Is Riemannian Geometry Better than Euclidean in Averaging Covariance Matrices for CSP-based Subject-Independent Classification of Motor Imagery?

Yassawe Kainolda\textsuperscript{1}, Berdakh Abibullaev\textsuperscript{2}, Reza Sameni\textsuperscript{3}, Amin Zollanvari\textsuperscript{1}

Abstract—Common Spatial Pattern (CSP) is a popular feature extraction algorithm used for electroencephalogram (EEG) data classification in brain-computer interfaces. One of the critical operations used in CSP is taking the average of trial covariance matrices for each class. In this regard, the arithmetic mean, which minimizes the sum of squared Euclidean distances to the data points, is conventionally used; however, this operation ignores the Riemannian geometry in the manifold of covariance matrices. To alleviate this problem, Fréchet mean determined using different Riemannian distances have been used. In this paper, we are primarily concerned with the following question: Does using the Fréchet mean with Riemannian distances instead of arithmetic mean in averaging CSP covariance matrices improve the subject-independent classification of motor imagery (MI)? To answer this question we conduct a comparative study using the largest MI dataset to date, with 54 subjects and a total of 21,600 trials of left- and right-hand MI. The results indicate a general trend of having a statistically significant better performance when the Riemannian geometry is used.

I. INTRODUCTION

Brain-Computer Interface (BCI) systems aim to infer different brain activity patterns of a user as accurately as possible and translate them into appropriate commands [1].

Most studies use electroencephalography (EEG) signals generated by the collective action of millions of cortical cells in the development of BCI systems. The key component of such BCI systems are the signal processing and classification methods that allow the extraction of discriminative features of EEG in which human thoughts are best encoded. Although there has been significant progress in the field, most studies still focus on subject-dependent classification of user mental states from EEG data [2], [3]. A subject-independent classification of users’ mental states remains one of the major challenges in designing practical BCIs. This is due to enormous variability in data; in other words, EEG distribution is highly variable between different users, experimental sessions, and even between different trials [4]. As a result, a classifier model trained on one subject’s data will not generalize to the data of another subject. Such variability dramatically affects most BCI performances and requires tedious subject-specific calibration and classifier modeling. This study focuses on the problem of subject-independent classification of EEG recordings acquired from motor imagery (MI) experiments.

Common Spatial Pattern (CSP) is one of the most popular feature extraction methods for EEG data, which creates a set of spatial filters that transform the data to be more discriminative in terms of variances [5], [6]. It is instrumental in the MI paradigm, much of useful EEG features in the spatial domain, since brain areas responsible for different limb movements are spatially separated.

CSP works by simultaneously diagonalizing the average covariance matrices of trials of each class. Most commonly, average covariance matrices are estimated by arithmetic averaging. Arithmetic averaging, however, suffers from a series of problems, the most significant of which is the swelling effect [7], resulting in the determinant of the average covariance matrix (which is a measure of the overall scatter of the samples in the feature space), being much larger than determinants of individual matrices that comprised the average. Such a distortion is an undesirable artifact of the computation. Since covariance matrices are by construction symmetric positive definite (SPD) and therefore lie on an SPD manifold, their underlying geometry is better captured by Riemannian methods [8], [9]. This work applies that reasoning and compares the results produced by CSP with arithmetic averaging and CSP with Riemannian means in the subject-independent classification framework.

The rest of the paper is organized as follows: in Section II the relevant concepts of the Common Spatial Pattern algorithm and Riemannian geometry will be reviewed, in Section III the methodology of the experiment will be explained, and Section IV presents the results of the experiment.

