t-WDA: A novel Discriminant Analysis applied to EEG classification

Imen Ayadi, Florent Bouchard, Frédéric Pascal

t-$\mathcal{W}$DA: A novel Discriminant Analysis applied to EEG classification

Imen Ayadi, Florent Bouchard, Frédéric Pascal
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France

***Abstract*—This paper provides a new classification method of covariance matrices exploiting the t-Wishart distribution, which generalizes the Wishart distribution. Compared to the Wishart distribution, it is more robust to aberrant covariance matrices and more flexible to distribution mismatch. Following recent developments on this matrix-variate distribution, the proposed classifier is obtained by leveraging the Discriminant Analysis framework and providing original decision rules. The practical interest of our approach is shown thanks to numerical experiments on real data. More precisely, the proposed classifier yields the best results on two standard electroencephalography datasets compared to the best state-of-the-art minimum distance-to-mean (MDM) classifiers.**

***Index Terms*—EEG, Covariance matrices, t-Wishart, Bayseian classification, Discriminant Analysis, BCI.**

I. Introduction

In signal processing, covariance matrices have recently gained interest in classification problems. Not only do they embed relevant information from signals, but they also fall into the domain of Riemannian geometry. They have proved their merit in radar and image processing [1], biomedical signals analysis, *etc*. For instance, their use has revolutionized the field of Electroencephalography (EEG), which consists in recording brain signals for medical purposes or to help people with motor impairment via brain-computer interfaces (BCI) [2], [3]. The first generation of EEG classification methods uses spatial filters paired with Euclidean classifiers [2]. Such an approach has been outperformed for a decade using spatial covariance matrices as features [4]. There are two main methods for classifying covariance matrices either directly on their native space of symmetric positive definite (SPD) matrices via the Minimum Distance to Mean (MDM) or by projecting them onto a tangent space at a reference SPD matrix and then using a Euclidean classifier for the projected matrices [4].

In order to face the challenges of the intrinsic non-Gaussianity of signals, covariance matrices are often estimated thanks to robust techniques [5] before proceeding to MDM or tangent space projection methods. However, this trick needs to be more efficient in the case of aberrant recordings or mislabelling. Such situations, which often occur when manipulating signals, explain the motivation to consider the outliers at the scale of the covariance matrices and not only at the scale of a single measurement. Accordingly, it is interesting to design probabilistic classifiers that exploit statistics over SPD [6], particularly heavy-tailed matrix-variate distributions. We would mention that such a strategy was already implemented on matrices (not specifically SPD) [7] where an Expectation-Maximization algorithm is suggested using matrix-variate t-distributions. Among the distributions handling outliers over SPD matrices, figure the t-Wishart ones [8], [9]. They are known to generalize the most classical distribution on SPD matrices, the Wishart distribution [10]. The latter is the distribution of sample covariance matrices of random vectors drawn from a multivariate Gaussian distribution. The generalization brought by the t-Wishart distributions can be similarly interpreted as the extension to the multivariate Gaussian distribution brought by the multivariate t-distributions [11].

The main contribution of this paper is the development of a new Bayesian classifier, t-$\mathcal{W}$DA (t-Wishart Discriminant Analysis), that uses the t-Wishart distribution for the likelihood of observed data. Moreover, the Discriminant Analysis framework offers a novel interpretation of the most classical versions of MDM. A particular focus is drawn on the merit of t-$\mathcal{W}$DA for EEG classification.

The paper is organized as follows. Section II provides a brief background about EEG classification, mainly on the methods based on covariance matrices as features. Section III reviews the t-Wishart distributions. Then, Section IV derives the t-Wishart Discriminant Analysis classifier. Some numerical experiments in Section V validate its merit for BCI. Finally, concluding remarks and perspectives are drawn in Section VI.

II. Background on EEG classification: MDM classifier

As the introduction mentions, various methods have been considered to classify EEG signals [2]. The most efficient ones usually rely on covariance matrices. In particular, the reference method in the BCI community is the MDM classifier [4]. This section introduces this EEG classification method, which will be used as a baseline in this work.

