Modeling complex EEG data distribution on the Riemannian manifold toward outlier detection and multimodal classification

Maria Sayu Yamamoto, Khadijeh Sadatnejad, Toshihisa Tanaka, *Senior Member, IEEE*, Md. Rabiul Islam, *Member, IEEE*, Frédéric Dehais, Yuichi Tanaka, *Senior Member, IEEE*, and Fabien Lotte

***Abstract*— *Objective*: The usage of Riemannian geometry for Brain-computer interfaces (BCIs) has gained momentum in recent years. Most of the machine learning techniques proposed for Riemannian BCIs consider the data distribution on a manifold to be unimodal. However, the distribution is likely to be multimodal rather than unimodal since high-data variability is a crucial limitation of electroencephalography (EEG). In this paper, we propose a novel data modeling method for considering complex data distributions on a Riemannian manifold of EEG covariance matrices, aiming to improve BCI reliability. *Methods*: Our method, *Riemannian spectral clustering* (RiSC), represents EEG covariance matrix distribution on a manifold using a graph with proposed similarity measurement based on geodesic distances, then clusters the graph nodes through spectral clustering. This allows flexibility to model both a unimodal and a multimodal distribution on a manifold. RiSC can be used as a basis to design an outlier detector named *outlier detection Riemannian spectral clustering* (odenRiSC) and a multimodal classifier named *multimodal classifier Riemannian spectral clustering* (mcRiSC). All required parameters of odenRiSC/mcRiSC are selected in data-driven manner. Moreover, there is no need to pre-set a threshold for outlier detection and the number of modes for multimodal classification. *Results*: The experimental evaluation revealed odenRiSC can detect EEG outliers more accurately than existing methods and mcRiSC outperformed the standard unimodal classifier, especially on high-variability datasets. *Conclusion*: odenRiSC/mcRiSC are anticipated to contribute to making real-life BCIs outside labs and neuroergonomics applications more robust. *Significance*: RiSC can work as a robust EEG outlier detector and multimodal classifier.**

***Index Terms*— Brain-computer interface (BCI), electroencephalography (EEG), multimodal distributions, outlier detection, Riemannian geometry, spectral clustering**

I. Introduction

Brain-Computer Interfaces (BCIs) offer a direct communication pathway between a user and a machine without requiring any muscular engagement [1], [2]. Since the first BCI system was proposed [3], BCI technology has diversified over the past several decades and shown potential to revolutionize numerous applications. For instance, it is promising as an assistive device for motor-impaired users' life [4], [5], a new control device for gaming for the general public [6], [7] and a detector for intraoperative awareness of general anesthesia [8]. Furthermore, the technology used to design BCIs can significantly benefit in neuroergonomics [9], [10], which is the emerging science of how human brain works in everyday life, such as a real-time monitoring of pilot's mental states [11].

To translate brain activity into meaningful computer commands, first, brain activities are recorded typically by using electroencephalography (EEG) [1]. Then, the user's intent is identified according to a pattern recognition pipeline, *i.e.,* extracting features from recorded EEG and classifying them using machine-learning approaches. Recently, the use of Riemannian geometry has gained prominence in the BCI pattern recognition pipeline, referred to as Riemannian BCIs [12]–[14]. In this pipeline, features are extracted by describing EEG signals as covariance matrices, and classified by considering properties of the space where the covariance matrices belong, called Riemannian manifold. Riemannian BCIs have shown significant improvements for many subjects [15] and have won multiple EEG classification competitions [13], [14], [16].

Despite their many potential applications and the efficacy of the use of Riemannian geometry, most BCI studies are still confined to laboratories and generally report about 20% of their participants failing to operate BCIs, a phenomenon known as BCI illiteracy/deficiency [17], [18]. There are still many issues that the BCI research community needs to overcome for making BCIs practically useful to many users in real-life settings. Among them, one can cite the issue of large intra-user data variability [19]. This variability problem can be caused by a variety of factors. In general, EEG recordings are very sensitive to artifacts and label noises. Artifacts are often caused by the user's muscular activity [20], while label noises are caused when the user performs a different task than the given instruction [21]. Also, EEG are sensitive to changes in users' cognitive states such as fatigue and vigilance [22]. These factors can cause outliers in a dataset, and/or make a data distribution more complex, and possibly multimodal. As a result, this may impede a successful pattern recognition because common Riemannian methods are based on unimodal distribution modeling. Moreover, the variability issue can be even more crucial when EEG are recorded in noisy environments in the real world.

As such, even though data variability remains a serious issue, there has been less progress in Riemannian BCIs for developing systems that can handle high data variability. Therefore, we have suggested the need for adapting a traditional unimodal Riemannian BCI classifier, *i.e.*, Minimum Distance to Riemannian Mean (MDRM) [23] to multimodal data distributions in [12]. To the best of our knowledge, the present paper is the first work to tackle this open research challenge. In this study, we propose a modeling method of complex data distributions on a Riemannian manifold toward improving BCI reliability. Our method, named Riemannian spectral clustering (RiSC), captures a geometrical data distribution on a manifold via a similarity graph and then performs spectral clustering on this graph. This allows us to represent an EEG covariance matrices distribution on a manifold as multimodal when the distribution has multiple modes. This modeling can be used as a basis to design

This work was supported by the European Research Council with project BrainConquest (grant ERC-2016-STG-714567), JSPS KAKENHI Grant Number 17H01760, and the French National Research Agency (ANR) with the program UDOPIA (grant ANR-20-THIA-0013-01).

M.S. Yamamoto is with LISN, Univ. Paris-Saclay, Gif-sur-Yvette, France (e-mail: maria-sayu.yamamoto@universite-paris-saclay.fr).

K. Sadatnejad and F. Lotte are with Inria Center at Univ. Bordeaux / LaBRI, Talence, France (e-mail: sadatnejad.kh@gmail.com, fabien.lotte@inria.fr).

T. Tanaka is with Tokyo University of Agriculture and Technology, Koganei–shi, Tokyo, Japan (e-mail: tanakat@cc.tuat.ac.jp).

M.R. Islam is with Center for Precision Medicine, University of Texas Health, San Antonio, USA (e-mail: islamr@uthscsa.edu).

F. Dehais is with ISAE-SUPAERO, Univ. Fédérale Toulouse, France, Artificial and Natural Intelligence Toulouse Institute, Univ. Toulouse, France, School of Biomedical Engineering, Science Health Systems, Drexel Univ., Philadelphia, PA, USA (e-mail: frederic.dehais@isae-supaero.fr).

Y. Tanaka is with Osaka University, Suita, Osaka, 565-0871, Japan (e-mail: ytanaka@comm.eng.osaka-u.ac.jp).

an outlier detector named outlier detection Riemannian spectral clustering (odenRiSC) or a multimodal classifier named multimodal classifier Riemannian spectral clustering (mcRiSC).

OdenRiSC allows detecting outliers without the need for a reference matrix and a threshold, which are traditionally required in existing outlier detection methods for Riemannian BCIs [24], [25]. It should be mentioned that our preliminary results of odenRiSC have been presented in a short conference paper [26] and confirmed that odenRiSC could detect outliers significantly better than existing methods on the real EEG data with artificial outliers. In the present paper, we report further studies with this approach on real EEG datasets including actual outliers.

McRiSC removes the need for prior mode number setting, which is required for a geometrical probability-based multimodal classifier [27]. The performance of mcRiSC was evaluated using a dataset recorded inside a lab, in which data variability is expected to be low, and with another dataset collected in actual flight condition, which is expected to be noisy and with large data variability.

This paper is organized as follows: Section II presents the basis of Riemannian geometry. Then, Section III describes existing methods for outlier detection and mutimodal classification. Section IV describes the theoretical background and implementation of our method. The data sets used for numerical experiments and the evaluation way of the proposed method are described in Section V and VI. Then, Section VII describes the results while Section VIII discusses them. We also provide a guideline about how to apply our proposed methods for practitioners in Section IX and conclude this paper in Section X.

II. Riemannian geometry for EEG covariance matrices

In this section, we introduce the basic principles of Riemannian geometry involved in EEG covariance matrices analysis.

A. EEG covariance matrix

Let $X \in \mathbb{R}^{M \times T}$ be a band-pass filtered EEG signal with M channels and T time samples. The covariance matrix of X is defined as: $P_X = \frac{1}{T-1} X X^\top$. This estimated covariance matrix is known to be empirically Symmetric Positive-Definite (SPD). Its diagonal entries represent the variance of each channel, *i.e.,* the band power of each channel, while the off-diagonal terms contain the covariance of each channel pair.

