AVERAGING COVARIANCE MATRICES FOR EEG SIGNAL CLASSIFICATION BASED ON THE CSP: AN EMPIRICAL STUDY

Florian Yger[†]*, Fabien Lotte*[*]*, Masashi Sugiyama*[†]

[†] Dept of Complexity Science and Engineering
Graduate School of Frontier Sciences
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

[*] Inria Bordeaux Sud-Ouest
LaBRI, Potioc team
200 rue de la vieille tour - 33405 Talence, France

ABSTRACT

This paper presents an empirical comparison of covariance matrix averaging methods for EEG signal classification. Indeed, averaging EEG signal covariance matrices is a key step in designing brain-computer interfaces (BCI) based on the popular common spatial pattern (CSP) algorithm. BCI paradigms are typically structured into trials and we argue that this structure should be taken into account. Moreover, the non-Euclidean structure of covariance matrices should be taken into consideration as well. We review several approaches from the literature for averaging covariance matrices in CSP and compare them empirically on three publicly available datasets. Our results show that using Riemannian geometry for averaging covariance matrices improves performances for small dimensional problems, but also the limits of this approach when the dimensionality increases.

***Index Terms*—** common spatial pattern, SPD matrices, robust averaging, Riemannian geometry, EEG signal classification, brain-computer interface (BCI)

1. INTRODUCTION

Brain-computer interface (BCI) systems [1] aim at translating brain signals into control signals, such brain signals being usually processed using machine learning methods. Among the possible paradigms, in recent years, BCI approaches based on motor imagery (MI) have been developing rapidly [2]. In this paradigm, subjects are asked to imagine the movements of their limbs, e.g., hand movements. Different areas of the brain show an alteration in the regular activity according to the imagined movement performed, and this activity can be measured through an electroencephalogram (EEG). The main motivation of BCI is to establish a novel communication channel for healthy and disabled people to interact with the environment. In fact, the information of the mental state of a subject can be used for controlling a computer application or a robotic device such as a wheelchair.

A very challenging task with BCI is to find a reliable representation of brain signals. To do so, a feature extraction method is necessary and, among all, Common Spatial Pattern (CSP) [3, 4] is certainly the most popular for classifying oscillatory EEG signals such as those observed during MI. The idea behind CSP is to compute the most suitable spatial filters to discriminate between different types of EEG signals in a BCI protocol based on changes in oscillations (e.g., motor imagery, steady state visually evoked potential, etc.). Practically, it reduces the volume conduction effect — the spatial spread of information after the electrical signals go through the skull and skin— on the filtered signal.

The core idea of CSP is to simultaneously diagonalize the average covariance matrices (averaged over trials) of the class-related signals. Thus, using poorly estimated or noisy covariance matrices often leads to poor spatial filters, and thus poor BCI performances [5]. Hence, improving covariance matrix estimators should improve CSP performance.

Moreover, covariance matrices are symmetric positive-definite (SPD) and Riemannian geometry has been shown effective for handling such data [6–9]. Links between CSP and Riemannian approaches have even been discussed in [10]. Altogether, several methods are available and theoretically relevant to average covariance matrices. However, it is not known yet which method is the most appropropriate for which EEG classification context (e.g., for which data dimensionality). Therefore, in this study, we review and experimentally compare the pros and cons of various methods for averaging covariance matrices in BCI design.

2. MOTIVATIONS

2.1. Common Spatial Pattern

Based on the pioneering work of Fukunaga et al. in 1970 [3], the CSP is nowadays the most popular algorithm for spatial filtering in motor imagery experiments. The main idea is to use a linear transformation to project the multi-channel EEG

The authors would like to thank the reviewers for their valuable comments. FY was supported by a JSPS fellowship (KAKENHI 26.04730) and MS was supported by KAKENHI 23120004. This work was partially funded by Inria Project Lab BCI-LIFT.

data into a low-dimensional subspace. The aimed transformation maximizes the variance of signals of one class and at the same time minimizes the variance of signals of the other class [1, 4, 11].