The MDM classifier can be seen as a supervised k-means classification method: it aims to minimize the dispersion inside classes according to a given distance or divergence. In EEG, signals $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ (p, number of electrodes; n number of samples) are supposed centered, *i.e.*, $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$.

"This work is partially supported by a public grant overseen by the French National Research Agency (ANR) through the program UDOPIA, project funded by the ANR-20-THIA-0013-01".

The covariance matrix of $\boldsymbol{X}$ is estimated with the sample covariance matrix (SCM), defined as

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T. \tag{1}$$

By construction, $\boldsymbol{C} \in \mathcal{S}_p^{++}$, the set of $p \times p$ SPD matrices. In the training step, a barycenter $\hat{\boldsymbol{\Sigma}}_k$ is computed for each class k. Given covariance matrices $\{\boldsymbol{C}_i\}_{i=1}^{N_k}$ in class k, it is

$$\hat{\boldsymbol{\Sigma}}_k = \underset{\boldsymbol{\Sigma}\in\mathcal{S}_p^{++}}{\operatorname{argmin}} \sum_{i=1}^{N_k} d(\boldsymbol{C}_i, \boldsymbol{\Sigma}), \tag{2}$$

where d is a divergence or the square of a distance on $\mathcal{S}_p^{++}$. In the testing step , a covariance matrix $\boldsymbol{C}$ is assigned to the class y whose barycenter $\hat{\boldsymbol{\Sigma}}_y$ is the closest. Formally, the decision rule is:

$$y = \underset{k\in[\![1,K]\!]}{\operatorname{argmin}}\, d(\boldsymbol{C}, \hat{\boldsymbol{\Sigma}}_k). \tag{3}$$

Several divergences/distances have been considered; see *e.g.*, [3] for a review. The original choice, which is also one of the most natural since it corresponds to the Fisher distance of the multivariate Gaussian distribution [12], is the natural Riemannian – or affine-invariant – distance defined as

$$\delta_{\mathrm{R}}^2(\boldsymbol{A}, \boldsymbol{B}) = \|\log(\boldsymbol{B}^{-1}\boldsymbol{A})\|_2^2, \tag{4}$$

where $\|\cdot\|_2$ denotes the Frobenius norm and $\log(\cdot)$ is the matrix logarithm. The corresponding barycenter of $\{\boldsymbol{C}_i\}_{i=1}^N$ is the so-called geometric mean, which is the unique solution to the fixed-point equation

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}_i)\right). \tag{5}$$

where $\exp(\cdot)$ is the matrix exponential. As no closed-form expression is known for $p > 2$, it is computed using a recursive algorithm based on (5) (see, e.g., [4]).

Another choice of particular interest for d is the Kullback-Leibler divergence between two centered multivariate Gaussian distributions [13], defined as

$$d_{\mathrm{KL}}(\boldsymbol{A}, \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^{-1}) - \log|\boldsymbol{A}\boldsymbol{B}^{-1}| - p, \tag{6}$$

where $\operatorname{tr}(\cdot)$ and $|\cdot|$ denote the trace and determinant operators, respectively. Since the Kullback-Leibler divergence is not symmetric, one can obtain two barycenters. Here, we use the "right" barycenter, *i.e.*, we choose $\boldsymbol{A} = \boldsymbol{C}_i$ and $\boldsymbol{B} = \boldsymbol{\Sigma}$ as in (2) (see [14] for a review on the barycenters of this Kullback-Leibler divergence). The corresponding right barycenter of $\{\boldsymbol{C}_i\}_{i=1}^N$ is simply the arithmetic mean

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{C}_i. \tag{7}$$

The MDM classifiers associated with (4) and (6) are denoted RMDM and rKLMDM, respectively. In Section IV, it is shown that they can be seen as Discriminant Analysis classifiers under specific assumptions.

III. Wishart and t-Wishart distributions

In this section, we present the Wishart and t-Wishart distributions. After providing definitions, we give estimation methods of their parameter, which lies in $\mathcal{S}_p^{++}$. While the maximum likelihood estimator (MLE) of the Wishart distribution is known in closed form, one has to employ an iterative algorithm to obtain the t-Wishart's one. In this work, we propose a method relying on Riemannian optimization that is very similar to the one in [15]. However, it is a bit simpler as it relies on the usual affine-invariant metric on $\mathcal{S}_p^{++}$ rather than on the Fisher information metric of the distribution.

