACIES FOR ALL CHANNELS

	M-M	AIV	M-M/VP	M-GM/V	All channels
S1	**81.88**	84.10	84.10	84.10	83.78
S2	74.11	74.58	73.99	73.87	**74.70**
S3	61.81	**66.11**	63.13	63.84	63.84
S5	53.93	54.76	53.69	**54.52**	54.17
S8	57.93	64.60	**64.72**	64.12	61.98
S9	57.14	54.29	60	60	**57.78**
S16	74.73	71.39	73.18	73.18	**75.57**
S19	47.76	52.24	53.19	**52.24**	**52.24**
S4	50.64	44.09	46.65	47.12	**50.80**
S6	72.90	73.15	73.15	73.15	**74.65**
S10	53.65	52.70	52.06	52.06	**55.24**
S11	56.44	58.51	**58.66**	**58.66**	57.87
S14	61.28	55.37	**61.64**	**61.64**	60.92
S15	50.83	52.26	51.55	50.48	**53.21**
S17	49.92	**55.20**	52.80	46.40	54.08
S18	**51.95**	46.10	47.08	46.43	50.17
Mean	59.80	59.9656	60.60	60.11	**61.31**
Avg. No. of channels	7.3± 2.4	8.9± 1.5	7.9± 2.0	7.1± 3.0	30

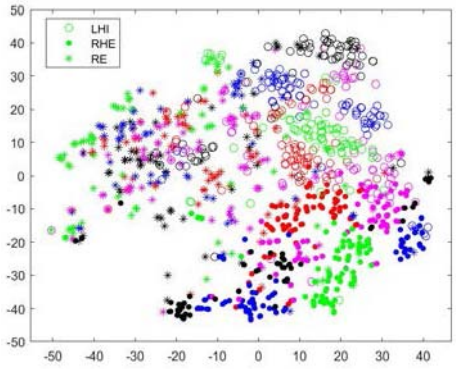
where $C_i^{(c_k)}$ is the i^{th} sample of c_k class and N_{c_k} is the number of samples in c_k class. For multi-class problem, we have the summation of N_c times $d_R\left(\bar{C}^{(c_i)}, \bar{C}\right)$ in numerator and summation of N_c times $\sigma^{(c_i)}$ in the denominator, where N_c is the number of classes. For upper bound we have:

$$(N * max_{ij}\left(\log\left(M_{ij}^{c_k}\right)\right))/(Nmin_{ij}(\log(X_{ij}^{c_k}))) = max_{ij}\left(\log\left(M_{ij}^{c_k}\right)\right))/min_{ij}(\log(X_{ij}^{c_k}))) \quad (13)$$

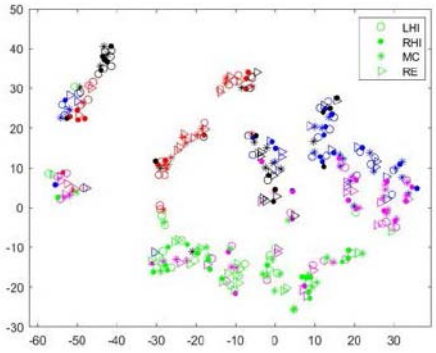
Each of $M_{ij}^{c_k}$ and $X_{ij}^{c_k}$ is from order $o(N^2)$ so in upper and lower bound on M-GM/V we have:

$$M - \frac{GM}{V} \leq \frac{log\left(N^2\right) + max_{i,j,n,l} log m_{ij} m_{jn} m_{nl}}{log\left(N^2\right) + min_{i,j,n,l} log x_{ij} x_{jn} x_{nl}} = \frac{2log\left(N\right) + max_{i,j,n,l} log m_{ij} m_{jn} m_{nl}}{2log\left(N\right) + min_{i,j,n,l} log x_{ij} x_{jn} x_{nl}}$$
$$M - \frac{GM}{V} \geq \frac{log\left(N^2\right) + min_{i,j,n,l} log m_{ij} m_{jn} m_{nl}}{log\left(N^2\right) + max_{i,j,n,l} log x_{ij} x_{jn} x_{nl}} = \frac{2log\left(N\right) + min_{i,j,n,l} log m_{ij} m_{jn} m_{nl}}{2log\left(N\right) + max_{i,j,n,l} log x_{ij} x_{jn} x_{nl}} \quad (14)$$