Formally, let $X \in R^{N \times C}$ be the data matrix which corresponds to a trial of imaginary movement; N is the number of observations in a trial and C is the number of channels. We want the linear transformation $X_{\mathrm{CSP}} = X \cdot W^T$ where the m spatial filters $w_j \in R^C$, composing the projection matrix $W \in R^{m \times C}$, extremize the following Rayleigh quotient:

$$J(w) = \frac{w^T \Sigma_1 w}{w^T \Sigma_2 w}. \tag{1}$$

$\Sigma_i \in R^{C \times C}$ is the spatial covariance matrix of the band-pass filtered EEG signals from class i. For a given trial matrix X, the empirical covariance estimator is

$$S = \frac{1}{N-1} X^T X. \tag{2}$$

Σ_i, the spatial covariance for class i, is usually computed by averaging the trial covariance matrices as

$$\Sigma_i = \frac{1}{|\varphi_i|} \sum_{j \in \varphi_i} S_j, \tag{3}$$

where φ_i is the set of trials belonging to each class and $|\varphi|$ denotes the cardinality of φ. Notably, such a computation of covariance matrices assumes the EEG signals to have a zero mean (which is true in practice for band-pass filtered signals).

The problem in Eq. (1) can be solved as a generalized eigenvalue problem involving the matrices Σ_1 and Σ_2 and has lead to various extensions and variants [4].

In this article, we do not focus on the CSP algorithm itself but rather on the estimation of Σ from the trial covariance matrices. Indeed, every variant of CSP is based on covariance estimations and we argue that this point is too often underestimated and that the implicit choice of averaging as in Eq. (3) should be considered more carefully.

It should be stated that the maximum likelihood estimator (MLE) of the covariance matrix as shown in Eq. (2) can be very sensitive to outliers. From a sample perspective, having robust estimations of the trial covariance matrices is also likely to improve the CSP performances. Although this point is very important, we are adopting a trial perspective, as also studied in [12], and we try to find a robust way of down-weighting noisy trials in the class-covariance estimation. In this view, we do not try to down-weight or discard individual EEG samples that may be noisy but rather an entire trial (i.e. groups of samples), e.g., artifactual trials.

Using the trial covariance matrices as features, some recent works [6–8] have shown the usefulness of using particular geometries for manipulating this kind of data.

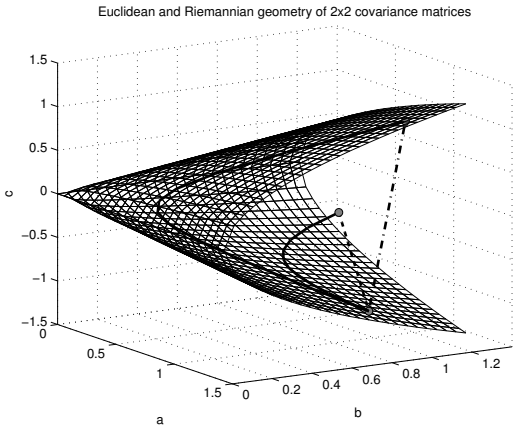

Fig. 1. Comparison between Euclidean (straight dashed lines) and Riemannian (curved solid lines) distances measured between points of the space $\mathcal{P}_2$.

2.2. Different geometries for SPD matrices

SPD matrices —covariance matrices in our case— belong to a Euclidean[1] space (namely the space of symmetric matrices). For example, 2×2 SPD matrix A can be written as $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ with $ac - b^2 > 0$, $a > 0$ and $c > 0$. Then symmetric matrices can be represented as points in $\mathbb{R}^3$ and the constraints can be plotted as a cone, inside which SPD matrices lie strictly (see Fig. 1). A straightforward approach to averaging matrices in this space would be to simply use the Euclidean distance δ_e:

$$\delta_e(A, B) = \|A - B\|_{\mathcal{F}}, \tag{4}$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm. The Euclidean geometry of symmetric matrices implies that distances are computed along straight lines according to δ_e (see Fig. 1 again).

Implicitly, averaging covariance matrices based on Euclidean geometry resorts to the formula in Eq. (3). However, this Euclidean geometry suffers from disavantages. First, as noted in [13], the space of SPD matrices equipped with a Euclidean geometry produces a non-complete space. As illustrated in Fig. 1, extrapolating between two SPD matrices may lead to indefinite matrices. As averaging is just an interpolation problem, this is not a major issue but the so-called swelling effect highlighted in [14] is more problematic. The effect translates the fact that the determinant of the average of two matrices can be bigger than both of their determinants. The implied distortion is then an artifact from the geometry.

To avoid this, we use a more natural metric to compare

[1] Note that, in order to be a Euclidean space, the set if SPD matrices should be equipped with the Frobenius inner product, $\langle A, B \rangle_{\mathcal{F}} = \mathrm{Tr}(A^\top B)$ and the derived norm $||A||_{\mathcal{F}} = \sqrt{\langle A, A \rangle_{\mathcal{F}}}$.