The probability density function (pdf) of a random matrix $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T \in \mathcal{S}_p^{++}$, with $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ $(n \geq p)$, following the Wishart distribution $\mathcal{W}(n, \boldsymbol{\Sigma})$ with center $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$ is, up to a normalization factor,

$$f^{\mathcal{W}}(\boldsymbol{S}|\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}} |\boldsymbol{S}|^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right). \tag{8}$$

As for the usual multivariate t-distribution, the t-Wishart distribution is obtained by replacing the exponential in (8) with another density generator. The pdf of $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T$ following the t-Wishart distribution t-$\mathcal{W}(n, \boldsymbol{\Sigma}, \nu)$ with center $\boldsymbol{\Sigma}$ and degree of freedom (d.o.f.) $\nu > 0$ is, up to a normalization factor,

$$f^{t\text{-}\mathcal{W}}(\boldsymbol{S}|\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}} |\boldsymbol{S}|^{\frac{n-p-1}{2}} \left(1 + \frac{\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})}{\nu}\right)^{-\frac{\nu+np}{2}}. \tag{9}$$

Notice that, compared to the Wishart distribution, the t-Wishart distribution adds a dependence on the columns of $\boldsymbol{X}$. The independence in the case of the Wishart distribution is a direct consequence of the properties of the exponential function. Thus, substituting it with another density generator cancels this property.

Given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{S}_i\}_{i=1}^N$, the log-likelihood $\mathcal{L}$ associated to the distribution with pdf f is $\mathcal{L}(\boldsymbol{\Sigma}) = \sum_i \log f(\boldsymbol{S}_i|\boldsymbol{\Sigma})$. The log-likelihoods of the Wishart and t-Wishart distributions are denoted $\mathcal{L}^{\mathcal{W}}$ and $\mathcal{L}^{t\text{-}\mathcal{W}}$, respectively. To obtain the MLE, one needs to solve the optimization problem $\hat{\boldsymbol{\Sigma}} = \operatorname{argmax}_{\boldsymbol{\Sigma}} \mathcal{L}(\boldsymbol{\Sigma})$. In the case of the Wishart distribution, it is

$$\hat{\boldsymbol{\Sigma}}^{\mathcal{W}} = \frac{1}{nN}\sum_{i=1}^{N} \boldsymbol{S}_i. \tag{10}$$

For the t-Wishart distribution, we must develop an iterative algorithm. Here, we perform a Riemannian gradient descent [16] on $\mathcal{S}_p^{++}$ to obtain the MLE $\hat{\boldsymbol{\Sigma}}^{t\text{-}\mathcal{W}}$. Given a sequence of iterates $\{\hat{\boldsymbol{\Sigma}}^{(j)}\}$, the next iterate is

$$\hat{\boldsymbol{\Sigma}}^{(j+1)} = R_{\hat{\boldsymbol{\Sigma}}^{(j)}}\left(t_j \nabla\mathcal{L}^{t\text{-}\mathcal{W}}\left(\hat{\boldsymbol{\Sigma}}^{(j)}\right)\right), \tag{11}$$

where $\nabla\mathcal{L}^{t\text{-}\mathcal{W}}$ is the Riemannian gradient of $\mathcal{L}^{t\text{-}\mathcal{W}}$ in $\mathcal{S}_p^{++}$, $R_{\cdot}(\cdot)$ is a retraction on $\mathcal{S}_p^{++}$ (mapping from tangent spaces back onto the manifold), and t_j is a stepsize computed through a linesearch [16]. To define the Riemannian gradient, one has to choose a Riemannian metric on $\mathcal{S}_p^{++}$. Here, we select the usual affine-invariant metric. Given $\hat{\boldsymbol{\Sigma}} \in \mathcal{S}_p^{++}$ and