B. Geodesic distance on Riemannian manifold

The $M \times M$ SPD matrix belongs to a differentiable manifold $\mathcal{P}(M)$, called Riemannian manifold, which is a smooth curved space and equipped with a Riemannian metric [28]. The Affine Invariant Riemannian Metric (AIRM) is a geodesic measurement that respects the original curvature of a manifold [29]. The AIRM distance δ_r between two SPD matrices P_1 and P_2 on manifold is defined as:

$$\delta_r(P_1, P_2) = \|\log({P_1}^{-\frac{1}{2}} P_2 {P_1}^{-\frac{1}{2}})\|_F = \left(\sum_{i=1}^{n} \log^2 \lambda_i\right)^{1/2} \quad (1)$$

where λ_i are positive eigenvalues of ${P_1}^{-\frac{1}{2}} P_2 {P_1}^{-\frac{1}{2}}$ and $\|.\|_F$ is the Frobenius norm. We used the AIRM distance for similarity measures and classification.

C. Riemannian basic classification: Minimum Distance to Riemannian Mean (MDRM)

The most basic Riemannian classifier relies on the intra-class means of SPD matrices, referred to as Minimum Distance to Riemannian Mean (MDRM) [23].

Given n SPD matrices $\{P_1, P_2, \ldots, P_n\}$ for class c, the Riemannian mean $\overline{P}_c$ on a manifold $\mathcal{P}(M)$ is defined as:

$$\overline{P}_c \left(P_1, P_2, \ldots, P_n\right) := \underset{P \in \mathcal{P}(M)}{\arg\min} \sum_{k=1}^{n} \delta_r^2(P, P_k). \quad (2)$$

where $\overline{P}_c$ is typically estimated through an optimization algorithm using training data [30]. Then, a new observation is classified according to the shortest AIRM distance to each class mean $\overline{P}_c$, *i.e.,* $\hat{c} = \arg\min_c \delta_r \left(P, \overline{P}_c\right)$.

III. Related work

A. Outlier detection in Riemannian BCIs

1) Riemannian Potato: This is the first outlier detection method for Riemannian BCIs, which was originally proposed for online EEG signal quality monitoring [24]. The detection algorithm requires two parameters: A reference matrix and a threshold. Samples whose distances from a reference matrix are larger than the threshold are rejected as outliers. The reference matrix is the Riemannian mean of all samples. The threshold th to reject samples is estimated as $th = \mu + 2.5\sigma$, where μ and σ are the mean and the standard deviation respectively of the Riemannian distances between each EEG covariance matrix and the reference matrix. In the experimental evaluation with P300-based BCI, Riemannian Potato detected artifacts of different origins, such as eye movements or electrodes movements.

2) Median-based Trimmed Averages: This is one of the geometric trimmed average techniques that were proposed to improve tangent point estimation for Tangent Space Mapping (TSM) based Riemannian BCIs [25]. The tangent point for mapping was conventionally set by using a plain geometric average derived from all samples. However, this way of estimation is strongly affected by outliers. Thus, to estimate a more robust tangent point, this approach eliminated as outliers the d % (a user-specified threshold) of samples that exhibited the largest Riemannian distance to the geometric median of all samples. This showed higher classification accuracies than the plain geometric average in TSM-based Riemannian BCIs.

However, both methods suffer from one main limitation. They need a reference matrix to characterize the distribution of EEG covariance matrices on a Riemannian manifold and a threshold to detect outliers. Thus, an inappropriate value of such parameters may decrease their outlier detection performance. In other words, there is a risk that some true outliers may not be detected as outliers or that some non-outlier samples may be erroneously rejected as outliers. In contrast, our method, odenRiSC, is free from this limitation as it describes data distributions using a graph and detects outliers by clustering nodes of this graph.

B. Multimodal Riemannian EEG classification

Gaussian distributions and mixtures of Gaussian distributions were generalized from an Euclidean space to a Riemannian manifold in [31]. The probability density function of the Riemannian Gaussian distribution can be written as:

$$f(P \mid \bar{P}, \sigma) = \frac{1}{\zeta(\sigma)} \exp\left(-\frac{\delta_r^2(P, \bar{P})}{\sigma^2}\right) \quad (3)$$

where $\zeta(\sigma)$ is a normalization function with $\bar{P} \in \mathcal{P}(M)$ and $\sigma > 0$. Note that the maximum likelihood estimator of $\bar{P}$ coincides with the center of mass $\overline{P}_c$. This density function can be extended to mixtures of Riemannian Gaussian distribution by: $g(P) = \sum_{h=1}^{H} w_h f\left(P \mid \bar{P}_h, \sigma_h\right)$ where H is the number of mixture components, $w_1, \ldots, w_H$ are positive weights summing to 1. The two parameters $\bar{P}_h \in \mathcal{P}(M)$ and $\sigma_h > 0$ are estimated by the expectation-maximization algorithm.

Zanini et al. [27] applied mixtures of Riemannian Gaussian distribution to build a probabilistic Bayesian classifier, for EEG classification to consider the shape of covariance matrix distributions on a manifold. The classification accuracy of the best Bayesian classifier was at least as accurate as the MDRM for all subjects. Interestingly, the results also revealed the optimal number of modes was different among subjects. This implies that, even if the BCI experiment is common for every subject, the shapes and numbers of data clouds are highly variable between individuals.

However, as a limitation, this method required to set the number of mixture components in advance manually, even though it is difficult to know how many modes there are in a given dataset in practice. In contrast, mcRiSC, our proposed multimodal classification model, does not need to handcraft the number of modes as it detects modes in a data-driven manner using spectral clustering.

IV. Methodology

Our method models SPD matrices on a manifold as a set of clusters through spectral clustering. Then, these clusters are utilized for outlier detection and multimodal classification. Spectral clustering, a core tool of our method, is a data clustering method relying on a graph [32]. The relationships between data samples is represented via a similarity graph. Then, the samples are grouped through calculation of eigenvalues of the graph, *i.e,* the spectrum of the graph Laplacian.

In this section, we introduce the way to construct the graph, the two different spectral clustering methods: unnormalized and normalized spectral clustering from the graph partitioning problem point of view, and finally, the detail of our data modeling method.

A. Graphs

Given a set of data points $x_1, x_2, \ldots, x_n$, an undirected graph $\mathcal{G}$ is represented by two elements $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ denotes a set of nodes representing data points and $\mathcal{E}$ denotes a set of edges connecting nodes. If two data points x_i and x_j have some similarities, their nodes are connected by an edge with a non-negative weight w_{ij}. The connectivity of each node is represented by the adjacency matrix W whose entries are as follows:

$$W(i,j) = \begin{cases} w_{ij} & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Note that we constrain $\mathcal{G}$ to be undirected, *i.e.*, $w_{ij} = w_{ji}$.

There are several popular methods to construct a similarity graph. In this paper, we use two different methods: A fully connected graph and a k-nearest neighbor (k-nn) graph. With a fully connected graph, all pairs of nodes are connected with their similarity weights. With the k-nn graph, the goal is to connect node v_i with node v_j if v_j is in the k-nn of v_i. However, this definition leads to a directed graph, as the neighborhood relationship is not symmetric. To satisfy the undirected constraint, we ignore the directions of the edges, so that v_i and v_j are connected with an undirected edge if v_i is among the k-nn of v_j or if v_j is among the k-nn of v_i.

Algebraically, a graph topology is formulated through a graph Laplacian. There are several types of graph Laplacians. In this paper, we only use an unnormalized graph Laplacian L [33], which are defined as $L = D - W$ where D is a diagonal matrix called degree matrix whose diagonal elements are given as $d_i = \sum_{j=1}^{n} w_{ij}$. L is symmetric positive semi-definite, *i.e.,* all sorted eigenvalues are non-negative and real-valued $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

B. Graph cut theory for spectral clustering

Spectral clustering algorithms are based on the solution of graph cut problems. Its goal is to produce well-separated clusters, represented by h connected subgraphs. The two most common objective functions for it are RatioCut [34] and Ncut [35]. Both functions try to include sufficient samples in each cluster but it is achieved based on different measures. RatioCut [34] solves the following optimization problem to find the optimal h connected subgraphs $A_1, A_2 \ldots, A_h$:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\sum_{j=1}^{h} \frac{W(A_j, \bar{A}_j)}{|A_j|} \\ &\text{subject to} && A_1, A_2 \ldots, A_h \end{aligned} \tag{5}$$

where $\bar{A}_j$ is the complement of A_j, $W\left(A_j, \bar{A}_j\right) = \sum_{i \in A_j, l \in \bar{A}_j} w_{ij}$ with the shorthand notation i and l for the set of nodes $\{i \mid v_i \in A_j\}$ and $\{l \mid v_l \in \bar{A}_j\}$, and the factor $\frac{1}{2}$ is for avoiding counting each edge twice in the cut. $|A|$ denotes the number of nodes to measure the size of subset A in $\mathcal{G}$. RatioCut tries to find the optimal clusters by balancing the size of clusters according to their number of nodes.