II. BACKGROUND

A. Common Spatial Pattern

Let $X_i$ denote the $i^{th}$ trial data matrix of size $C \times N$ where $C$ is the number of EEG channels and $N$ is the number of observations. The spatial covariance matrix for each trial is then

$$C_i = \frac{X_i X_i^T}{\text{tr}(X_i X_i^T)},$$

(1)
where $T$ and $\text{tr}(.)$ denote the transpose and trace operators, respectively. In the case of a two-class paradigm, CSP performs a linear transformation that maps the data from the original sensor space to a new surrogate feature space, in which the variance of one class is maximized while the variance of the other is minimized [10]. To achieve this transformation, CSP finds a set of $M$ spatial filters given by an $C \times M$ matrix $W$ to linearly transform an input signal. Finding $W$ involves solving the generalized eigenvalue problem, which involves the class-specific “average” covariance matrices given by the usual arithmetic mean [10]:

$$
\Sigma_{i,a} = \frac{1}{T_i} \sum_{t=1}^{T_i} C_t,
$$

where $i = 1, 2$ denote class indexes, $T_i$ is the number of trials in class $i$ and the subscript “a” is used to denote the use of arithmetic mean. An input data matrix $X$ is then transformed as

$$
S = W^T \text{cropped} X,
$$

where $W_{\text{cropped}}$ is a matrix of the shape $C \times (2n)$, which in order to achieve the most separability it consists of only the first and the last $n$ columns of $W$ [10]—it is common to choose a small value of $n$ (e.g., $n = 3$). The feature vector $f = [f_1, f_2, ..., f_{2n}]$ is then obtained as:

$$
\begin{equation}
\begin{aligned}
f_r & = \log \frac{\text{var}(s_r)}{\sum_{k=1}^{2n} \text{var}(s_k)},
\end{aligned}
\end{equation}
$$

where $\text{var}(\cdot)$ denotes the variance and $s_r$ is the $r$th row of $S$, $r = 1, ..., 2n$. This feature vector is then used as the input to a classifier of choice.

**B. Riemannian Mean**

The use of Euclidean geometry in CSP is manifested through the arithmetic mean (2) [11]. However, employing Euclidean geometry in manipulating SPD covariance matrices has multiple drawbacks: 1) it forms a non-complete space; that is, Euclidean computations in general may lead to indefinite matrices [12]; 2) arithmetic mean is not invariant to matrix inversion, which is an undesirable property in some applications [13]; and 3) it leads to the well-known swelling effect, which implies the determinant of the arithmetic mean matrix may become larger than the original determinants [7] – in data analysis this may add harmful spurious variations to the data [12].

The aforementioned issues are due to the application of Euclidean geometry in the space of SPD matrices, which is a differentiable Riemannian manifold [14]. To overcome difficulties associated with the arithmetic mean of SPDs in (2), one may use instead the Fréchet mean (also known as Riemannian mean) defined as [7]

$$
\begin{equation}
\begin{aligned}
\Sigma_{i,r} & = \arg\min_{C \in P(C)} \sum_{t=1}^{T_i} \delta_r^2(C, C_t),
\end{aligned}
\end{equation}
$$

where $P(C)$ is the set of all $C \times C$ SPD matrices, $C_1, ..., C_{T_i}$ is the set of given SPD covariance matrices for $T_i$ trials, $\delta_r(A, B)$ is a Riemannian distance between two SPD matrices $A$ and $B$, and the subscript “r” is used to denote the use of Riemannian mean. Depending on the choice of scalar multiplication to equip the tangent spaces, two Riemannian distances are commonly used: 1) the Log-Euclidean metric (LEM) [7]; and 2) the Affine-Invariant Riemannian metric (AIRM) [15].

Formally, for two SPD matrices $A$ and $B$, LEM distance $\delta_{\text{LEM}}(A, B)$ is given by [7]:

$$
\begin{equation}
\begin{aligned}
\delta_{\text{LEM}}(A, B) &= \| \log(A) - \log(B) \|_F,
\end{aligned}
\end{equation}
$$

where $\log(A)$ is the matrix logarithm, $\|A\|_F$ is the Frobenius norm of $A$ defined as $\sqrt{\text{tr}(A^T A)}$. Using LEM distance leads to a closed-form solution for calculating the mean as follows [7]:

$$
\begin{equation}
\begin{aligned}
\Sigma_{r,\text{LEM}} & = \exp\left( \frac{1}{T_i} \sum_{t=1}^{T_i} \log(C_t) \right),
\end{aligned}
\end{equation}
$$

where $\exp(\cdot)$ denotes the matrix exponential. AIRM distance, on the other hand, is defined as [16]

$$
\begin{equation}
\begin{aligned}
\delta_{\text{AIRM}}(A, B) &= \| \log(A^{-1/2}B A^{-1/2}) \|_F.
\end{aligned}
\end{equation}
$$

In contrast with the LEM distance, the Fréchet mean determined using AIRM distance has no closed-form solution. In this regard, optimization methods such as incremental Fréchet mean estimation can be used to find the mean [17].