tangent vectors $\boldsymbol{\xi}$, $\boldsymbol{\eta} \in \mathcal{S}_p$ (set of $p \times p$ symmetric matrices), it is $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\hat{\boldsymbol{\Sigma}}} = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\xi} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\eta})$. The Riemannian gradient $\nabla \mathcal{L}^{t\text{-}\mathcal{W}}(\hat{\boldsymbol{\Sigma}})$ at $\hat{\boldsymbol{\Sigma}}$ is then defined as the only tangent vector such that, for all $\boldsymbol{\xi} \in \mathcal{S}_p$, $\langle \nabla \mathcal{L}^{t\text{-}\mathcal{W}}(\hat{\boldsymbol{\Sigma}}), \boldsymbol{\xi} \rangle_{\hat{\boldsymbol{\Sigma}}} = \mathrm{D}\, \mathcal{L}^{t\text{-}\mathcal{W}}(\hat{\boldsymbol{\Sigma}})[\boldsymbol{\xi}]$, where D denotes the directional derivative. One can show

$$\nabla \mathcal{L}^{t\text{-}\mathcal{W}}(\hat{\boldsymbol{\Sigma}}) = \frac{1}{2} \sum_{i=1}^{N} \frac{\nu + np}{\nu + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_i)} \boldsymbol{S}_i - \frac{nN}{2} \hat{\boldsymbol{\Sigma}}. \quad (12)$$

Finally, the retraction we choose is the second-order approximation of geodesics on $\mathcal{S}_p^{++}$ (a generalization of straight lines to manifolds), which is arguably the best choice from a numerical perspective [17]. Given $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\xi}$, it is

$$R_{\hat{\boldsymbol{\Sigma}}}(\boldsymbol{\xi}) = \hat{\boldsymbol{\Sigma}} + \boldsymbol{\xi} + \frac{1}{2} \boldsymbol{\xi} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\xi}. \quad (13)$$

IV. Discriminant Analysis classifiers

In this section, a novel Bayesian classifier is proposed for signal classification. Inspired by Linear/Quadratic Discriminant Analysis (LDA/QDA) in the multivariate case [18], the introduced method generalizes the concept of Discriminant Analysis to the matrix-variate case. It exploits Wishart and t-Wishart distributions to model the likelihood of observed samples knowing their associated labels.

A. Wishart and t-Wishart Discriminant Analysis classifiers

Rather than considering SCMs, $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T$ is used. The following proposition introduces the t-Wishart Discriminant Analysis, denoted t-$\mathcal{W}$DA.

Proposition 1 (t-$\mathcal{W}$DA)**.** *The decision rule of the t-Wishart Discriminant Analysis, for a testing covariance matrix $\boldsymbol{S}$, is given by*

$$\hat{y}(\boldsymbol{S}) = \underset{k \in [\![1,K]\!]}{\mathrm{argmax}} \quad \delta_k^{t\text{-}\mathcal{W}}(\boldsymbol{S}), \quad (14)$$

with the discriminant function

$$\delta_k^{t\text{-}\mathcal{W}}(\boldsymbol{S}) = \log(\hat{\pi}_k) - \frac{n}{2} \log |\hat{\boldsymbol{\Sigma}}_k| - \frac{\nu + np}{2} \log \left(1 + \frac{\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{S})}{\nu} \right), \quad (15)$$

where $\hat{\pi}_k$ is the proportion of the class k in the training set; $\hat{\boldsymbol{\Sigma}}_k$ is the estimated barycenter of training covariance matrices of the class k, computed thanks to (11)*, and ν is the degree of freedom of the model*[1].

Proof. Being a probabilistic classifier, the t-$\mathcal{W}$DA has the following discrimination rule

$$\hat{y}(\boldsymbol{S}) = \underset{k \in [\![1,K]\!]}{\mathrm{argmax}} \quad P(y = k | \boldsymbol{S}). \quad (16)$$

Using Bayes formula, $P(y = k|\boldsymbol{S}) \propto \pi_k f_k(\boldsymbol{S})$, where π_k is the prior probability of the class k and $f_k(\boldsymbol{S})$ is the conditional probability density function observed on the class k.

t-$\mathcal{W}DA$ approaches the classification problem by assuming that $\boldsymbol{S}|y = k$ follows the t-Wishart distribution t-$\mathcal{W}(n, \boldsymbol{\Sigma}_k, \nu)$ where $\boldsymbol{\Sigma}_k$ is considered as the center of the class k. Hence, by plugging in the pdf defined in (9) and applying the logarithm function, (16) becomes

$$\hat{y}(\boldsymbol{S}) = \underset{k \in [\![1,K]\!]}{\mathrm{argmax}} \quad \log \left(\pi_k \, f^{t\text{-}\mathcal{W}}(\boldsymbol{S}|\boldsymbol{\Sigma_k}) \right). \quad (17)$$

where $f^{t\text{-}\mathcal{W}}(\boldsymbol{S}|\boldsymbol{\Sigma_k})$ is defined in (9).

