Distance Covariance: A Nonlinear Extension of Riemannian Geometry for EEG-based Brain-Computer Interfacing

Jiachen Xu[1,*], Alex Markham[1], Anja Meunier[1], Philipp Raggam[1,2] and Moritz Grosse-Wentrup[1,3,4]

***Abstract*— Riemannian frameworks are the basis for some of the best-performing decoding methods in EEG-based Brain-Computer Interfacing. In this work, we consider whether a nonlinear extension of the Riemannian framework, obtained by replacing the channel-wise covariance of the EEG signal with the nonlinear distance covariance, improves decoding performance. We study the theoretical properties of the distance covariance metric in this framework, in particular invariance to affine transformations, and compare the proposed method with established Riemannian methods on three different EEG data sets. We do not find evidence that the distance covariance extension improves decoding performance in comparison to the linear Riemannian framework.**

I. INTRODUCTION

Brain-Computer Interfaces (BCIs) decode cognitive states from recordings of brain imaging data, e.g., for communication with severely paralyzed patients [1], [2], to induce neural plasticity in stroke rehabilitation [3], or to passively monitor and react to their users' mental state [4], [5]. For BCIs based on non-invasive recordings of the electroencephalogram (EEG), the decoding pipeline typically consists of three components. In a first step, the EEG channels are linearly combined to attenuate irrelevant EEG sources and retain those that are most informative for a given paradigm and subject [6]–[8]. In a second step, informative features are estimated from the spatially filtered timeseries, e.g., by computing log-bandpowers in canonical EEG frequency bands. In the final decoding step, these features are then processed by a classification algorithm [9].

Among the wide variety of available BCI decoding pipelines, Riemannian methods have evolved as one of the best performing frameworks [10]–[12]. Riemannian frameworks combine the spatial filtering and feature extraction steps of a decoding pipeline by representing each trial of recorded EEG data by a channel-wise covariance matrix. Distances between covariance matrices of different trials are then measured on the Riemannian manifold of symmetric positive definite matrices (SPDs). In this way, the Riemannian frameworks measure similarity between EEG trials not only by the channel-wise power of individual EEG channels but also by the spatial dependence structure between channels. These similarity estimates then form the feature space for the final classification step, for which standard machine learning algorithms, e.g., support vector machines (SVMs), can be employed.

It has been a long-standing discussion in the BCI community whether linear classifiers are sufficient for EEG decoding, with the general consensus that nonlinear methods may provide small benefits in certain settings [13]. Nonlinear methods, however, have mostly been explored in the final classification step, e.g., by kernelizing SVMs [9]. Spatial filtering and feature extraction algorithms, on the other hand, are predominantly based on linear methods [6], with recent deep learning methods forming a notable exception [14]. To study whether nonlinear extensions of spatial filtering methods may improve decoding performance, we replaced the linear covariance metric in the Riemannian framework with the nonlinear distance covariance metric [15]. The distance covariance is a generalization of the covariance that provides a measure of association similar to the covariance in case of bivariate normal data but is further able to represent nonlinear associations. As such, we hypothesized that an extension of Riemannian frameworks by the distance covariance would perform at least as well as standard, linear Riemannian frameworks while offering the potential to further leverage nonlinear relationships, if they should be present in the data.

After briefly introducing Riemannian frameworks and the distance covariance in Sections II-A and II-B, we present the nonlinear extension of the Riemannian framework in Section II-C. We then study its theoretical properties in Section II-D, where we show that the desirable property of invariance to affine transformations only generalizes in part to the nonlinear case. We then describe in Section II-E how we compare the linear and nonlinear frameworks on three publicly available EEG data sets, and present our results in Section III. We find that the nonlinear Riemannian framework performs on par with the linear one but does not substantially outperform it. We conclude this work by discussing the implications of our findings in Section IV.