SPD matrices, namely, the *LogEuclidean distance* δ_l:

$$\delta_l(A,B) = \|\log(A) - \log(B)\|_{\mathcal{F}}, \tag{5}$$

where $\log(\cdot)$ stands for the matrix logarithm. The LogEuclidean metric (and the derived distance and kernel) has been used in the literature [6, 14].

Finally, as shown in [15], using the proper Riemannian metric, a distance between two SPD matrices A and B can be computed along curves (namely geodesic). In this Riemannian geometry, the SPD matrics becomes a complete space and the distance between its members is defined as:

$$\delta_r(A,B) = \|\log(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})\|_{\mathcal{F}}. \tag{6}$$

As illustrated in Fig. 1, when using this Riemannian geometry, the space $\mathcal{P}_d$ becomes a complete manifold. As already stated in [14], this distance is immune to the swelling effect. It could be a good candidate for averaging covariance matrices.

3. AVERAGE OF COVARIANCE MATRICES

As discussed in [16] for diffusion tensor imaging, averaging covariance matrices can be tackled from several different ways, depending on the chosen geometry. Independently of the chosen geometry, the problem of averaging a set of objects in a metric space can be expressed as Eq. (7) and it generalizes the well-known least squares principle in Euclidean spaces.

$$\min_{\Sigma} \sum_i \delta^2(S_i, \Sigma). \tag{7}$$

Using the Euclidean distance δ_e (in Eq. (4)), the Euclidean average Σ_E is obtained with the closed-form solution in Eq. (3). On the other hand, as shown in [14], when using the LogEuclidean distance δ_l (as in Eq. (5)), we have the following closed-form solution:

$$\Sigma_L = \exp\left(\sum_i \log(S_i)\right), \tag{8}$$

where $\exp(.)$ denotes the matrix exponential.

However, when the Riemannian distance δ_r is used, there is no closed-form solution for computing the Karcher mean Σ_R and optimization techniques [17,18] are used. In practice, as it was numerically found stable and fast to converge, we use the algorithm proposed in [19].

Even when a suitable distance is used in the Eq. (7), the square in the formula makes the problem sensitive to outliers. To remedy this problem, the square is removed from the formula. Then, computing a median is done by solving

$$\min_{\Sigma} \sum_i \delta(S_i, \Sigma). \tag{9}$$

Owing to the curvature properties of the space, the existence and uniqueness of the median Σ_m for δ_r has been studied in [20] and an iterative algorithm has been proposed.

Using information geometry [21], divergences can also be used to handle structured objects such as covariance matrices. Adopting such a point of view, a trial approach has been used in [12]. From a set of observed covariance matrices S_i, under the assumption that the matrices follow a Wishart distribution with μ degrees of freedom, the robust estimator Σ_d is the matrix minimizing its β-divergence with the observations. As stated in [12], this estimator can be computed by an iterative procedure. Implicitly, this approach compares Wishart distributions and it finds a distribution —in practice, only its parameter— that is the closest to the distributions that most likely generated the observations.

Here, we compare these 5 approaches[2], i.e., the standard MLE averaging (Eq. (3)) minimizing δ_e^2, the LogEuclidean mean minimizing δ_l^2, the Karcher mean [19] minimizing δ_r^2, the Riemannian median minimizing δ_r and the divergence based averaging [12] minimizing β-div.

4. NUMERICAL EXPERIMENTS

4.1. Data description

In order to compare the covariance averaging algorithms, we used EEG data from 17 subjects, from 3 publicly available data sets of BCI competitions, as in [4]. These three datasets contain motor imagery (MI) EEG signals. The first two datasets were collected in a multi-class setting, with the subjects performing more than 2 different MI tasks. For these 2 datasets, we evaluate our algorithms on two-class problems by selecting only signals of left- and right-hand MI trials.

- *BCI competition IV dataset IIa* [22] contains EEG signals (recorded from 22 electrodes) from 9 subjects who performed left-hand, right-hand, foot and tongue MI. A training and a testing sets were available for each subject, both sets containing 72 trials for each class.
- *BCI competition III dataset IIIa* [23] comprises EEG signals (recorded from 60 electrodes) from 3 subjects who performed left-hand, right-hand, foot and tongue MI. A training and a testing sets were available for each subject. Both sets contain 45 trials per class for subject 1, and 30 trials per class for subjects 2 and 3.
- *BCI competition III dataset IVa* [24] consists of EEG signals (recorded from 118 electrodes) from 5 subjects, who performed right hand and foot MI. A training set and a testing set were available for each subject, with different sizes. More precisely, 280 trials were available for each subject, among which 168, 224, 84, 56 and 28 composed

[2] In Table 1, we gather the results of the different methods and refer to them by the cost they minimize.

the training set for subject A1, A2, A3, A4 and A5 respectively, the remaining trials composing their test set.