Concerning the prior distribution of the class k, π_k is simply estimated by the proportion of training samples belonging to class k, denoted as $\hat{\pi}_k$. Notice that $\hat{\pi}_k$ is a consistent estimate of π_k, a direct consequence of the strong law of large numbers.

The centers $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ are also estimated during the training step. A natural estimation choice is the MLE of t-$\mathcal{W}(n, \boldsymbol{\Sigma}, \nu)$, denoted $\{\hat{\boldsymbol{\Sigma}}_k\}_{k=1}^K$. They are computed as described in Section III. We assume that the MLE of the center for the t-Wishart distribution is consistent [2]. As a result of the continuous mapping theorem and the Slutsky lemma, $\log \left(\hat{\pi}_k \, f^{t\text{-}\mathcal{W}}(\boldsymbol{S}|\hat{\boldsymbol{\Sigma}}_k) \right)$ converges in probability to $\log \left(\pi_k \, f^{t\text{-}\mathcal{W}}(\boldsymbol{S}|\boldsymbol{\Sigma}_k) \right)$ when the size of the training set tends to $+\infty$. The term depending only on $\boldsymbol{S}$ is then neglected due to the argmax operator, which concludes the proof. □

Remark 1. *As mentioned in Proposition 1, the d.o.f. ν is a hyperparameter; estimating it during the training step is beyond the scope of this paper. A possible way to tune it is to perform a grid search. Furthermore, we would highlight that in the actual model, a unique d.o.f. is chosen for all classes for simplicity. It would be more realistic to estimate different d.o.f. for each class to improve discrimination, especially when the centers of classes are close.*

Similarly, the Wishart Discriminant Analysis classifier denoted $\mathcal{W}$DA, is derived in the following proposition.

Proposition 2 ($\mathcal{W}$DA)**.** *The decision rule of the Wishart Discriminant Analysis, for a testing covariance matrix $\boldsymbol{S}$, is given by*

$$\hat{y}(\boldsymbol{S}) = \underset{k \in [\![1,K]\!]}{\mathrm{argmax}} \quad \delta_k^{\mathcal{W}}(\boldsymbol{S}), \quad (18)$$

with the discriminant function

$$\delta_k^{\mathcal{W}}(\boldsymbol{S}) = log(\hat{\pi}_k) - \frac{n}{2} \log |\hat{\boldsymbol{\Sigma}}_k| - \frac{1}{2} \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{S}).$$

where $\hat{\pi}_k$ is the proportion of the class k in the training set, and $\hat{\boldsymbol{\Sigma}}_k$ is the estimated barycenter of training covariance matrices of the class k, computed thanks to Eq. (10)*.*

Proof. The proof follows the same steps as for the t-$\mathcal{W}$DA. □

The discriminant function of the t-$\mathcal{W}$DA approximates the $\mathcal{W}$DA's one when the dof ν tends to $+\infty$.

Remark 2. *Let us examine the decision boundary between two classes k and l for t-$\mathcal{W}DA$ and $\mathcal{W}DA$. We recall that it corresponds to the region $\{\boldsymbol{S} \in \mathcal{S}_p^{++} : \delta_k(\boldsymbol{S}) = \delta_l(\boldsymbol{S})\}$, where $\delta_k := \delta_k^{t\text{-}\mathcal{W}}$ for t-$\mathcal{W}DA$ and $\delta_k := \delta_k^{\mathcal{W}}$ for $\mathcal{W}DA$.*

[1] considered here as a hyperparameter

[2] Due to the lack of space, this will be proved in a forthcoming paper

The decision boundary is linear in $\boldsymbol{S}$ since it is a solution to an equation involving the trace of $\boldsymbol{M}_{kl}\boldsymbol{S}$, where $\boldsymbol{M}_{kl}$ is a symmetric matrix that depends on the model parameters. This makes the proposed classifier very close to LDA. However, finding decision boundaries for $\mathcal{W}$DA/t-$\mathcal{W}$DA is more complex due the constraint $\boldsymbol{S} \in \mathcal{S}_p^{++}$.