II. METHODS

A. Riemannian Methods in BCI

For classification in BCIs, short-time epochs of the EEG signal, also called trials, are usually used to generate the feature space for machine learning algorithms. A trial is represented by the matrix $\mathbf{X}_i \in \mathbb{R}^{N \times T}$, where N denotes the number of channels and T the number of sampled time points. In this context, Riemannian geometry-based methods make use of the covariance matrix $\Sigma_i \in \mathbb{R}^{N \times N}$

[1]Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna, Austria
[2]Department of Neurology & Stroke, Hertie Institute for Clinical Brain Research, University of Tübingen, Germany
[3]Research Platform Data Science @ Uni Vienna
[4]Vienna Cognitive Science Hub
*Email: jiachen.xu@univie.ac.at

[10], estimated by the sample covariance matrix (SCM)

$$\hat{\Sigma}_i = \frac{1}{T-1}\mathbf{X}_i\mathbf{X}_i^\top. \tag{1}$$

Covariance matrices do not live in Euclidean space but rather in the smooth and differentiable Riemannian manifold $\mathcal{M}$ of Symmetric Positive Definite (SPD) matrices [16].

Trials can be compared with each other by calculating the distance between their SCMs, which are represented by points in the manifold [17]. The distance between two points can be calculated using the affine-invariant Riemannian metric (AIRM), which generalizes the properties of the Euclidean metric in standard vector spaces to the Riemannian manifold [18]. The distance between two covariance matrices (Σ_1, Σ_2) is thus defined as

$$\mathrm{d}_{\mathrm{AIRM}}(\Sigma_1, \Sigma_2) = \left\| \log\left(\lambda\left(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\right)\right\|_2, \tag{2}$$

where $\lambda(\cdot)$ represents the eigenvalues of its argument and $\|\cdot\|_2$ the L2 norm. This Riemannian distance can be used for classification with the Minimum Distance to Mean (MDM) procedure [10]. For each class $k \in \{1, \ldots, K\}$, the Riemannian mean of its I_k intraclass SCMs $(\Sigma_i^k)_{i\in\{1,\ldots,I_k\}}$ is calculated by

$$\overline{\Sigma}^k = \underset{\Sigma\in\mathcal{M}}{\arg\min} \sum_{i=1}^{I_k} \mathrm{d}^2_{\mathrm{AIRM}}(\Sigma, \Sigma_i^k). \tag{3}$$

For each new trial $\mathbf{X_j}$, the Riemannian distance between the SCM of the trial Σ_j and the Riemannian mean of each class $\overline{\Sigma}^k$ is calculated. The new trial is then assigned to the class with the minimum distance according to

$$\hat{k} = \underset{k\in\{1,\ldots,K\}}{\arg\min} \left\{\mathrm{d}_{\mathrm{AIRM}}\left(\Sigma_j, \overline{\Sigma}^k\right)\right\}. \tag{4}$$

Alternatively, a tangent space mapping can be employed to obtain standard feature vectors and combine Riemannian methods with traditional classification algorithms [19]. The mapping of Σ_j from the Riemannian manifold to the tangent space with Fréchet mean reference $\overline{\Sigma}$ is done via matrix logarithmic mapping:

$$\mathrm{Logm}_{\overline{\Sigma}}(\Sigma_j) = \overline{\Sigma}^{\frac{1}{2}} \log\left(\overline{\Sigma}^{-\frac{1}{2}}\Sigma_j\overline{\Sigma}^{-\frac{1}{2}}\right)\overline{\Sigma}^{\frac{1}{2}}. \tag{5}$$