Meanwhile, Ncut [35] defines the optimization problem as follows:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\sum_{j=1}^{h} \frac{W(A_j, \bar{A}_j)}{vol(A_j)} \\ &\text{subject to} && A_1, A_2 \ldots, A_h \end{aligned} \tag{6}$$

where $vol(A_j)$ is a sum of weights of all edges attached to nodes in subset A_j. Ncut tries to find optimal clusters by balancing the cluster according to the weight of edges.

Note that Eq. 5 and 6 are NP-hard combinatorial optimization problems. They aim to find the optimal combination of sets $A_1, A_2, \ldots, A_h$ by assigning a discrete value $\{-1, 1\}$, which indicates a hard cluster membership, to each graph node. To make the problems tractable, they are relaxed to be continuous real-valued optimization problems: finding a real value between -1 and 1, *i.e.*, $[-1, 1]$ - *i.e.*, a soft cluster membership. The optimal solution of this continuous optimization is a matrix whose columns are h eigenvectors of L for relaxed RatioCut and h generalized eigenvectors of L for relaxed Ncut based on the Rayleigh-Ritz theorem [36, Section 5.5.2].

Finally, this real-valued solution is converted to a discrete partition by applying k-means algorithms on rows of this optimal solution. These methodology is so-called spectral clustering. Relaxing RatioCut is solved by unnormalized spectral clustering and relaxing Ncut by normalized spectral clustering, respectively.

In this paper, we used these two different spectral clustering methods based on an unnormalized graph Laplacian [35], which are summarized in Algorithm 1. The difference between these methods is only the eigenvalue decomposition method for L.

Algorithm 1 General spectral clustering algorithm

Input: A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Output: Graph partitions $C_1, \ldots, C_h$
1: Compute an unnormalized Laplacian matrix L
2: **For unnormalized spectral clustering**:
Compute the first h eigenvectors $u_1, u_2, ..., u_h$ of L
For normalized spectral clustering:
Compute the first h generalized eigenvectors $u_1, u_2, ..., u_h$ of L
3: Let $U \in \mathbb{R}^{n \times h}$ be the matrix containing h vectors $u_1, \ldots, u_h$ as columns and let $y_i \in \mathbb{R}^h$ be a vector corresponding to the i^{th} row of U for $i = 1, \ldots, n$.
4: Cluster $(y_i)_{i=1,\ldots,n}$ into partitions $C_1, \ldots, C_h$ applying the k-means algorithm in the feature space $\mathbb{R}^h$

C. RiSC: Riemannian spectral clustering

The geometrical relations between n EEG covariance matrices of class c on a manifold are described by a similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{P_1, P_2, \ldots, P_n\}$ is a set of nodes, here EEG covariance

matrices, and $\mathcal{E}$ is a set of edges connecting nodes in $\mathcal{V}$. The similarity graph is built for each class by either a fully connected graph or a k-nn graph. To construct the k-nn graph, we select $k = \log(n)$, as recommended in [32]. The edges are weighted by Gaussian similarity w_{ij} based on a pairwise AIRM distance between P_i and P_j:

$$w_{ij} = \begin{cases} \exp\left(-\frac{\delta_r^2(P_i,P_j)}{2q^2}\right) & i \neq j, \\ 0 & i = j, \end{cases} \tag{7}$$

where $q \in [0, 1]$ controls the similarity between EEG covariance matrices. In general, spectral clustering can be quite sensitive to q values since the similarity topology can be changed according to this parameter [32]. For instance, a too small q may lead to too small similarity weights and w_{ij} would be very close to 0. As a result, all nodes of the similarity graph would appear equally far away. To set parameter q in data-driven manner, another graph $\mathcal{G}_q$ is built, which has the same graph topology as $\mathcal{G}$ with the pairwise AIRM distance $\delta_r(P_i, P_j)$ as edge weight. Then, the median of the edge lengths of the minimum spanning tree (MST) of $\mathcal{G}_q$ was selected as q. This MST has the same n nodes as $\mathcal{G}_q$ and all nodes are connected by a route that minimizes the total edge weights without any cycle. We assume the median of their length to be a reasonable radius of adjacency among different distributions. Thus, this q magnifies the similarity within a cluster and weaken it between clusters.

Another important point for spectral clustering is the choice of the number of clusters h. In this study, the optimal h is selected according to the eigengap heuristic [32]. The eigengap heuristic is based on the spectral decomposition of the Laplacian of graph $\mathcal{G}$. The index of the i^{th} eigenvalue λ_i which has the maximum gap is set as suitable h, *i.e.,* $h = \arg\max_i(\lambda_{i+1} - \lambda_i)$. After determining h (at maximum $h = 5$), the spectral clustering algorithm is applied on L as Algorithm 1. Depending on how the resulting clusters are used, RiSC can act as an outlier detector named odenRiSC, or as a multimodal classifier named mcRiSC. The RiSC algorithms (odenRiSC and mcRiSC) were implemented with Matlab R2022b and code are available at https://github.com/msyamamoto/RiSC.

1) odenRiSC – outlier detection Riemannian spectral clustering: odenRiSC is an outlier detector working on a Riemannian manifold of EEG covariance matrices. As illustrated in Fig. 1, all resulting clusters of spectral clustering except the cluster with the largest size are identified as outlier clusters. Thus, if there is no outlier in a given dataset, a single cluster should be detected, and no data will be rejected as outliers. This is in contrast to other existing methods, which always reject data, even when there is no outlier.

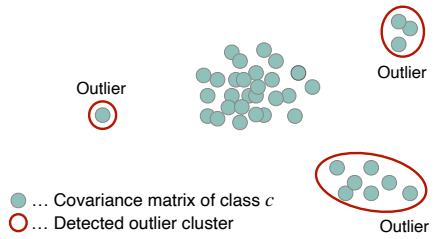

Fig. 1. Illustration of odenRiSC idea

2) mcRiSC – multimodal classification Riemannian spectral clustering: mcRiSC allows MDRM to be multimodal thanks to RiSC. To train a classifier, first, the data are clustered through RiSC per each class. Then, the clusters with only one node are removed as outlier cluster. Next, the cluster centroids of the remaining B clusters for all classes combined $\{\overline{P}^{(1)}, \cdots, \overline{P}^{(B)}\}$ are estimated using the Riemannian mean. Then, for applying the classifier to test set, the AIRM distances from a new observation P to those cluster centroids are computed. As illustrated in Fig. 2, the prediction $\hat{b}$ is provided according to the class to which the nearest cluster belongs, *i.e.,*

$$\hat{b} = \arg \min_{b \in \{1,\ldots,B\}} \delta_r \left(P, \overline{P}^{(b)}\right) \tag{8}$$

The distinctive point compared to the geometrical probability-based classifier based on mixtures of Riemannian Gaussian distribution, which needs manually selection of the number of modes, is that RiSC can select the number of modes in a fully data-driven way by using spectral clustering. Another advantage of our automatic mode number selection is that if the data does not have well-separated modes, RiSC can theoretically detects a single big cluster, and mcRiSC will thus act as a unimodal classifier. Thus, mcRiSC could be used flexibly without prior knowledge on whether the data has a unimodal or multimodal distribution. It is also important to note that even though the current paper focuses on MDRM, it is possible to combine mcRiSC with other classifiers such as k-nn by assigning the majority class among the k-nearest cluster centroids.

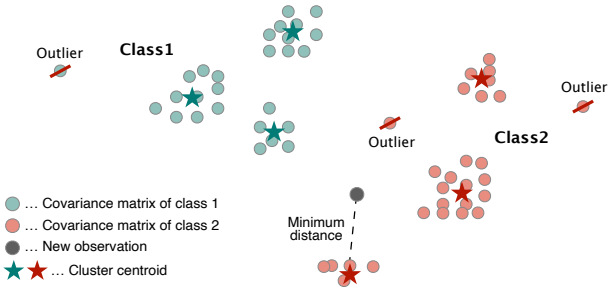

Fig. 2. Illustration of mcRiSC idea

V. Experiments for Outlier Detection

A. Data description for outlier detection experiment

To evaluate RiSC, we conducted two numerical experiments: with EEG dataset with artificial outliers or with actual motion outliers. The dataset with artificial outliers were generated based on real EEG data and concatenated to the real EEG dataset. This enabled us to have a ground truth to assess objectively how well each method detects outliers. The concatenated original EEG data may still contain some outliers, which we assume here are mostly not substantial because the data was recorded in careful laboratory conditions. For the experiment with actual motion outliers, we used a public dataset whose EEG epochs were contaminated by motion artifacts. Those artifacts locations were labeled. This provided us with a ground truth to assess outlier detection performance.