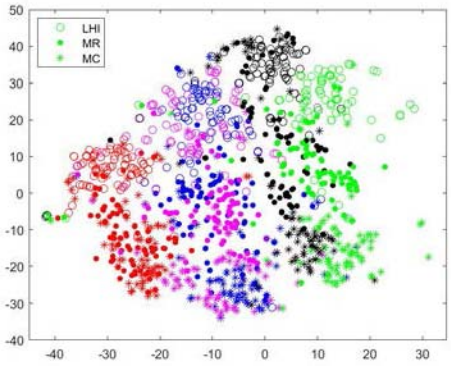
Both upper bound and lower-bounds are from order O(log(N)) in numerator and denominator. Computing the derivative of the



(a)



(b)



(c)

Fig. 6. Two-dimensional representation of a randomly selected subject of (a) MIE, (b) CYBATHLON, and (c) MT datasets after channel selection based on AIV criterion. Different colors represent different sessions (Session 1 to 5 in MT and MIE and session 14 to session 18 are in red, blue, magenta, black, green colors respectively), different shapes represent different classes. As compared to Figure 4 of the main text, these Figures show more overlap between classes after channel selection, i.e., less between-session non-stationarity.

bounds with respect to the N variable at upper bound leads to:

$$\frac{\partial}{\partial N}\left(\frac{2log\,(N)+maxlogm_{ij}m_{jn}m_{nl}}{2log\,(N)+minlogx_{ij}x_{jn}x_{nl}}\right)$$
$$=\frac{2}{N}(2log\,(N)$$
$$+minlogx_{ij}x_{jn}x_{nl}-2log\,(N)-maxlogm_{ij}m_{jn}m_{nl})$$
$$/(2log\,(N)+minlogx_{ij}x_{jn}x_{nl})^2$$
$$=\frac{2(minlogx_{ij}x_{jn}x_{nl}-maxlogm_{ij}m_{jn}m_{nl})}{N(2log\,(N)+minlogx_{ij}x_{jn}x_{nl})^2} \quad (15)$$

TABLE VIII
CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON CYBATHLON DATASET, CHANNEL SUBSET SELECTED BASED ON LEAVE-ONE SESSION – OUT CROSS-VALIDATION OVER TRAINING SET & DIFFERENT CRITERIA, AND ACCURACY FOR ALL CHANNELS

	M-M	AIV	M-M/VP	M-GM/V	All channels
Accuracy	33.74	28.22	39.36	45.4	31.90
No. of channels	5	36	15	8	46

TABLE IX
AVERAGE CLASSIFICATION ACCURACY AND NUMBER OF CHANNELS ON MT DATASET AT NEAR-OPTIMAL CHANNEL SUBSET

	M-M	AIV	M-M/VP	M-GM/V	All channels
S12	58.53	59.75	**60.98**	59.75	58.19
S13	60.29	59.73	**62.30**	58.28	52.01
S14	56.92	**57.37**	56.92	51.34	46.43
S15	57.91	**58.13**	55.46	57.91	55.23
S16	64.11	**71.00**	68.11	68.11	67.67
S17	**59.11**	58.67	58.67	58.22	58.67
S18	60.56	54.78	58.56	**60.67**	56.11
S19	**56.56**	52.22	52.33	52.33	50.67
S21	**79.28**	77.48	78.83	78.83	79.28
S23	55.00	**59.11**	58.67	57.44	57.00
S24	**53.33**	52.60	52.89	52.89	51.70
S25	**67.95**	63.50	62.17	63.20	60.83
S26	60.45	**65.91**	62.73	62.73	60.00
S27	63.11	**69.33**	65.78	67.11	61.78
S28	62.09	**66.42**	63.73	63.73	60.00
S29	73.16	72.74	71.26	72.30	**74.22**
Mean Accuracy	61.77	**62.42**	61.84	61.55	59.36
Avg. Channel No.	16.1±4.2	22.7±2.6	22.3±2.9	21.5±3.4	30

And at lower bound leads to:

$$\frac{2(maxlogx_{ij}x_{jn}x_{nl}-minlogm_{ij}m_{jn}m_{nl})}{N(2log\,(N)+maxlogx_{ij}x_{jn}x_{nl})^2} \quad (16)$$

The sign of the derivative determines the increasing and decreasing trend in bounds. Since the derivative of bounds in denominator is positive, the increasing or decreasing trend of upper or lower bound of M-GM/V is determined only by its numerator that is based on X and M matrix elements. Since over N the M and X matrix changes depends on selection criterion, therefore the trend of M-GM/V is only affected by the selection criteria, and not by changes in dimensionality.