B. Distance Covariance

The *distance covariance* [15], dCov, can be thought of as a nonlinear extension of the more familiar product-moment covariance, cov, which is only a measure of linear association. Before we define dCov, first recall that the population covariance of random variables X and Y can be given in terms of expected values as

$$\mathrm{cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{6}$$

The distance covariance between random variables X and Y is defined as

$$\begin{aligned}\mathrm{dCov}^2(X, Y) = \mathrm{cov}(|X - X'|, |Y - Y'|) \\ - 2\mathrm{cov}(|X - X'|, |Y - Y''|),\end{aligned} \tag{7}$$

where (X', Y') is an independent and identically distributed (iid) copy of (X, Y), and Y'' is an iid copy of Y. Note that it can be defined more generally, as in [15], however we give a restricted definition here so that it is more obviously analogous to cov and applicable to Riemannian methods. The key intuition here is that it makes use of distances, such as $|X - X'|$: this can be thought of as nonlinearly projecting X and Y into a kernel space so that defining dCov using the linear product-moment covariance (cov) in the kernel space nevertheless results in a nonlinear measure of dependence in the original space—indeed, [20] proves that dCov is equivalent to kernel-based dependence measures such as the Hilbert-Schmidt Inpendence Criterion (HSIC) [21].

Analogous to (1), the distance covariance matrix, $\Delta_{\mathbf{X}}$, for an N-dimensional random vector can be estimated from the observations $\mathbf{X}_i \in \mathbb{R}^{N\times T}$ of each trial. First, use the sample $\mathbf{X}_i$ to construct the flattened distances matrix F: for each row $\mathbf{x}_i^l = (x_1, \ldots, x_n)$ in $\mathbf{X}_i$, the corresponding uncentered distance matrix is defined element-wise to be $U_{p,q} := |x_p - x_q|$; then U must be doubly-centered, $C_{p,q} := U_{p,q} - U_{p,\cdot} - U_{\cdot,q} + U_{\cdot,\cdot}$, where replacing an index p or q with $\cdot$ denotes taking the mean over that index; lastly, form the flattened distances matrix F so that each row F^l is the corresponding vectorized (flattened) doubly-centered matrix C. We can finally define the distance covariance matrix:

$$\hat{\Delta}_{\mathbf{X}} = \frac{1}{N^2}FF^\top. \tag{8}$$

The construction in (8) makes it clear that distance covariance matrix, like the covariance matrix, is the product of a matrix and its transpose and is thus SPD if and only if F is of full row rank. However, as we discuss in Section II-D, the distance covariance matrix does not have the same affine invariance properties of the covariance matrix.

C. Extending Riemannian Decoding with the Distance Covariance

Because the distance covariance gives rise to SPD matrices, we can simply replace the covariance matrices in equations (2) to (5) with the appropriate distance covariance matrices. In practice, we can mitigate non-SPD cases, which are induced by numerical errors, by adding regularization. We thereby obtain a nonlinear generalization of the Riemannian frameworks. The distance covariance provides a measure of association similar to the square of the covariance in case of bivariate normal data but is further able to represent nonlinear associations. In this way, the distcov-Riemannian framework is able to assess nonlinear spatial dependence structure between EEG channels.

It has been argued that invariance under affine transformations is one of the reasons for the superior decoding performance achieved by Riemannian methods [11]. In the following section, we study whether this property also applies to its nonlinear generalization.

D. Invariance Under Affine Transformation

Traditional Riemannian feature extraction based on the covariance matrix has the desirable property of being invariant to affine transformations of the underlying data. This immediately follows from properties of the covariance matrix

$$\Sigma_{A\mathbf{X}+b} = A\Sigma_{\mathbf{X}}A^\top \tag{9}$$

for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$, and the affine invariant Riemannian distance [18]

$$\mathrm{d_{AIRM}}(A\Sigma_1 A^\top, A\Sigma_2 A^\top) = \mathrm{d_{AIRM}}(\Sigma_1, \Sigma_2) \tag{10}$$

for $A \in \mathrm{GL}_n(\mathbb{R})$, the group of n-dimensional invertible matrices. In the context of BCI, this means that the features are invariant to scaling (e.g. changes in channel connectivity) and remixing of the sources (e.g. shifts in electrode placement). Unfortunately, this property cannot be simply generalized to the nonlinear case, since (9) does not generally hold for the distance covariance matrix. However, weaker invariance properties can be shown:
For $A \in \mathrm{GL}_n(\mathbb{R}), b \in \mathbb{R}^n$, we have