**III. Material and Methods**

A. Dataset

The dataset consists of EEG recordings of 54 subjects recorded using 62 Ag/AgCl electrodes with a sampling rate of 1000 Hz. Electrodes were placed according to the international 10-20 system [18]. The dataset contains two motor imagery classes: 1) imagined movement of the left hand; and 2) imagined movement of the right hand. Furthermore, the dataset is organized in two sessions, where each session contains 200 trials for each class and subject. For more information on the experimental protocol, the readers are referred to [18].

There are in total 21600 trials collected across 54 subjects in this dataset, which to the best of our knowledge makes it the largest publicly available EEG MI dataset to date. For the sake of comparison, there are a total of 5184 trials across 9 subjects in BCI Competition IV 2a [19] and 10400 trials across 52 subjects in [20] datasets, which have been frequently used in EEG-based MI studies before [21], [22], [23].

In our simulations, we used the 20 channels from the motor cortex region that were used in [18] (FC-5/3/1/2/4/6, C-5/3/1/2/4/6, and CP-5/3/1/2/4/6). The placement of these channels is shown in Fig. 1. The data was further bandpass filtered within 8-30 Hz using a fifth order Butterworth filter.
B. Methods

Depending on the approach used to compute the class-
specific average covariance matrices, we use and compare the
following three types of CSP feature extraction: I) CSP with
\( \Sigma_{i,a} \) computed from (2); II) CSP with \( \Sigma_{i,LEM} \) computed from
(7); and III) CSP with \( \Sigma_{i,AIRM} \), which is obtained by using
\( d_{AIRM}(A, B) \) in (5). Hereafter, we refer to these methods as
conventional CSP, LEM-CSP, and AIRM-CSP, respectively,
whereas by Riemannian-CSP we refer to both LEM-CSP and
AIRM-CSP.

In order to simulate the subject-independent classification,
we use a Leave-One-Subject-Out Cross-Validation (LOSO-
CV) procedure where we successfully hold out the ob-
servations (trials) collected from each subject, construct a
classifier using the pooled set of observations from the re-
aining subjects, and test the performance of the constructed
classifier on data from the held-out subject. To construct our
classifiers, each training and test observation was trans-
formed into a feature vector using (4) for \( n = 2, ..., 5 \). These
feature vectors were used as the input to linear discriminant
analysis (LDA) classifier [24].

IV. RESULTS

Here, we compare the subject-independent classification
performance obtained using the Riemannian-CSP with re-
spect to the conventional CSP. In this regard, we first estimate
the distributions of the difference between the LOSO-CV
classification accuracy obtained using Riemannian-CSP and
conventional CSP (i.e., empirical deviation distribution). The
distribution deviations provide a comprehensive picture of
the conducted comparisons in each setting.

Fig. 2 shows the plots of the empirical deviation distribu-
tion across session 1 and 2 for different \( n \). The plots were
obtained by fitting a beta density to the raw histograms of
classification accuracies estimated using all test observations
for each of the 54 subjects used as part of the LOSO-
CV procedure. Therefore, the density of a point larger
(less) than zero shows the likelihood of an improvement
in classification accuracy equivalent to that point obtained
using the Riemannian-CSP (conventional CSP) with respect
to the conventional CSP (Riemannian-CSP). As a result,
the area of the shaded region (the grey region) in each
density plot is an estimate of the probability of achieving
a higher classification accuracy using the Riemannian-CSP
with respect to the conventional CSP (we simply refer to this
probability as the probability of improvement denoted as \( P_I \)).
As we can see in Fig. 2, out of 8 scenarios (2 sessions \( \times 4 \)
values of \( n \)) to compare the LEM-CSP and the conventional
CSP, in 7 cases \( P_I > 54\% \) with 4 cases \( P_I > 60\% \). This
shows a clear advantage of the LEM-CSP with respect to
the conventional CSP. We observe a similar situation when
AIRM-CSP is used. In this case, out of 8 similar scenarios,
in 7 cases \( P_I > 53\% \) with 4 cases \( P_I > 60\% \).