APPENDIX B

To justify this inconsistency between performance in terms of M-GM/V and classification accuracy for M-GM/V-based

TABLE X
AVERAGE CLASSIFICATION ACCURACY AND NO. OF CHANNELS ON THE MIE DATASET AT A NEAR-OPTIMAL CHANNEL SELECTED BASED ON DIFFERENT CRITERIA AND FOR ALL CHANNELS

	M-M	AIV	M-M/VP	M-GM/V	All channels
S1	81.86	82.51	82.51	82.51	**83.78**
S2	73.75	73.87	**73.99**	73.87	74.70
S3	61.58	**65.27**	63.13	63.60	63.84
S5	53.93	**54.17**	**54.17**	52.98	54.17
S8	**65.55**	**65.55**	**65.55**	**65.55**	61.98
S9	57.14	52.70	**60.00**	**60.00**	57.78
S16	72.76	**74.26**	**74.26**	**74.26**	75.57
S19	47.76	50.64	**52.56**	52.40	52.24
S4	**48.08**	44.09	44.73	44.73	50.80
S6	**73.65**	72.02	72.02	72.02	74.65
S10	54.13	50.79	54.13	54.13	**55.24**
S11	53.74	**58.51**	**58.51**	**58.51**	57.87
S14	**61.28**	61.16	61.16	61.16	60.92
S15	52.86	**53.21**	**53.21**	**53.21**	53.21
S17	51.04	**55.2**	**55.2**	**55.2**	54.08
S18	50.65	46.10	**51.79**	**51.79**	50.17
Mean Accuracy	59.98	60.00	61.06	60.99	**61.31**
Avg. Channel No.	7.5±1.4	8.9±0.5	8.7± 0.6	8.7±0.6	11

selection, one possible reason is more discrimination for some pairs of classes and excessive overlapping for some other pairs. Fig. 5. illustrates the trend of pair-wise class discrimination averaged over all subjects. Since the most discriminative class is not the same for all subjects, at each channel subset we compute the average over the first most discriminative pair, second-most discriminative pair, and least discriminative pair for each subject of MT dataset. Higher M-GM/V for M-GM/V-based selection in small channel set size as compared with M-M/VP–based selection coincides with higher discrimination in two more discriminative pairs and more overlapping in less discriminative pairs (Fig. 5. (b)).

APPENDIX C

See Tables VI–VIII.

APPENDIX D

See Fig. 6.

APPENDIX E

Tables IX and X provides the individual subject performances, with the different criteria and our proposed channel subset and number selection approaches, for the MT and MIE data sets. These tables are the detailed versions (with subject-specific results) of the average performances Tables II and III in the main body.

ACKNOWLEDGMENT

The authors thank Fatemeh Alimardali, Aline Roc, Léa Pillette, and Thibaut Monseigne for reading and commenting the manuscript; Léa Pillette and Camille Jeunet for sharing their EEG data and Fatemeh Alimardani for useful discussion.

REFERENCES

[1] M. Clerc, L. Bougrain, and F. Lotte, *Brain-Computer Interfaces 1: Foundations and Methods*. Hoboken, NJ, USA: Wiley ISTE, 2016.
[2] M. Congedo *et al.*, "'Brain invaders': A prototype of an open-source P300-based video game working with the OpenViBE platform," in *Proc. 5th Int. Brain-Comput. Interface Conf. (BCI)*, 2011, pp. 280–283.
[3] S. Haufe, M. S. Treder, M. F. Gugler, M. Sagebaum, G. Curio, and B. Blankertz, "EEG potentials predict upcoming emergency brakings during simulated driving," *J. Neural Eng.*, vol. 8, no. 5, Jul. 2011, Art. no. 056001.
[4] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. Oxford, U.K.: OUP USA, 2012.
[5] 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 Process.*, vol. 85, no. 11, pp. 2190–2212, 2005.
[6] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, Jun. 2011.
[7] S. Saha and M. Baumert, "Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: A review," *Frontiers Comput. Neurosci.*, vol. 13, p. 87, Jan. 2020.
[8] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 031005.
[9] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: A review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1753–1762, Oct. 2017.
[10] M. S. Yamamoto, K. Sadatnejad, T. Tanaka, R. Islam, Y. Tanaka, and F. Lotte, "Detecting EEG outliers for BCI on the Riemannian manifold using spectral clustering," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 438–441.
[11] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.
[12] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2390–2401, Aug. 2019.
[13] S. Perdikis, R. Leeb, and J. D. R. Millán, "Context-aware adaptive spelling in motor imagery BCI," *J. Neural Eng.*, vol. 13, no. 3, May 2016, Art. no. 036018.
[14] S. Kumar, F. Yger, and F. Lotte, "Towards adaptive classification using Riemannian geometry approaches in brain-computer interfaces," in *Proc. 7th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2019, pp. 1–6.
[15] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
[16] B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K.-R. Müller, "The Berlin brain-computer interface: Accurate performance from first-session in BCI-naive subjects," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp. 2452–2462, Oct. 2008.
[17] F. Lotte and C. Jeunet, "Defining and quantifying users' mental imagery-based BCI skills: A first step," *J. Neural Eng.*, vol. 15, no. 4, 2018, Art. no. 46030.
[18] A. Barachant and S. Bonnet, "Channel selection procedure using Riemannian distance for BCI applications," in *Proc. 5th Int. IEEE/EMBS Conf. Neural Eng.*, Apr. 2011, pp. 348–351.
[19] A. Abu-Rmileh, E. Zakkay, L. Shmuelof, and O. Shriki, "Co-adaptive training improves efficacy of a multi-day EEG-based motor imagery BCI training," *Frontiers Hum. Neurosci.*, vol. 13, p. 362, Oct. 2019.
[20] G. John and R. Kohavi, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 272–324, 1997.