For all data sets, EEG signals were band-pass filtered in 8-30 Hz, using a 5^{th} order Butterworth filter. For each trial, we extracted features from the time segment located from 0.5s to 2.5s after the cue instructing the subject to perform MI.

4.2. Classifier setup

For the minimum divergence estimator from [12], we set as recommended $\mu = \frac{N}{20}$ and $\beta = 0.001$, respectively the degrees of freedom of the Wishart distributions in the model and the parameter of the β-divergence. The feature extracted from every trial consists of the log-variance EEG signal projected on the 6 selected CSP filters (as recommended in [11]). Then those features are fed to a linear support vector machine (SVM) [25] with C parameter chosen among $[0.1, 1, 10, 100, 1000]$ by a cross-validation procedure (on 30 iterations with $80\% - 20\%$ splits).

4.3. Results and discussion

The first column of Table 1 shows the classification results obtained after various averaging on small dimensional covariance matrices (22 EEG channels). In this case, the Riemannian geometry (either Riemannian mean or median and LogEuclidean mean) seems to have a clear advantage in the mean accuracy (over subjects) ranging from 78.78% up to 79.24% against 76.31% for the Euclidean geometry. This may be due to the fact that the estimated covariance matrices are completely SPD (enough time samples were available for estimating each trial covariance properly) and then the averaging methods behave well. With the default setting, the divergence-based approach also shows a good empirical behaviour. However, it should be noted that the impact of the geometry seems to be subject dependent (as for example the Riemannian geometry seems to particularly suit to Subject 4).

As shown in the two right columns of Table 1, when the dimensionnality of the covariance matrices grows (60 or 118 EEG channels), the situation seems to be less in favour of the Riemannian geometry. Indeed, as the dimension grew, we reach the limit of the SPD assumption of empirical covariance matrices. Then, the Riemannian geometry becomes less efficient[3]. For example, the Euclidean geometry completely outperforms its Riemannian counterparts with a mean accuracy (over subjects) of 79.27% against 74.03% for the best of its competitors. Independently of this, our implementation of the divergence based approach encountered numerical difficulties as the dimension grows. Indeed, the iterative approach described in [12] involves the determinant of a covariance matrix elevated at the power $(\mu - C - 1)\beta$ and some Γ function is numerically ill-behaved.

[3]Note that in the datasets, as the dimensions grows, the number of trials gets smaller and this also affects the performance of the averaging methods.

5. CONCLUSIONS

In this empirical study, we adopted a trial perspective on EEG data and empirically compared several approaches for averaging trials. From our evaluations on three different datasets, Riemannian geometry appears useful for averaging covariance matrices for small dimensional problems. When the dimensionality grows, numerical problems appear and the Euclidean geometry seems to be more suited.

In this study, we left aside the already difficult problem of estimating the trial covariance matrices. In BCI data, when enough time samples are available, robust estimators such as the one proposed in [12] or in [26] (although it has been criticized in [27]) could be investigated as a substitute to the MLE. Otherwise, when the number of time samples is insufficient (as for example in the third experiment) and in case of correlated time samples, the matrices are almost indefinite, shrinkage —a form of regularization— should be used as in [28,29]. Also, the interaction between these estimators and the averaging remains unclear and should be empirically investigated.

In this paper, we studied the different averaging methods for covariance matrices under various scenarii. As a future work, some more numerical experiments should be carried out, by changing the rate of outliers and plotting the CSP patterns, in order to strengthen our analysis of the problem.

We try to evaluate what other geometries could bring to CSP methods but a promising way would be to extract smaller dimensional covariance matrices. As such, approaches like [30] could bridge the gap that currently separate CSP-based approaches and Riemannian approaches in BCI.

REFERENCES

[1] G. Dornhege, J. del R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*, The MIT Press, 2007.

[2] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proc. IEEE*, 2001.

[3] K. Fukunaga and W. Koontz, "Application of the Karhunen-Loève expansion to feature selection and ordering," *IEEE Trans. Comp.*, 1970.

[4] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, 2011.