$$\begin{aligned}(\Delta_{A\mathbf{X}+b})_{rs} &= \mathrm{dCov}^2(A_r\mathbf{X}+b_r, A_s\mathbf{X}+b_s)\\ &= \mathrm{E}[|A_r(\mathbf{X}-\mathbf{X}')|\cdot|A_s(\mathbf{X}-\mathbf{X}')|]\\ &+ \mathrm{E}[|A_r(\mathbf{X}-\mathbf{X}')|]\cdot\mathrm{E}[|A_s(\mathbf{X}-\mathbf{X}')|]\\ &- 2\mathrm{E}[|A_r(\mathbf{X}-\mathbf{X}')|\cdot|A_s(\mathbf{X}-\mathbf{X}'')|],\end{aligned}$$

where $A_r = (a_{r1}, ..., a_{rn})$ is the r-th row vector of A. Furthermore, we have

$$|A_r(\mathbf{X}-\mathbf{X}')| = |\sum_{j=0}^{n} a_{rs}(X_s - X'_s)| \le \sum_{j=0}^{n}|a_{rs}(X_s - X'_s)|. \tag{11}$$

Equality holds if $a_{rs}(X_s - X'_s) \ge 0$ or all $a_{rs}(X_s - X'_s) \le 0$ for all $s \in \{0, ..., n\}$. This can only be fulfilled if for all rows A_r of A, $a_{rs} \neq 0$ for only one $s \in \{1, ..., n\}$. In this case there exists a permutation $\pi : \{1, ..., n\} \to \{1, ..., n\}$ and a corresponding permutation matrix P, as well as a diagonal matrix D, such that $A = PD$. Then it follows with

$$a_{rs} = P_r D_s = \begin{cases} d_{\pi(r)\pi(r)} & \text{if } \pi(r) = s \\ 0 & \text{otherwise}\end{cases} \tag{12}$$

and

$$|A_r(\mathbf{X}-\mathbf{X}')| = |d_{\pi(r)\pi(r)}|\cdot|X_{\pi(r)} - X'_{\pi(r)}| \tag{13}$$

that

$$\begin{aligned}(\Delta_{PD\mathbf{X}+b})_{rs} &= |d_{\pi(r)\pi(r)}d_{\pi(s)\pi(s)}|\\ &\cdot\Big(\mathrm{E}[|X_{\pi(r)} - X'_{\pi(r)}|\cdot|X_{\pi(s)} - X'_{\pi(s)}|]\\ &+ \mathrm{E}[|X_{\pi(r)} - X'_{\pi(r)}|]\cdot\mathrm{E}[|X_{\pi(s)} - X'_{\pi(s)}|]\\ &- 2\mathrm{E}[|X_{\pi(r)} - X'_{\pi(r)}|\cdot|X_{\pi(s)} - X''_{\pi(s)}|]\Big)\\ &= |d_{\pi(r)\pi(r)}d_{\pi(s)\pi(s)}|\cdot(\Delta_{PD\mathbf{X}+b})_{\pi(r)\pi(s)}\end{aligned}$$

and therefore

$$\Delta_{PD\mathbf{X}+b} = PD\Delta_{\mathbf{X}}(PD)^\top. \tag{14}$$

We have thus shown that (9) generalizes to the distance covariance matrix for permutations and scaling of the channels, and that therefore the proposed nonlinear Riemannian feature extraction is invariant to such transformations, but not to remixing of the sources.

E. Evaluation Pipeline

After justifying the theoretical legitimacy of nonlinearly extending Riemannian methods by utilizing distance covariance matrices, we validate their practical performance against sample covariance matrices with real-world EEG data. Considering the well-understood neurophysiological mechanism behind the motor imagery (MI) paradigm [22], we adopt the left- versus right-hand motor imagery task as their cortical activities are highly discriminative [23]. Subsequently, we chose three datasets with the different number of electrodes (listed in TABLE I), and the analysis is restricted to the α and β band (8-32 Hz) since they are the most relevant frequency bands to the MI task. For convenience, we base our evaluation setup on the open-source benchmark, Mother of all BCI Benchmark (MOABB) [24]. After epoching the data based on individual task intervals, we estimate the trial-wise distance- and sample covariance matrices according to procedures in Sections II-A and II-B.