1) Real EEG data for artificial outlier generation: In order to create contaminated dataset with artificial outliers, we used dataset IIa from BCI competition IV, provided by TU Graz, Austria [37]. This set comprises EEG signals from nine subjects who performed left hand, right hand, both feet, and tongue Motor Imagery (MI). EEG signals were recorded using 22 EEG channels. The presence of ocular artifacts, *i.e.*, eye movement artifacts was marked. Usually, eye movement affect frequency bands lower than 7 Hz [20]. Here, EEG signals were band-pass filtered in the 7–30 Hz, using a 4^{th} order Butterworth filter, thus, ocular artifacts should not affect the resulting EEG covariance matrices much. Training and testing sets are available for each subject. Both sets contain 72 trials for each class. Each trial lasts for 7 sec and subjects performed MI within $t = 3$ to 6 sec. In this work, we only used training data of left hand MI from subjects *A01*, *A02*, *A03*, and *A07* because these datasets

contain the fewest ocular artifacts. For computing EEG covariance matrices, EEG signals from the whole MI interval was used.

2) Contamination scenario of artificial outliers: Contaminated datasets were generated by adding artificial artifacts to reference EEG trials by following the reference paper about probability modeling for time series additive outliers [38]. In actual EEG, outliers are often generated due to artifacts affecting specific channels and time periods. For instance, a common artifact type is facial muscle artifacts (*e.g.*, from frowning), which affect frontal channels EEG [20]. Thus, once the reference EEG trials were randomly selected from the real dataset, contaminated EEG, *i.e.* outliers were generated, by adding artifacts to frontal channels (Fz, FC3, FC1, FCz, FC2, & FC4), for a fixed time interval, to the reference EEG trials. Based on these trials, z outlier trials $Y \in \mathbb{R}^{M\times N}$ were generated. The contaminated EEG signal $Y_t^{(m)}$ for channel m and time instance t is formulated as $Y_t^{(m)} = x_t^{(m)} + \gamma_t v_t$, where $x_t^{(m)}$ is a reference trial signal and $\gamma_t \in \{0, 1\}$ controls the proportion of time points to which we add artifacts. The simulated artifact v_t is added (*i.e.* $\gamma_t = 1$) with probabilities ε. v_t is drown from a multivariate normal distribution $v_t \sim N(\mu, 2\sigma^2 I)$ in which $N(\cdot)$ is a normal distribution with mean μ and variance σ^2 of a reference channel selected randomly from each reference EEG trial. In this study, three different outlier numbers $z = 5, 10, 25$ and outlier strength $\varepsilon = 0.10,\ 0.30,\ 0.50$ were set. For each generating condition, 30 datasets were randomly generated, for a total of 270 datasets per subject.

3) Real EEG datasets with motion artifact labels: To evaluate whether outlier removal does really remove true outliers and only outliers, each trial should be ideally labeled as outlier/non-outlier. For this, we used the motion artifact contaminated EEG public dataset by Sweeney et al. [39]. This dataset was collected to serve as a benchmark for artifact removal techniques. EEG signals were recorded from six subjects who did not perform any activity and kept their eyes closed to limit the number of artifacts. In addition, users' head position was maintained stationary throughout the experiment to avoid head motion. EEG signals were recorded from two frontal channels (FPz and FP1h using 10–5 system). There are 23 trials totally, each lasting 9 min, with motion-induced artefacts to one of the electrodes at regular 2 min intervals for 1 min long. This artifact was induced by mechanically pulling the connecting lead of the electrode.

In our experiment, EEG signals were band-pass filtered in the 0.5–30 Hz, using a 4^{th} order Butterworth filter. The first 59 sec of each alternating clean and contaminated time period were epoched, and 4 epochs were extracted for each time period in one initial trial. Thus, a total of 92 trials were created for clean data and 92 trials for outlier data, from 23 initial trials. Then, 3 different quantities of outliers (5, 10, and 25) were randomly selected from the created outlier dataset. For each outlier number condition, 30 datasets were generated.

B. Evaluation for outlier detection

The performance of odenRiSC was compared with *Riemannian Potato* (RP) and *Median-Based Trimming* (MBT) that were introduced in Section III-A. Regarding odenRiSC, we used 2 different models: odenRiSC with a fully connected graph with either unnormalized spectral clustering (unnormalized odenRiSC) or normalized spectral clustering (normalized odenRiSC). For MBT, we used 95% trimming as a common threshold for statistical outlier detection.

The performance of each method was evaluated by Hit-False Difference (HFD) [40]. The HFD is a single metric to quantify the performance of a detection algorithm, which is calculated by subtracting the False Positive Rate (FPR) from the True Positive Rate (TPR). Here, the FPR is the percentage of non-outliers that were detected as outliers whereas the TPR is the percentage of actual outliers that were detected as outliers. In this sense, a larger HFD means the model detects true outliers more accurately and precisely.

To compare each method detection performances by the outlier strengths or numbers, we used a three-way ANOVA for repeated measures with factors Method (unnormalized odenRiSC, normalized odenRiSC, RP, MBT (95%)), Outlier Strength (10%, 30% or 50%), and Outlier Number (5, 10 or 25) for the artificial outliers dataset. For the actual motion outliers dataset, we performed a two-way ANOVA for repeated measures with two factors: Method (unnormalized odenRiSC, normalized odenRiSC, RP, MBT (95%) and Outlier Number, (5, 10 or 25). Sphericity was confirmed using Mauchly's sphericity test, and corrected using Greenhouse-Geisser if needed. When statistical significance was observed, a post-hoc analysis was performed using Tukey's honestly significant difference test.

VI. Experiments for Multimodal Classification

A. Data description for multimodal classification model

To evaluate mcRiSC, we conducted experiments under 2 different data variability conditions. One is EEG data recorded under a laboratory setting. This well-controlled experimental environment allows to record mostly clean EEG data, thus we assume that it has low variability. The other dataset was recorded in actual flight and is expected to be particularly contamined by several artifacts such as electromagnetic interferences (*e.g.*, GPS antenna, radio communication), vibrations (*e.g.*, engines), and pilots' muscular activity. Thus, we assumed this dataset to have many outliers and higher variability.

1) Inside-the-lab data: We used our two EEG datasets [41], [42], which consist of EEG signals recorded while subjects performed right and left-hand MI tasks. We used 56 subjects data with 27 channels (Fz, FCz, Cz, CPz, Pz, C1, C3, C5, C2, C4, C6, F4, FC2, FC4, FC6, CP2, CP4, CP6, P4, F3, FC1, FC3, FC5, CP1, CP3, CP5 & P3 sites) from [41], and 20 subjects with 30 channels (C6, CP4, CPz, CP3, P5, P3, P1, Pz, P2, P4, P6, PO7, PO8, Oz, F3, Fz, F4, FT8, FC6, FC4, FCz, FC3, FC5, FT7, C5, C3, C1, Cz, C2 & C4 sites) from [42] for our analysis, thus totally 76 subjects. The dataset from [41] consists of 6 runs and each run includes 20 trials for each class. The EEG signals were filtered by using a 4^{th} order Butterworth filter in a frequency band selected for each user according to the algorithm proposed in [43]. The dataset from [42] contains 5 runs and each run includes 20 trials per class. The EEG signals were filtered between 8–30 Hz, using a 4^{th} order Butterworth filter. In both datasets, a single time window was extracted in each trial from 0.5 to 2.5 sec after the MI instruction cue. The first two runs were used as training sets and the remaining runs were used as test sets.

2) Out-of-the-lab dataset: We used the dataset collected in real flight conditions [44]. Twenty-two pilots equipped with the 6 dry-electrode Enobio Neuroelectrics system (Fz, Cz, Pz, Oz, P3 & P4 sites) had to perform one low load and one high load navigation task (*i.e.* traffic pattern) along with a passive auditory oddball in a real single-engine aircraft. In the low load condition ($\sim 8min$), the volunteers had to monitor the flight path handled by the safety pilot while in the high load condition ($\sim 8min$), they had to operate the aircraft under the supervision of the flight instructor. Each traffic pattern lasted almost eight minutes in each workload condition.

Two hundred and thirty five epochs were extracted from successive and non overlapping epochs of 2 seconds for each flying condition. Each epoch was then band-pass filtered in the theta (4–7 Hz) and alpha (8–12 Hz) bands, using FIR filter with the filter order 450. These two frequency bands were chosen because they are known to be collaboratively associated with mental workload changes [44]. Next, EEG covariance matrices were estimated for each frequency band. Then, those theta-band and alpha-band covariance matrices were arranged diagonally as block matrices in a big covariance matrix.

B. Evaluation for multimodal classification

The performance of mcRiSC was compared with a standard unimodal classifier *i.e.*, MDRM [23]. Those method were evaluated in term of classification accuracy on the test set. The original train-test split (as performed in the original online experiment [41]) for the inside-of-the-lab dataset and 5-fold cross-validation for the out-of-the-lab dataset were used. The 4 different mcRiSC models were set; unnormalized mcRiSC with a k-nn graph, normalized mcRiSC with a k-nn graph, unnormalized mcRiSC with a fully connected graph and normalized mcRiSC with a fully connected graph.