B. Interpretation of RMDM and rKLMDM in the framework of Discriminant Analysis

As mentioned in Section II, rKLMDM can be related to the proposed $\mathcal{W}$DA classifier. In fact, if the training set is balanced, rKLMDM corresponds exactly to $\mathcal{W}$DA. First, one can observe that the MLE for the center of the i.i.d. samples $\{\boldsymbol{S}_i\}_{i=1}^N$ following the Wishart distribution is the arithmetic mean of $\{\boldsymbol{C}_i\}_{i=1}^N$ (with the notation $\boldsymbol{S}_i = n\boldsymbol{C}_i$). Therefore, rKLMDM and $\mathcal{W}$DA share the same barycenters of classes. Furthermore, one can write

$$n \log |\hat{\boldsymbol{\Sigma}}_k| + \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{S}) = n(d_{\text{KL}}(\boldsymbol{C}|\hat{\boldsymbol{\Sigma}}_k) + p + \log|\boldsymbol{C}|).$$

Hence, rKLMDM and $\mathcal{W}$DA have the same decision rule.

Similarly, under the condition of a balanced training set, RMDM can be seen as the Discriminant Analysis classifier associated with the Riemannian Gaussian distribution on $\mathcal{S}_p^{++}$ defined in [19]. We recall that the pdf of the Riemannian Gaussian distribution $\mathcal{G}(\boldsymbol{\Sigma}, \sigma^2)$ with center $\boldsymbol{\Sigma}$ and dispersion parameter $\sigma > 0$ is, up to a normalization factor, $f(\boldsymbol{C}|\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2\sigma^2}\delta_{\text{R}}^2(\boldsymbol{\Sigma}, \boldsymbol{C})\right)$. Moreover, its MLE corresponds to the geometric mean of the samples. The observed likelihood associated with RMDM in a Discriminant Analysis framework is obtained with the model $\boldsymbol{C}|y = k \sim \mathcal{G}(\boldsymbol{\Sigma}_k, \sigma^2)$ (or equivalently, $\boldsymbol{S}|y = k \sim \mathcal{G}(n\boldsymbol{\Sigma}_k, \sigma^2)$). The barycenters of classes in the MDM classifier are exactly the MLEs for centers of Riemannian Gaussian distributions. Under the assumption of equal proportions π_k's, the decision rule for the MDM given by (3) with the affine-invariant Riemannian distance (4) is exactly recovered.

V. Numerical Experiments

This section aims to show the practical interest of the proposed t-$\mathcal{W}$DA classifier. To do so, numerical experiments are conducted on two different real EEG datasets. The datasets, which are available on the MOABB platform[3], correspond to two different BCI paradigms. The first one [3] contains steady-state visually evoked potentials (SSVEP) while the second one [20] concerns motor imagery (MI). In both cases, the classification performances of RMDM, $\mathcal{W}$DA (or rKLMDM) and t-$\mathcal{W}$DA are compared.

A. SSVEP dataset

SSVEP signals are natural responses to repetitive visual stimuli at specific frequencies, *i.e.*, the visual cortex synchronizes with the stimuli. In this work, we employ the dataset used in [3], where 12 subjects are asked to look at light emitting diodes (LEDs) blinking at three different frequencies: 17, 13 and 21 Hz. EEG signals are acquired on eight electrodes located around the visual cortex and the sampling rate is set to 256 Hz. For each subject, the recordings of two to five sessions are available. A session contains 32 labeled trials of five seconds each. Data are equally divided into four classes: a class per stimulus frequency and a resting class where no light is blinking.