TABLE I: Overview of adopted datasets

Dataset Name	#Channels	#Subjects	#Sessions	Citations
BNCI 2014-001	22	9	2	[25]
Cho	64	49	1	[26]
Munich MI	128	10	1	[7]

In the Riemannian framework, we classify the feature matrices via either measuring the geodesic distance between them directly on the manifold, which is known as minimum distance to the mean (MDM) algorithm, or by projecting them to the tangent space and then applying classifiers to the tangent vectors. We adopt both ways to investigate the benefit of bringing in nonlinearity. Further, we choose two types of Riemannian metrics, the affine invariant Riemannian metric (AIRM) and the Log-Euclidean metric (LogE), for the tangent space based classification. The Log-Euclidean metric can also be interpreted as the simplified version of AIRM because the reference point under the LogE metric is an identity matrix rather than the Fréchet mean point. By comparing both, we can explore the influence of affine-invariance.

As for the choice of the classifier on the tangent space, we choose support vector machines (SVM) with both linear and nonlinear kernels, including the polynomial and Gaussian kernel. By comparing their performances, we can study the effect of the nonlinearity introduced by the classifier. There are several hyperparameters to be set in the kernel SVM. For the regularization parameter (L2 penalty) used by all three SVM classifiers, we use a grid search to find the optimal value within the range from 0.01 to 100. For the parameter γ in both nonlinear kernels, we set it to $\frac{1}{l_{\text{ts-vec}} * \mathrm{Var}(X)}$, where $l_{\text{ts-vec}}$ is the length of the tangent vector and $\mathrm{Var}(X)$ is

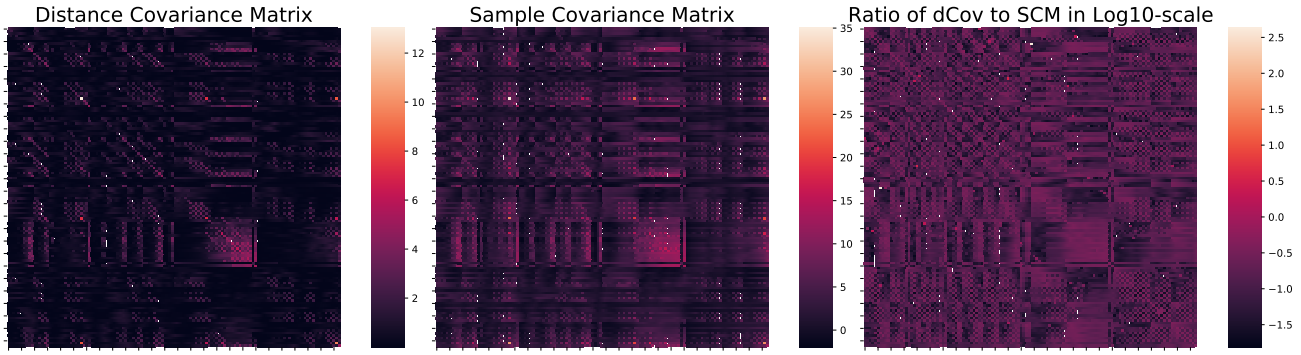


Fig. 1: Matrix structure comparison between distance covariance and sample covariance. These matrices are generated from one randomly selected trial in the Munich data set. The left one shows the structure of the distance covariance matrix computed according to Section II-B, and the middle one is the regular sample covariance matrix. The right one shows the elementwise absolute ratio of the distance covariance matrix (dCov) to sample covariance matrix (SCM) in log10 scale.

the variance of these tangent vectors. Regarding the degree of the polynomial kernel, we fix it to three to achieve a reasonable trade-off between model complexity and the risk of overfitting.