To investigate statistical differences, we first examined the normality of the results using the Shapiro-Wilk test. If they were not normally distributed, the Friedman test was performed. Otherwise, sphericity was then confirmed using Mauchly's sphericity test, corrected using Greenhouse-Geisser if needed, and finally, a one-way ANOVA for repeated measures was performed. When statistical significance was observed, a post-hoc analysis was performed using Tukey's honestly significant difference test.

VII. Results

A. Outlier detection results for artificial outliers

Average HFDs for each contaminated condition are summarized in Table I. As we see diagonally the table from the top left to the bottom right (5 outliers with 10% to 25 outliers with 25%), when contamination became more severe, unnormalized and normalized odenRiSC performed better while existing methods worsened.

Three-way ANOVA for repeated measure with sphericity corrections revealed main effects of "*Method*" [$F(2.26, 267.43) = 277; p < 0.001$], "*Outlier Strength*" [$F(1.85, 220.74) = 1177; p < 0.001$], and "*Outlier Number*" [$F(1.82, 216.13) = 173; p < 0.001$]. It also revealed interactions for "*Method X Outlier Strength*" [$F(3.19, 380.05) = 434; p < 0.001$], and "*Method X Outlier Number*" [$F(3.27, 388.80) = 4011; p < 0.001$] and "*Method X Outlier Strength X Outlier Number*" [$F(4.88, 581.00) = 161; p < 0.001$]. Post-hoc analyses of "*Method*" showed unnormalized odenRiSC was significantly better than normalized odenRiSC [$MD = 8.40; p < 0.001$], RP [$MD = 12.5; p < 0.001$] and MBT (95%) [$MD = 8.74; p < 0.001$]. Normalized odenRiSC was significantly better than RP [$MD = 4.06; p < 0.001$] but did not show significant difference with MBT (95%) [$MD = 0.35; p = 0.751$]. Fig. 3 shows the distributions of HFDs for each method, outlier strength and outlier number. In term of outlier strength, HFDs with unnormalized and normalized odenRiSC increased as outlier strength increased, while RP and MBT (95%) did not show substantial change according to outlier strength. In term of outlier numbers, HFDs of unnormalized and normalized odenRiSC increased overall as outlier numbers increased. Meanwhile, RP and MBT (95%) worsened with increasing outlier numbers.

B. Outlier detection results for actual motion outliers

The average HFDs for each outlier number are summarized in Table II. Unnormalized odenRiSC showed constantly higher HFD than other methods. Two-way ANOVA for repeated measure with sphericity correction revealed main effects of "*Method*" [$F(1.90, 55.09) = 505; p < 0.001$] and "*Outlier Number*" [$F(1.57, 45.64) = 240; p < 0.001$]. It also revealed interactions for "*Method X Outlier Number*" [$F(2.73, 79.07) = 158; p < 0.001$]. Post-hoc analyses of "*Method*" showed unnormalized odenRiSC was significantly better than normalized odenRiSC [$MD = 7.40; p < 0.001$], RP [$MD = 44.77; p < 0.001$] and MBT (95%) [$MD = 38.71; p < 0.001$]. Also, normalized odenRiSC was significantly better than RP [$MD = 37.38; p < 0.001$] and MBT (95%) [$MD = 31.31; p < 0.001$]. Fig. 5 compares the HFD for each outlier detection method and each outlier number. HFDs of both odenRiSC models were constantly high, regardless of the amount of outliers. In contrast, performances of RP and MBT (95%) decreased with increasing outlier numbers.

TABLE I
AVERAGE HFD FOR EACH CONTAMINATED CONDITION OF ARTIFICIAL OUTLIERS [%]

Method	Outlier Number	Outlier Strength		
		10%	30%	50%
odenRiSC (unnormalized)	5 outliers	1.33 ± 8.10	38.7 ± 33.9	63.0 ± 19.5
	10 outliers	0.25 ± 2.03	57.8 ± 23.3	74.6 ± 11.3
	25 outliers	72.3 ± 19.3	87.5 ± 2.96	90.5 ± 2.94
odenRiSC (normalized)	5 outliers	0.00 ± 0.00	0.00 ± 0.00	4.83 ± 18.7
	10 outliers	0.00 ± 0.00	68.8 ± 23.1	77.8 ± 6.88
	25 outliers	80.7 ± 11.6	87.7 ± 2.63	90.5 ± 2.75
RP	5 outliers	54.2 ± 16.2	64.5 ± 11.1	67.5 ± 10.4
	10 outliers	38.8 ± 9.00	47.2 ± 5.53	49.4 ± 6.39
	25 outliers	17.3 ± 3.48	17.2 ± 3.10	17.6 ± 2.51
MBT(95%)	5 outliers	68.9 ± 14.7	78.6 ± 5.36	80.0 ± 0.00
	10 outliers	39.7 ± 1.79	40.0 ± 0.00	40.0 ± 0.00
	25 outliers	20.0 ± 0.00	20.0 ± 0.00	20.0 ± 0.00

TABLE II
AVERAGE HFDS FOR EACH OUTLIER NUMBER IN REAL EEG DATASET WITH ACTUAL MOTION OUTLIERS [%]

Method	Outlier Number		
	5 outliers	10 outliers	25 outliers
odenRiSC (unnormalized)	93.9 ± 3.45	90.1 ± 8.99	92.7 ± 3.11
odenRiSC (normalized)	80.1 ± 16.9	83.0 ± 13.3	91.5 ± 4.02
RP	82.6 ± 16.0	48.7 ± 10.4	11.2 ± 3.99
MBT(95%)	86.6 ± 13.0	50.0 ± 0.00	24.0 ± 0.00

C. Multimodal classification results - inside-the-lab condition

The average accuracies for each method are summarized in Table IV. As the accuracies were not normally distributed, a non-parametric Friedman test was performed: significant differences were not observed ($p = 0.468$). Normalized mcRiSC with a fully connected graph showed slightly better accuracy among mcRiSC models. This is achieved by detecting several modes for only two subjects, which increased accuracy for both subjects (mean gain: $5.84 \pm 3.34\%$).

D. Multimodal classification results - out-of-the-lab condition

The average classification accuracies of each method are summarized in the second row in Table IV. As shown in the table, normalized mcRiSC with k-nn graph achieved the highest average accuracy.

As the accuracy of each method were normally distributed, the one-way ANOVA for repeated measure was applied. The result revealed a main effect of "*Method*" [$F(1.43, 24.31) = 8.42; p = 0.004$] with sphericity corrections. Only normalized mcRiSC with a k-nn graph showed significantly better accuracy than unimodal MDRM [$MD = 4.27; p = 0.028$], while other mcRiSC models did not show statistical improvement. Within mcRiSC, normalized mcRiSC with a k-nn graph was significantly better than unnormalized mcRiSC with a fully connected graph [$MD = 4.40; p = 0.019$]. In terms of individual subjects, normalized mcRiSC with a k-nn graph showed the highest accuracy in 13 out of 18 subjects. From those results, normalized mcRiSC with a k-nn graph was the best model for the high-variability dataset recorded outside-the-lab.

E. Complexity of the multimodal classifier

Table III presents the running times of each mcRiSC model for the inside-the-lab and the out-of-the-lab datasets. They were measured

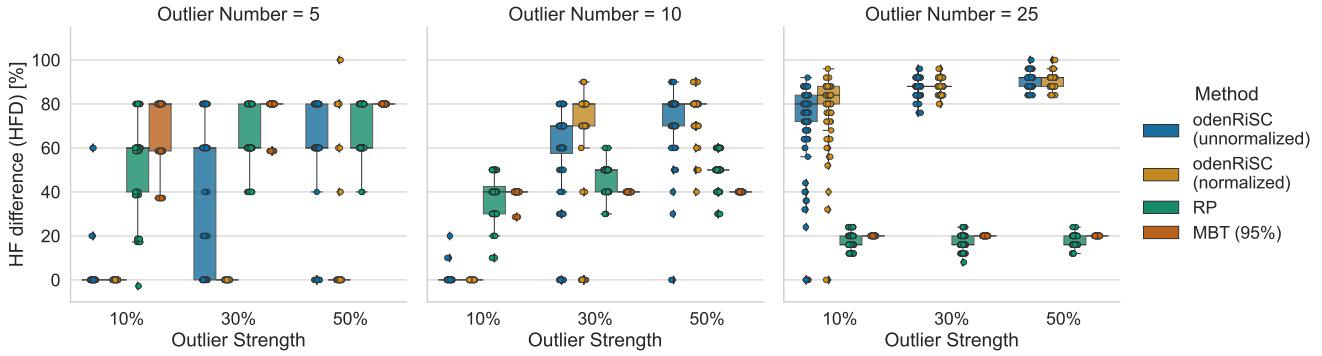

Fig. 3. HFD distribution for each method, according to outlier strength and outlier number of artificial outliers.