[5] B. Reuderink and M. Poel, "Robustness of the common spatial patterns algorithm in the BCI-pipeline," Tech. Rep., University of Twente, 2008.

[6] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, 2013.

Table 1. Accuracy obtained on three datasets.

	BCI IV dataset IIa										BCI III dataset IIIa				BCI III dataset IVa					
	1	2	3	4	5	6	7	8	9	mean	1	2	3	mean	1	2	3	4	5	mean
δ_e^2	**81.25**	53.47	95.83	65.97	57.64	67.36	77.78	94.44	**93.06**	76.31	96.67	**61.67**	**98.33**	85.56	**69.64**	**100**	**73.47**	**86.16**	**67.06**	**79.27**
δ_r^2	79.86	**55.56**	95.83	75.69	**61.81**	**72.92**	**82.64**	**97.22**	91.67	**79.24**	**97.78**	**61.67**	**98.33**	**85.93**	67.86	**100**	66.84	69.20	66.27	74.03
δ_r	80.56	54.86	95.83	**77.78**	61.11	73.61	81.94	**97.22**	89.58	79.17	**97.78**	58.33	**98.33**	84.81	67.86	**100**	66.84	67.86	64.68	73.45
δ_l^2	80.56	54.86	**96.53**	74.31	**61.81**	71.53	**82.64**	96.53	90.28	78.78	96.67	55.00	**98.33**	83.33	68.75	96.43	72.45	51.79	53.57	68.60
β-div	**81.25**	52.08	95.83	74.31	58.33	69.44	78.47	95.83	**93.0**6	77.62	-	-	-	-	-	-	-	-	-	-

[7] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *LVA/ICA*, 2010.

[8] F. Yger, "A review of kernels on covariance matrices for BCI applications," in *Proc. IEEE MLSP*, 2013.

[9] F. Yger and M. Sugiyama, "Supervised LogEuclidean metric learning for symmetric positive definite matrices," *preprint arXiv:1502.03505*, 2015.

[10] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Common spatial pattern revisited by Riemannian geometry," in *Proc. IEEE MMSP*, 2010.

[11] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Sig. Proc. Mag.*, 2008.

[12] W. Samek and M. Kawanabe, "Robust common spatial patterns by minimum divergence covariance estimator," in *Proc. ICASSP*, 2014.

[13] T. Fletcher and S. Joshi, "Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors," in *Proc. CVAMIA and MMBIA*. 2004.

[14] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, 2007.

[15] R. Bhatia, *Positive Definite Matrices*, Princeton University Press, 2009.

[16] I. Dryden, A. Koloydenko, and D. Zhou, "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *The Annals of Applied Statistics*, 2009.

[17] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM J. Mat. An. App.*, 2005.

[18] B. Jeuris, R. Vandebril, and B. Vandereycken, "A survey and comparison of contemporary algorithms for computing the matrix geometric mean," *Elec. Trans. Num. An.*, 2012.

[19] D. Bini and B. Iannazzo, "Computing the Karcher mean of symmetric positive definite matrices," *Linear Algebra and its Applications*, 2013.

[20] T. Fletcher, S. Venkatasubramanian, and S. Joshi, "The geometric median on Riemannian manifolds with application to robust atlas estimation," *NeuroImage*, 2009.

[21] S.-I. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 2010.

[22] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, "Seperability of four-class motor imagery data using independent components analysis," *J. Neur. Eng.*, 2006.

[23] A. Schlögl, F. Lee, H. Bischof, and G. Pfurtscheller, "Characterization of four-class motor imagery EEG data for the BCI-competition 2005," *J. Neur. Eng.*, 2005.

[24] B. Blankertz, K. Muller, D. Krusienski, G. Schalk, J. Wolpaw, A. Schlogl, G. Pfurtscheller, J. Millan, M Schroder, and N. Birbaumer, "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Trans. Neur. Syst. Rehab.*, 2006.

[25] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[26] P. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 1999.

[27] D. Olive, "Why the Rousseeuw Yohai paradigm is one of the largest and longest running scientific hoaxes in history," Tech. Rep., Southern Illinois University, 2012.

[28] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. An.*, 2004.

[29] D. Bartz and K.-R. Müller, "Covariance shrinkage for autocorrelated data," in *Proc. NIPS*, 2014.

[30] M. Harandi, M. Salzmann, and R. Hartley, "From manifold to manifold: geometry-aware dimensionality reduction for SPD matrices," in *Proc. ECCV*, 2014.