For scoring the predictions made by the classifiers (incl. MDM), we choose the ROC-AUC (receiver operating characteristic-area under the curve) metric. These scores are computed via five-fold cross-validation within each session and subject. To further detect the significant difference among these scores, we adopt the statistical test tools from MOABB [24].

III. RESULTS

Before going into the details of the decoding analysis, we first visualize the matrix structure of both distance and sample covariance matrix, as shown in the left and middle sub-figures in Fig. 1. By comparing these two figures, we can get some intuition about their similarities. These covariance matrices are computed from one randomly selected trial in the Munich data set, which has the largest number of channels. Comparing the left and middle sub-figures, notice that they share a similar blockwise structure, though there does not seem to exist a universal scaling factor between them. To compare the positive-valued distance covariance matrix with the real valued covariance matrix we show the elementwise ratio of their absolute values in the right sub-figure, for better visualization in log10-scale. As seen in this sub-figure, most of the pixels are purple, which means their log-ratios are reasonably close to 0.0. However, we can also find some yellow pixels and black blocks in the same sub-figure, which means the distance covariance values differ substantially from the sample covariance. For instance, the yellow pixel is with a value larger than 2.5, which means their ratio is larger than $10^{2.5} \approx 316$.

The next interesting question is whether these differences can bring additional advantages regarding the feature separability. Therefore, we continue with a classification accuracy comparison between distance and sample covariance matrices as input features in the Riemannian framework. One proper Riemannian manifold-based method to investigate the raw separability of the input features is the minimum distance to the mean (MDM) classifier because it simply measures and compares the distance between the test point and the mean point of each class. The MDM classifier can be understood as the nearest neighbor algorithm using the AIRM based Riemmanian distance instead of the Euclidean distance. As shown in Fig. 2, the classification accuracy of both types of covariance matrices are highly tied with each other, and no significant differences can be found in this comparison. Even if no significance is found, an interesting tendency can be noticed that the difference between the median classification accuracy of both pipelines seems to be negatively correlated with the number of channels, i.e., the size of input features.

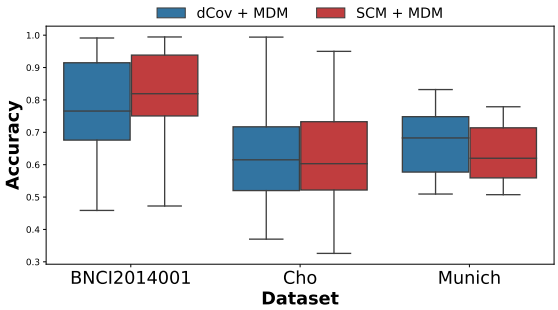
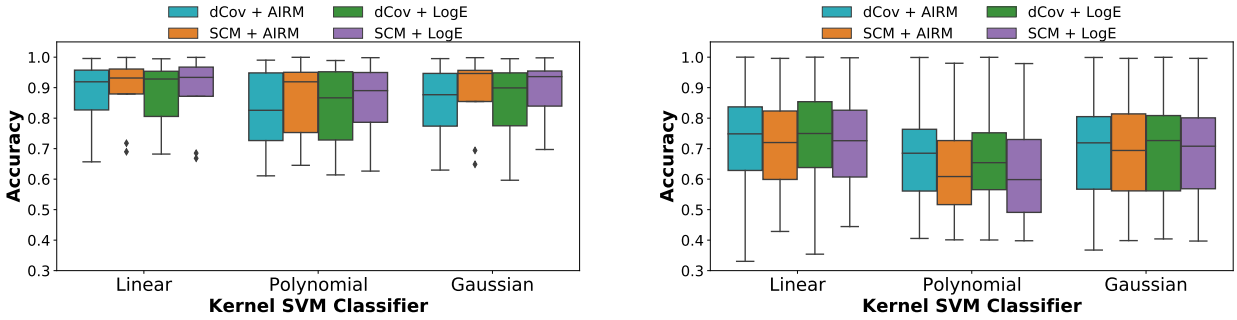