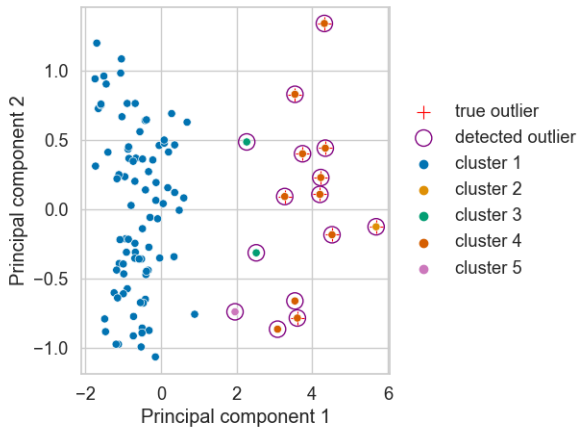

Fig. 4. Visualization of unnormalized odenRiSC modeling result for real EEG dataset with 10 outliers. The plot was generated by first projecting all data points on a tangent space using pyRiemann(v0.4) [45] and then applying principal component analysis (PCA) from Scikit-learn (v1.2.2) [46] to keep two components. For the tangent space mapping, the Riemannian mean of all data points was selected as the tangent point.

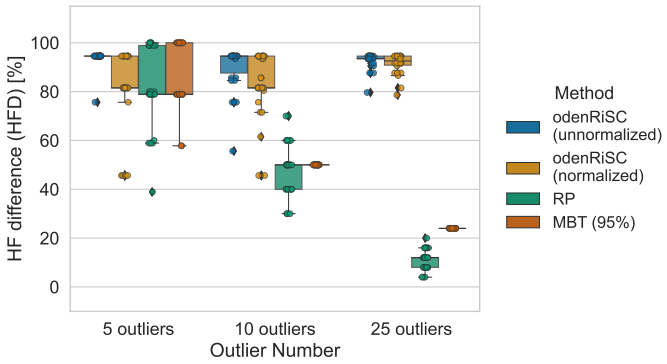

Fig. 5. HFDs distribution for each outlier number in the real EEG dataset with actual motion outliers.

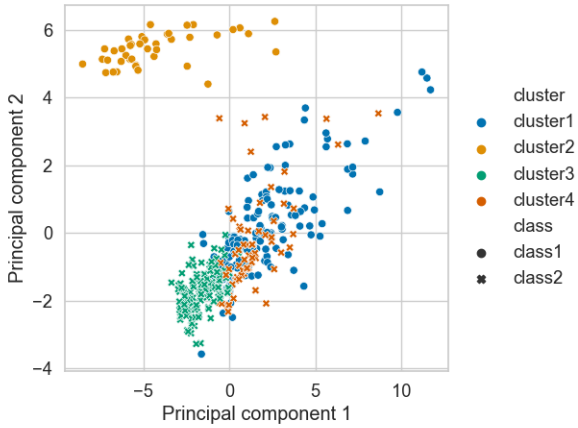

Fig. 6. Visualization of normalized mcRiSC(w/ k-nn graph) modeling result for out-of-the-lab dataset. The plot was made the same way as Fig. 4 using only training data from a CV-fold of a single subject.

on an Intel Core i5 2GHz CPU with Matlab R2022b. Note that we report the average time over 10 iterations with its standard deviation. All mcRiSC models were computationally more expensive than a unimodal classifier. However, it was still very fast, requiring less than one second and a half for training and less than a millisecond per trial for testing. Among mcRiSC, the models with fully connected graph were faster than the ones with k-nn graph overall.

VIII. Discussion

A. Outlier detection

1) Overview of odenRiSC: The aim of this outlier detection study was to assess the performance under different contamination conditions. In practice, when analyzing EEG dataset, we cannot know how many and how strong outliers exist in advance. Overall, our proposed method showed accurate performance when the dataset was highly contaminated in term of either outlier strength or number. This fact was clearly demonstrated with the dataset with actual motion outliers. Fig. 4 represents one of the modeling results of unnormalized odenRiSC for that dataset. RiSC detected some data labeled to be clean as outliers, but which exhibited outlier-like behavior (*i.e.*, the purple circles without a red cross in the figure). Based on these results, if the user suspects there may be moderate or strong outliers in the dataset, it makes sense to use odenRiSC to clean it up. The downside of the current odenRiSC is that we can only use it offline, as outlier detection is based on building a graph using all data points. The generalization to online setups will be tackled in the future.

2) Comparison between odenRiSC models: Among our proposed methods, unnormalized odenRiSC showed the best performance for both artificial outliers and actual motion outliers. For the artificial outliers, performances of unnormalized and normalized odenRiSC both increased as contamination became more severe, whereas existing methods decreased. However, HFDs were low for datasets more weakly contaminated. Notably, normalized odenRiSC did not detect any outlier for 5 outliers with 10% , 30% strength nor for 10 outliers with 10% strength. This may be due to EEG

TABLE III
AVERAGE RUNNING-TIMES [S] FOR LOW AND HIGH VARIABILITY DATASETS (INSIDE OR OUTSIDE THE LAB) OVER 10 ITERATIONS

		Method				
		unimodal	knn+unnorm	knn+norm	full+unnorm	full+norm
inside-the-lab (27-30 ch)	train (80 trials)	0.08 ± 0.02	0.40 ± 0.06	0.41 ± 0.05	0.28 ± 0.03	0.28 ± 0.02
	test (per trial)	$(3.3 \pm 0.2) \times 10^{-4}$	$(7.3 \pm 2.6) \times 10^{-4}$	$(7.3 \pm 2.8) \times 10^{-4}$	$(3.4 \pm 0.1) \times 10^{-4}$	$(3.4 \pm 0.3) \times 10^{-4}$
out-of-the-lab (4 ch)	train (376 trials)	0.10 ± 0.05	1.34 ± 0.08	1.31 ± 0.05	0.87 ± 0.06	0.88 ± 0.05
	test (per trial)	$(5.8 \pm 0.6) \times 10^{-5}$	$(16 \pm 5.3) \times 10^{-5}$	$(17 \pm 5.3) \times 10^{-5}$	$(6.4 \pm 2.4) \times 10^{-5}$	$(7.2 \pm 2.7) \times 10^{-5}$

covariance matrices being based on the variance computation of EEG signal. Indeed, when estimating variance of the EEG signal over a long time window, short artifacts may have a small impact on the resulting overall variance. In other words, if artifacts affect a small proportion of EEG signals, the resulting EEG covariance matrices may be rather similar to non-outlier EEG covariance matrices. Thus, those weakly contaminated covariance matrices may locate near the boundary of the non-outlier cluster on a manifold.

For the real data with artifact labels, unnormalized and normalized odenRiSC both showed overall higher performance. In this experiment, the entire time periods used to estimate covariance matrices were contaminated by motion artifacts. Thus, those outliers were highly contaminated. The observed results were thus consistent with the experimental results on artificial outliers with high-contamination. The HFD of normalized odenRiSC was lower than unnormalized odenRiSC as non-outlier data was also divided into several clusters, which led to the high FPR (*i.e.*, higher ratio of erroneous rejection of clean data). Normalized spectral clustering is widely known to work better than unnormalized one in general clustering problem [47]. This may explain why normalized spectral clustering tended to capture not only the difference between non-outliers and outliers, but also the difference among non-outliers. This needs to be further investigated in future dedicated experiments.

3) Comparison with existing methods: The HFD of RP decreased with increasing outlier numbers in both the artificial outliers dataset and the actual motion outliers dataset. This may be because as the number of outliers increases, standard deviation, one of the parameters defining the rejection threshold in RP, increases. This led to a larger rejection threshold and decreased the amount of detected outliers. MBT (95%) did not differ in the removal ability by changing the outlier strength in the artificial outliers dataset. This is because MBT (95%) determines outliers as a fixed percentage of data. In the artificial outliers dataset, RP and MBT (95%) both showed better performance than unnormalized and normalized odenRiSC for low-contamination datasets. This may be because they both estimate the location of each EEG covariance matrix using their distance from a reference matrix and then detect outliers with a threshold. Therefore, they can handle "weak" outliers interspersed near non-outliers. However, this may carry the risk of rejecting non-outliers as outliers. There is need of the further investigation to assess the risk of removing weak outliers on classification accuracy.