[21] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection. Pattern recognition letters," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[22] K. Sadatnejad, A. Roc, L. Pillette, A. Appriou, T. Monseigne, and F. Lotte, "Channel selection over Riemannian manifold with non-stationarity consideration for brain-computer interface applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1364–1368.

[23] L. Pillette, B. N'Kaoua, R. Sabau, B. Glize, and F. Lotte, "Multi-session influence of two modalities of feedback and their order of presentation on MI-BCI user training," *Multimodal Technol. Interact.*, vol. 5, no. 3, p. 12, Mar. 2021.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[25] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, pp. 629–636, 2010.

[26] V. Jayaram and A. Barachant, "MOABB: Trustworthy algorithm benchmarking for BCIs," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066011.

[27] S. Perdikis, L. Tonin, S. Saeedi, C. Schneider, and J. D. R. Millán, "The cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users," *PLOS Biol.*, vol. 16, no. 5, May 2018, Art. no. e2003787.

[28] R. Scherer, J. Faller, P. Sajda, and C. Vidaurre, "EEG-based endogenous online co-adaptive brain-computer interfaces: Strategy for success?" in *Proc. 10th Comput. Sci. Electron. Eng. (CEEC)*, Sep. 2018, pp. 299–304.

[29] Z. Qiu, J. Jin, H.-K. Lam, Y. Zhang, X. Wang, and A. Cichocki, "Improved SFFS method for channel selection in motor imagery based BCI," *Neurocomputing*, vol. 207, pp. 519–527, Sep. 2016.

[30] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Robust EEG channel selection across sessions in brain-computer interface involving stroke patients," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–6.

[31] M. Schröder *et al.*, "Robust EEG channel selection across subjects for brain-computer interfaces," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 19, pp. 1–10, Dec. 2005.

[32] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, Feb. 2004.

[33] W.-K. Tam, Z. Ke, and K.-Y. Tong, "Performance of common spatial pattern under a smaller set of EEG electrodes in brain-computer interface on chronic stroke patients: A multi-session dataset study," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 6344–6347.

[34] C. Jeunet, B. N'Kaoua, S. Subramanian, M. Hachet, and F. Lotte, "Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0143962.

[35] C. Benaroch *et al.*, "Long-term BCI training of a tetraplegic user: Adaptive Riemannian classifiers and user training," *Frontiers Human Neurosci.*, vol. 15, p. 118, Mar. 2021.

[36] A. Appriou, A. Cichocki, and F. Lotte, "Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals," *IEEE Syst., Man Cybern. Mag.*, vol. 6, no. 3, pp. 29–38, Jul. 2020.

[37] D. Freer and G. Z. Yang, "Data augmentation for self-paced motor imagery classification with C-LSTM," *J. Neural Eng.*, vol. 17, no. 1, 2020, Art. no. 016041.

[38] A. Chowdhury and J. Andreu-Perez, "Clinical brain–computer interface challenge 2020 (CBCIC at WCCI2020): Overview, methods and results," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 3, pp. 661–670, Aug. 2021.

[39] *3rd Neuroergonomics Conference Grand Challenge: Passive BCI Hackthon*. Accessed: Sep. 30, 2021. [Online]. Available: https://www.neuroergonomics conference.um.ifi.lmu.de/pbci/

[40] A. Barachant and M. Congedo, "A Plug&Play P300 BCI using information geometry," 2014, *arXiv:1409.0107*.

[41] E. K. Kalunga, S. Chevallier, Q. Barthélemy, K. Djouani, E. Monacelli, and Y. Hamam, "Online SSVEP-based BCI using Riemannian geometry," *Neurocomputing*, vol. 191, pp. 55–68, May 2016.

[42] *BCI Challenge @ NER 2015*. Accessed: Dec. 30, 2022. [Online]. Available: https://www.kaggle.com/c/inria-bci-challenge/