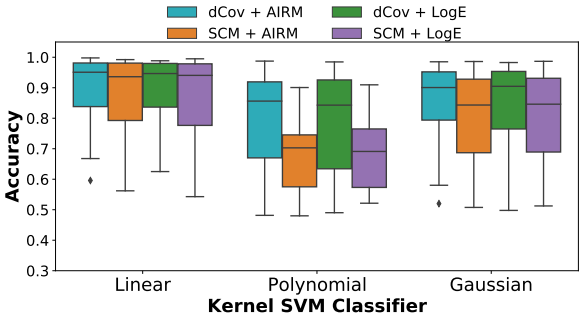
Fig. 2: Within session classification accuracy based on the minimum distance to the mean (MDM) classifier for three motor imagery (MI) datasets with the different number of channels (22, 64, 128 for BNCI2014001, Cho, and Munich, respectively). The blue color indicates the accuracy of using distance covariance matrices (dCov) as feature matrices, and the red color represents the results of using the sample covariance matrices (SCM).

This observation further questions whether the differences between the distance and sample covariance matrices can be better captured by leveraging more advanced classifiers. Therefore, we decide to classify the features on the tangent



(a) Data Set: BNCI2014001 (#Channels=22)

(b) Data Set: Cho (#Channels=64)



(c) Data Set: Munich (#Channels=128)

Fig. 3: Within session tangent space based classification accuracy with different kernel SVM classifiers for different datasets. X-axis represent the adopted classifier and color indicate the various combinations of Riemannian metric and feature matrices. E.g., cyan and orange show the accuracy of distance covariance matrices (dCov) and sample covariance matrices (SCM) based on affine-invariant Riemannian metric (AIRM) while yellow and purple represent using Log-Euclidean metric (LogE).

space, which can bring more flexibility to the choices of classifiers. To project both types of matrices from manifold to the tangent space, we choose two types of Riemannian metric, i.e., affine-invariant Riemannian metric (AIRM) and Log-Euclidean (LogE) metric, to explore the influence of preserving the property of affine-invariance. Further, considering the features in the high dimensional space are not necessary to be linearly separable, we choose the kernel support vector machine (SVM) as the classifiers. By using different kernels, we can investigate the influence of the nonlinearity introduced by the classifier. The results are depicted in Fig. 3, in which each sub-figure shows the classification accuracy for each data set and different color represents the accuracy of using various pipelines to classify.

We first look at the influence of different classifiers. Obviously, the classification accuracy within each data set mostly depends on its data quality, and the choice of classifier has a limited effect on the results. For instance, the linear kernel always performs the best, and the Gaussian kernel is a bit superior to the polynomial kernel. However, suppose we only focus on one specific classifier's results and compare the differences between distance and sample covariance matrices (cyan vs. orange and green vs. purple). In that case, the aforementioned negative correlation between median accuracy difference (distance covariance minus sample covariance) and the number of channels still seems to exist. In particular, in the data set Munich, which has the largest number of channels, an evident classification accuracy difference between distance and sample covariance matrices can be noticed when using a polynomial kernel. Next, by comparing the accuracy between AIRM and LogE based pipelines (cyan vs. green and orange vs. purple), their performances are so competitive that the median accuracy is almost the same.

IV. DISCUSSION

To discuss whether the Riemannian framework can benefit from the nonlinear generalization via the distance covariance, we need to consider three aspects: feature matrix, projection, and classifier. In the comparison in Fig. 2, we see that the nonlinearity does not lead to significant differences in decoding accuracies. One possible justification is that most elements roughly have the same value, which means that most of the information encoded within two types of covariance matrices remains the same. Thus, we cannot find any significant differences in the feature separability.