B. Multimodal classification

1) Overview of mcRiSC: The performance was evaluated by comparing our multimodal mcRiSC classifier to the unimodal MDRM classifier using two different datasets, one recorded inside a lab and one recorded outside a lab, where the degree of data variability is expected to be different. Results suggested mcRiSC, especially normalized mcRiSC with a k-nn graph, may be useful to improve BCI for highly contaminated and variable datasets. Fig. 6 shows one representative subject results for normalized mcRiSC with a k-nn graph, for the out-of-the-lab dataset. We can see that mcRiSC detected two different dense areas (*i.e.*, cluster 1 and cluster 2) in class 1. If a unimodal classifier had been used, the class centroid would have been estimated somewhere in the middle of those two parts, *i.e.*, in an area with fewer actual data points, which is not representative of the actual distribution. Theoretically, mcRiSC can be utilized for both online and offline BCI setups. For online setups, the training and testing times should meet the constraints of practical BCI use. While mcRiSC was computationally costlier than MDRM, it was still largely fast enough for real-time BCIs, needing less than one second and a half for training and less than a millisecond per trial for testing.

The strength of mcRiSC is its flexibility to adapt to both unimodal and multimodal distributions. In general, if data distribution is multimodal, a multimodal classifier is more representative than a unimodal classifier. However, a multimodal classifier may memorize too precisely the training data characteristics and is thus more likely to be influenced by spurious patterns or noises, i.e., overfitting. Ideally, the learning model should be balanced between underfitting and overfitting, in other words, rich enough to express underlying structure of data and simple enough to avoid learning spurious patterns, so called bias-variance trade-off [48]. Overfitting was not particularly observed in our experiments, which may be because mcRiSC removed outlier clusters from class centroid estimation, as illustrated in Fig.2, and because there was a constrain for the maximum cluster number. We will continue our investigation to see if our model balance well between underfitting and overfitting with different types of EEG data.

2) Low variability dataset (inside-the-lab): We could not observe a statistically significant difference in average classification accuracy between the unimodal classifier *i.e.* MDRM, and our multimodal classifiers *i.e.* mcRiSC. In fact, all mcRiSC models did not detect multiple modes for most subjects, in other words, mcRiSC worked as a unimodal classifier. This may be due to the experimental conditions of this EEG dataset. In this MI-BCI experiment, experimenters asked users to perform one fixed kinesthetic MI strategy during the calibration runs, which is expected to lead to low variability. McRiSC was applied to these calibration runs, which is probably why a single mode was detected in most cases.

3) High variability dataset (out-of-the-lab): We observed statistically significant differences between methods, with notably normalized mcRiSC with a k-nn graph showing the highest improvement from the unimodal classifier. This dataset was recorded in a real aircraft, which is particularly noisy in terms of electromagnetic interferences, vibrations and muscular activity. Thus, it is expected in this context that the computed covariance matrices have high variability on the manifold. Within mcRiSC, normalized mcRiSC with a k-nn graph showed the best performance, while a fully connected graph showed better performance than a k-nn graph for low variability dataset. This is because a k-nn graph can break into several disconnected components if there are high density regions which are reasonably far away from each other. Thus, a k-nn graph can capture different density parts and divide them precisely.

TABLE IV
AVERAGE CLASSIFICATION ACCURACY OF LOW AND HIGH VARIABILITY DATASET (INSIDE AND OUTSIDE THE LAB)

Experimental condition	Method				
	unimodal	knn+unnorm	knn+norm	full+unnorm	full+norm
inside-the-lab	64.0 ± 14.4	62.3 ± 13.6	62.6 ± 13.4	64.1 ± 14.5	$\mathbf{64.2 \pm 14.3}$
out-of-the-lab	68.8 ± 11.4	72.3 ± 11.3	$\mathbf{73.1 \pm 12.0}$	68.7 ± 11.4	69.6 ± 11.0

IX. PRACTICAL RECOMMENDATIONS FOR USING ODENRISC AND MCRISC MODEL ON YOUR PROBLEM

Our modeling method is modular, so we can easily substitute different choices, *e.g.*, choose different graph and spectral clustering methods. However, this leads to the question "which odenRiSC/mcRiSC model is optimal for our purpose?". Before concluding this paper, we provide a guideline for this question with a posterior analysis result.

Regarding the similarity graph choice, we recommend using a fully connected graph for odenRiSC, *i.e.*, outlier removal, and a k-nn graph for mcRiSC, *i.e.*, multimodal classification. From a theoretical point of view, generally, a fully connected graph connects all nodes regardless of the distribution density of the nodes. In contrast, a k-nn graph forms a graph separating the dense regions that are reasonably far from each other. As odenRiSC detects all clusters as outliers except the one with the maximum size, if we use a k-nn graph while clean samples are spread over multiple density regions, there is a risk of accidentally removing some of them as outliers. On the other hand, for multimodal classification, it is desirable to separate precisely high-density regions that are reasonably separated for estimating multiple centroids. Fig. 7 shows the posterior analysis results with the high variability dataset, in order to investigate how the average accuracy change from k-nn with small k-value to fully connection in mcRiSC. As we can see clearly, the average accuracy over subjects tended to globally decrease as the number of neighbors increased.

Another important point is the selection of k value for constructing a similarity graph. Indeed, we learned that the general heuristic used in our main experiments did not provide us with an optimal k-value in hindsight from the posterior analysis result (Fig. 7). Thus, the reader may want to build a k-nn graph by using the general heuristic provide but it is also worth trying different k-values around that given value.

Regarding the spectral clustering choice, unormalized spectral clustering is a safer choice for outlier removal. This is because if an outlier is an isolated point on a manifold, the denominator of Eq. 6 turns to be zero and consequently the objective function does not converge. In multimodal classification, normalized spectral clustering is more reasonable because the spectral clustering technique is used more like a general clustering problem and it generally performs better than unnormalized one for many clustering problems [47].

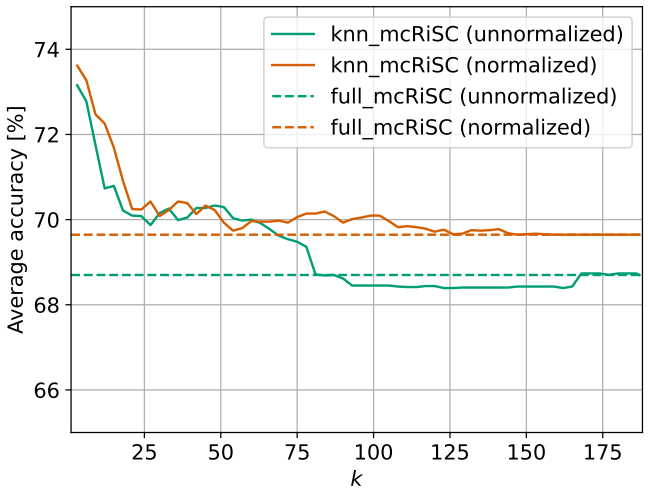

Fig. 7. Fluctuation of average classification accuracy over all subjects according to k-value in k-nn mcRiSC for the high variability dataset. The dotted lines indicate results of mcRiSC with fully connected graph.

X. CONCLUSION

In this paper, we have proposed modeling multimodal distributions based on clustering EEG covariance matrices on a Riemannian manifold using spectral clustering. Our modeling method, RiSC, can be used as a basis for outlier detection or for multimodal classification model design. As an outlier detector named odenRiSC, we successfully removed the necessity of setting two parameters: a reference matrix and a threshold which were required in existing methods. Instead of setting those parameters, odenRiSC clustered EEG covariance matrices into a non-outlier cluster and outlier clusters according to their geometrical similarity on a manifold. As a multimodal classification model, we tackled the open research challenge mentioned in [12], which is the need for MDRM with multiple modes per class. Our method, named mcRiSC, estimated multiple modes per class in a data-driven way and classified a new observation according to the nearest class centroid among multiple choices. This enabled us to consider the actual shape of covariance matrix distributions on a manifold for classification. Experimental evaluations revealed the superiority of odenRiSC and mcRiSC compared to existing methods. OdenRiSC could detect EEG outliers more accurately than existing Riemannian EEG outlier detection methods especially for the more contaminated data. This suggests that odenRiSC may contribute to build more robust classifiers by appropriately removing outliers from training dataset in the future, when the dataset is highly contaminated. Also, odenRiSC may contribute for EEG data screening in neuroscience study. McRiSC showed significantly higher classification accuracy than MDRM (*i.e.*, a unimodal Riemannian classifier) for the high variability dataset. From those aforementioned main results, we anticipate odenRiSC and mcRiSC will both help to lead BCIs outside laboratories, *e.g.*, for neuroergonomics applications [49], which are expected to suffer from various artifacts and variability sources.