For a more conventional comparison, for each scenario we
also recorded the the average classification accuracy obtained
across all 54 subjects and for session 1 and 2 (± standard
deviations) in Tables I and II, respectively. As we can see in
these tables, both the AIRM-CSP and the LEM-CSP show a
larger average accuracy than the conventional CSP in 7 out
of 8 scenarios. The results of a paired Wilcoxon signed rank
test show that in 7 and 6 out of 8 scenarios, the improvement
observed by LEM-CSP and AIRM-CSP, respectively, with
respect to the conventional CSP is statistically significant at
a significance level of 0.1 (see Table III).

TABLE I
AVERAGE SUBJECT-INDEPENDENT ACCURACIES FOR SESSION 1 (±
STANDARD DEVIATIONS)

<table>
<thead>
<tr>
<th>n</th>
<th>Conventional CSP</th>
<th>AIRM CSP</th>
<th>LEM CSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>62.64 ± 12.24 %</td>
<td>65.35 ± 13.80 %</td>
<td>65.63 ± 13.87 %</td>
</tr>
<tr>
<td>3</td>
<td>62.51 ± 13.06 %</td>
<td>65.56 ± 13.71 %</td>
<td>65.00 ± 13.60 %</td>
</tr>
<tr>
<td>4</td>
<td>62.60 ± 13.05 %</td>
<td>65.31 ± 13.96 %</td>
<td>65.11 ± 13.72 %</td>
</tr>
<tr>
<td>5</td>
<td>62.83 ± 12.99 %</td>
<td>65.26 ± 14.02 %</td>
<td>64.88 ± 13.58 %</td>
</tr>
</tbody>
</table>

TABLE II
AVERAGE SUBJECT-INDEPENDENT ACCURACIES FOR SESSION 2 (±
STANDARD DEVIATIONS)

<table>
<thead>
<tr>
<th>n</th>
<th>Conventional CSP</th>
<th>AIRM CSP</th>
<th>LEM CSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>66.25 ± 15.62 %</td>
<td>67.05 ± 15.42 %</td>
<td>67.19 ± 15.18 %</td>
</tr>
<tr>
<td>3</td>
<td>66.63 ± 15.72 %</td>
<td>67.26 ± 15.34 %</td>
<td>67.34 ± 15.08 %</td>
</tr>
<tr>
<td>4</td>
<td>66.05 ± 15.64 %</td>
<td>67.16 ± 14.95 %</td>
<td>67.01 ± 14.73 %</td>
</tr>
<tr>
<td>5</td>
<td>67.76 ± 15.24 %</td>
<td>67.07 ± 14.78 %</td>
<td>66.25 ± 14.92 %</td>
</tr>
</tbody>
</table>

TABLE III
OUTPERFORMANCE P-VALUES OF RIEMANNIAN VS CONVENTIONAL
CSPS, FOR SESSIONS 1 AND 2

<table>
<thead>
<tr>
<th>n</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIRM CSP</td>
<td>LEM CSP</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3</td>
<td>&lt; 0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>5</td>
<td>0.004</td>
<td>0.018</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this work, we compared the effect of using the Fréchet
mean with Riemannian distances with respect to the arith-
metic mean in averaging spatial covariance matrices on the
performance of CSP-based subject-independent classifiers of
motor imagery. Using the largest MI dataset collected to date, we showed that for a common number of CSP filters, using the Riemannian mean generally leads to a higher subject-independent classification accuracy with respect to the arithmetic mean. However, in this study we did not examine the effect of dimensionality (number of channels) on the joint behaviour of Riemannian mean and the classification performance. It has been previously observed that in subject-dependent classification, having a large number of channels is not in favour of Riemannian CSP [25]. Other aspects such as data completion and robustness to corrupted/missing data segments also remain as open problems [26]. We leave the investigation of these effects in the subject-independent context for future studies.

**REFERENCES**


[12] F. Yger, M. Berar, and F. Lotte, “Riemannian Approaches in Brain-