Second, we notice a negative correlation between the median accuracy difference and the number of channels. One possible explanation is that the total number of the elements with extremely different value has a quadratic growth when the number of channel increase. Furthermore, this change probably further results in the small differences in feature separability. However, the currently available datasets cannot support any significant conclusions about this tendency.

When comparing the results of manifold-based classification (Fig. 2) with tangent space based classification with affine-invariant Riemannian metric (AIRM) (Fig. 3), note that Riemannian methods take advantage of nonlinear projections. Both of them utilize the same Riemannian metric (AIRM) to measure the distance between matrices, while the latter outperforms the former and contains a nonlinear projection.

Next, as shown in Fig. 3, it seems that the classification performance cannot benefit from the nonlinearity of the classifier because the linear one outperforms both nonlinear classifiers regardless of the feature matrices. This finding indicates that the features after nonlinear projection seems to be linearly separable. However, the accuracy difference when using polynomial kernel also suggests the possibility that a linear classifier cannot capture the difference between distance and sample covariance matrices, while nonlinear classifiers seem to be capable of doing it.

Finally, we would like to discuss the importance of affine-invariance from the perspective of feature separability. First, in the Riemannian metric comparison, i.e., AIRM vs. LogE, we can barely find any difference among them. Second, as proved in Section II-D, even if distance covariance matrices only satisfy a weak affine-invariance, it is hard to find any loss of feature separability in comparison to the sample covariance matrices. Therefore, based on both comparisons, it seems that the affine-invariance is not as relevant for decoding performance as previously argued.

ACKNOWLEDGMENT

This project has received funding from the German Federal Ministry of Education and Research (BMBF, grant agreement No 13GW0213A), to P.R.

REFERENCES

[1] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, 1999.

[2] T. Fomina, G. Lohmann, M. Erb, T. Ethofer, B. Schölkopf, and M. Grosse-Wentrup, "Self-regulation of brain rhythms in the precuneus: A novel BCI paradigm for patients with ALS," *Journal of Neural Engineering*, vol. 13, no. 6, 2016.

[3] M. Grosse-Wentrup, D. Mattia, and K. Oweiss, "Using brain–computer interfaces to induce neural plasticity and restore function," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025004, 2011.

[4] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025005, 2011.

[5] T. O. Zander, L. R. Krol, N. P. Birbaumer, and K. Gramann, "Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity," *Proceedings of the National Academy of Sciences*, vol. 113, no. 52, pp. 14898–14903, 2016.

[8] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, 2008.

[6] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[7] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1209–1219, apr 2009.

[9] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, 2007.

[10] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2011.

[11] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[12] J. Xu, M. Grosse-Wentrup, and V. Jayaram, "Tangent space spatial filters for interpretable and efficient Riemannian classification," *Journal of Neural Engineering*, vol. 17, no. 2, p. 026043, 2020.

[13] K. Müller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, 2003.

[14] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, Dec 2007.

[16] R. Bhatia, *Positive definite matrices*. Princeton university press, 2009.

[17] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Geodesy-the Challenge of the 3rd Millennium*, pp. 299–309, Springer, 2003.

[18] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[19] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Common spatial pattern revisited by Riemannian geometry," in *2010 IEEE International Workshop on Multimedia Signal Processing*, pp. 472–476, IEEE, 2010.

[20] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, pp. 2263–2291, 2013.

[21] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *International Conference on Algorithmic Learning Theory*, pp. 63–77, Springer, 2005.

[22] G. Pfurtscheller and F. L. Da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[23] G. Pfurtscheller and C. Neuper, "Motor imagery activates primary sensorimotor area in humans," *Neuroscience Letters*, vol. 239, no. 2-3, pp. 65–68, 1997.

[24] V. Jayaram and A. Barachant, "MOABB: trustworthy algorithm benchmarking for BCIs," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066011, 2018.

[25] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz, *et al.*, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012.

[26] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain–computer interface," *GigaScience*, vol. 6, no. 7, p. gix034, 2017.