As future works, first, we will continue investigating mcRiSC performance, especially under inside-the-lab conditions. The inside-the-lab dataset we used might have a unimodal distribution due to the constraint of one fixed kinesthetic MI strategy during the training data recording. However, even if the dataset is recorded inside a lab, data variability can be large, such as dataset with no fixed kinesthetic MI strategies or dataset recorded cross-days. Thus, we will test if mcRiSC can detect multiple modes appropriately and contribute to improving classification accuracy for those datasets. We will also investigate alternative metrics for similarity measures. Here, we only used the AIRM distance, but it could be worth trying different Riemannian metrics such as the Log-Euclidean distance. Furthermore, we will explore the way to adapt odenRiSC for online BCI setups, for single-trial outlier detection in real-time. In addition,

as odenRiSC and mcRiSC are both generic approaches for any type of EEG signal, we will test our methods with Steady-State Visual Evoked Potentials, Event-Related Potentials, Sleep EEG, etc. As more general challenge, we are eager to collect new EEG datasets specialized for data variability study. We need a ground truth to understand what factors cause different modes on a manifold and to evaluate whether mcRiSC appropriately detects modes related to different variability sources. Those ground truth datasets will help us to identify what needs to be solved for the variability issue to make BCIs more reliable.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-Computer Interfaces for Communication and Control," *Clin. Neurophysiol.*, vol. 113(6), pp. 767–791, 2002.

[2] M. Clerc, L. Bougrain, and F. Lotte, *Brain-Computer Interfaces 1*, Wiley-ISTE, 2016.

[3] J. J. Vidal, "Toward direct Brain-Computer Communication," *Annu. Rev. Biophys.*, vol. 2, no. 1, pp. 157–180, 1973.

[4] G. Pfurtscheller, G. R. Müller-Putz, R. Scherer, and C. Neuper, "Rehabilitation with Brain-Computer Interface systems," *Computer*, vol. 41, no. 10, pp. 58–65, 2008.

[5] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An EEG-based Brain–Computer Interface for cursor control," *EEG and Clin. Neurophysiol.*, vol. 78, no. 3, pp. 252–259, 1991.

[6] D. Coyle, J. Principe, F. Lotte, and A. Nijholt, "Guest Editorial: Brain/neuronal-Computer game interfaces and interaction," *IEEE Trans. Comput. Intell. AI Games*, vol. 5, no. 2, pp. 77–81, 2013.

[7] J. Mladenovic, J. Frey, S. Pramij, J. Mattout, and F. Lotte, "Towards identifying optimal biased feedback for various user states and traits in motor imagery BCI," *IEEE Trans. Biomed. Eng.*, 2021.

[8] S. Rimbert, P. Riff, N. Gayraud, D. Schmartz, and L. Bougrain, "Median nerve stimulation based BCI: a new approach to detect intraoperative awareness during general anesthesia," *Front. Neurosci.*, vol. 13, 2019.

[9] F. Lotte and R. N. Roy, "Brain–Computer Interface contributions to neuroergonomics," in *Neuroergonomics*, pp. 43–48. Elsevier, 2019.

[10] F. Dehais, W. Karwowski, and H. Ayaz, "Brain at work and in everyday life as the next frontier: grand field challenges for neuroergonomics," *Front. Neuroergonomics*, p. 1, 2020.

[11] F. Dehais, S. Ladouce, L. Darmet, T-V. Nong, G. Ferraro, J. Torre Tresols, S. Velut, and P. Labedan, "Dual Passive Reactive Brain-Computer Interface: A Novel Approach to Human-Machine Symbiosis," *Front. Neuroergonomics*, vol. 3, 2022.

[12] 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," *J. Neural Eng.*, vol. 15, no. 3, pp. 031005, 2018.

[13] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in Brain-Computer Interfaces: a review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 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," *J. BCIs*, vol. 4, no. 3, 2017.

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

[16] R. N. Roy, M. F. Hinss, L. Darmet, S. Ladouce, E. S. Jahanpour, B. Somon, X. Xu, N. Drougard, F. Dehais, and F. Lotte, "Retrospective on the First Passive Brain-Computer Interface Competition on Cross-Session Workload Estimation," *Front. Neuroergonomics*, p. 4, 2022.

[17] B. Z. Allison and C. Neuper, "Could anyone use a BCI?," in *J. BCIs*, pp. 35–54. Springer, 2010.

[18] F. Lotte and C. Jeunet, "Towards improved BCI based on human learning principles," in *The 3rd Int. Winter Conf. on BCI*. IEEE, 2015, pp. 1–4.

[19] S. Saha and M. Baumert, "Intra-and inter-subject variability in EEG-based sensorimotor Brain Computer Interface: a review," *Front. Comput. Neurosci*, p. 87, 2020.

[20] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "EMG and EOG artifacts in Brain Computer Interface systems: A survey," *Clin. Neurophysiol.*, vol. 118, no. 3, pp. 480–494, 2007.

[21] A. K. Porbadnigk, N. Görnitz, C. Sannelli, A. Binder, M. Braun, M. Kloft, and K-R. Müller, "When brain and behavior disagree: Tackling systematic label noise in EEG data with machine learning," in *Int. Winter Workshop on BCI*, 2014, pp. 1–4.

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

[23] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer Interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59–4, pp. 920–928, 2011.

[24] A. Barachant, A. Andreev, and M. Congedo, "The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry," in *TOBI Workshop lV*, 2013, pp. 19–20.

[25] T. Uehara, M. Sartori, T. Tanaka, and S. Fiori, "Robust averaging of covariances for EEG recordings classification in motor imagery Brain-Computer Interfaces," *Neural Comput.*, vol. 29(6), pp. 1631–1666, 2017.

[26] MS. 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 *EMBC2020*. IEEE, 2020, pp. 438–441.

[27] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Parameters estimate of Riemannian Gaussian distribution in the manifold of covariance matrices," in *IEEE SAM*. IEEE, 2016, pp. 1–5.

[28] N. Boumal, "An introduction to optimization on smooth manifolds," *Available online, May*, vol. 3, 2020.

[29] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vision*, vol. 66–1, pp. 41–66, 2006.

[30] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *LVA/ICA 2010*. Springer, 2010, pp. 629–636.

[31] S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton, "Riemannian Gaussian distributions on the space of symmetric positive definite matrices," *IEEE Trans. Inf. Theory.*, vol. 63(4), pp. 2153–2170, 2017.

[32] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[33] B. Mohar, Y Alavi, G Chartrand, and OR Oellermann, "The Laplacian spectrum of graphs," *Graph theory, combinatorics, and applications*, vol. 2, no. 871-898, pp. 12, 1991.

[34] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 11, no. 9, pp. 1074–1085, 1992.

[35] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Patt. Anal. Mach. intel.*, vol. 22, no. 8, pp. 888–905, 2000.

[36] H. Lutkepohl, "Handbook of matrices.," *Comp. stat. Data anal.*, vol. 2, no. 25, pp. 243, 1997.

[37] M. Tangermann, KR. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, KJ. Miller, G. Mueller-Putz, and et. al, "Review of the BCI competition IV," *Front. Neurosci.*, 2012.

[38] R. A. Maronna, R. D. Martin, V. J. Yohai, and et. al, *Robust statistics: theory and methods (with R)*, John Wiley & Sons, 2019.

[39] K. T. Sweeney, H. Ayaz, T. E. Ward, M. Izzetoglu, S. F. McLoone, and B. Onaral, "A methodology for validating artifact removal techniques for physiological signals," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 5, pp. 918–926, 2012.

[40] S. Mason, J. Kronegg, J. Huggins, M. Fatourechi, and A. Schlögl, "Evaluating the performance of self-paced Brain-Computer Interface technology," *Neil Squire Soc., Vancouver, BC, Canada, Tech. Rep*, 2006.

[41] L. Pillette, A. Roc, B. N'kaoua, and F. Lotte, "Experimenters' influence on mental-imagery based Brain-Computer Interface user training," *Int. J. Hum. Comput. Stud.*, vol. 149, pp. 102603, 2021.

[42] C. Jeunet, E. Jahanpour, and F. Lotte, "Why standard Brain-Computer Interface (BCI) training protocols should be changed: an experimental study," *J. Neural Eng.*, vol. 13, no. 3, pp. 036024, 2016.

[43] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process Mag.*, vol. 25, no. 1, pp. 41–56, 2007.

[44] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. N. Roy, and F. Lotte, "Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions," *Sensors*, vol. 19, no. 6, pp. 1324, 2019.

[45] A. Barachant and et al., "pyriemann/pyriemann: v0.3," July 2022.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] U. Von Luxburg, O. Bousquet, and M. Belkin, "Limits of spectral clustering," *Adv Neural Inf Process Syst*, vol. 17, 2004.

[48] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *PNAS*, vol. 116, no. 32, pp. 15849–15854, 2019.

[49] R. Parasuraman, "Neuroergonomics: Research and practice," *Theor. Issues Ergon. Sci.*, vol. 4, no. 1-2, pp. 5–20, 2003.