Before proceeding to the classification, a preprocessing step is needed. This allows us to better extract the information at the frequencies of interest in the data. Following [3], each EEG trial $\boldsymbol{X}$ becomes

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}^{(17)} \\ \boldsymbol{X}^{(13)} \\ \boldsymbol{X}^{(21)} \end{bmatrix}, \tag{19}$$

where $\boldsymbol{X}^{(f)}$ is the result of backward-forward filtering of $\boldsymbol{X}$ with a band-pass region $[f - \Delta f, f + \Delta f]$, where $\Delta f = 0.5$ Hz. Therefore, $\tilde{\boldsymbol{X}} \in \mathbb{R}^{p \times n}$ with $p = 3 \times 8 = 24$ and $n = 5 \times 256 = 1280$. Transformed signals $\tilde{\boldsymbol{X}}$ are centered and their covariances are computed.

A within-session classification is run: for each subject, a session is randomly divided into a training set composed of 20 trials and a testing set of 12 trials such that all classes are equally represented in the training set. A Monte-Carlo approach of 100 permutations is performed in each session. For the t-$\mathcal{W}$DA, the degree of freedom ν is set to 10. The means and standard deviations of test accuracies for each subject are reported in Figure 1.

One can observe that the RMDM yields the best results for two subjects (2 and 4); $\mathcal{W}$DA is the most accurate for three subjects (6, 8, and 9); and t-$\mathcal{W}$DA features the best performance for the remaining seven subjects. Averaging over all the subjects, t-$\mathcal{W}$DA has a 3.11% gain over RMDM and a 1.19% gain over $\mathcal{W}$DA. Thus, our proposed t-$\mathcal{W}$DA classifier appears advantageous on these SSVEP data.

B. MI dataset

MI is a mental process during which a subject mentally simulates a physical action, *i.e.*, he/she imagines moving their right or left hand, feet, tongue, *etc*. In this work, we consider the BNCI2014001 dataset [20]. It contains the EEG recordings from 9 subjects. Signals are acquired via 22 electrodes with a sampling rate of 250 Hz. For each subject, two sessions composed of 288 trials are available. Only the four last seconds of the trial are taken into account. Thus, for each trial, we have $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ with $p = 22$ and $n = 4 \times 250 = 1000$.

A within-session classification is run: cross-validation of stratified 5-folds is considered on each session, dividing it into 228 training samples and 60 testing samples with a balanced representation of the 4 classes. For the t-$\mathcal{W}$DA, the degree of freedom ν is set to 10. The means and standard deviations of test accuracies for each subject are reported in Figure 2.

One can observe that $\mathcal{W}$DA never features the best performance on this dataset. RMDM is the most accurate for two subjects (1 and 5) and has a similar accuracy for subject 2 compared to t-$\mathcal{W}$DA. Finally, t-$\mathcal{W}$DA yields the best results

[3]https://github.com/NeuroTechX/moabb

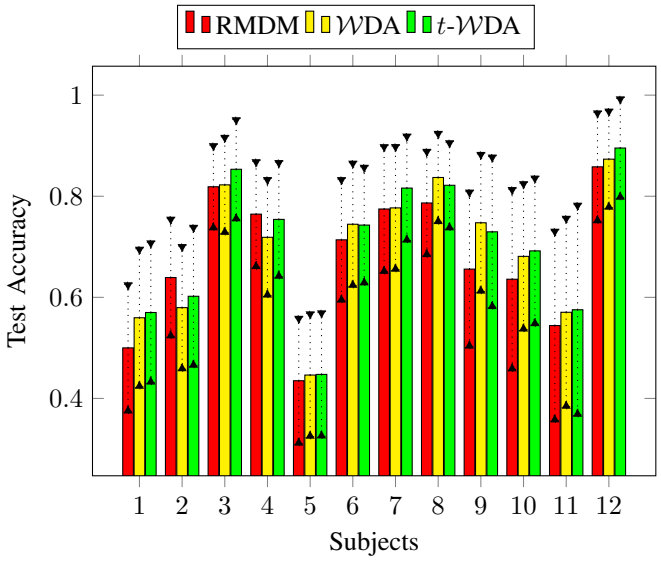

Fig. 1. Comparison of classifiers on the SSVEP dataset [3]. Standard deviations are plotted in dotted lines. The degree of freedom for t-$\mathcal{W}$DA is $\nu = 10$.

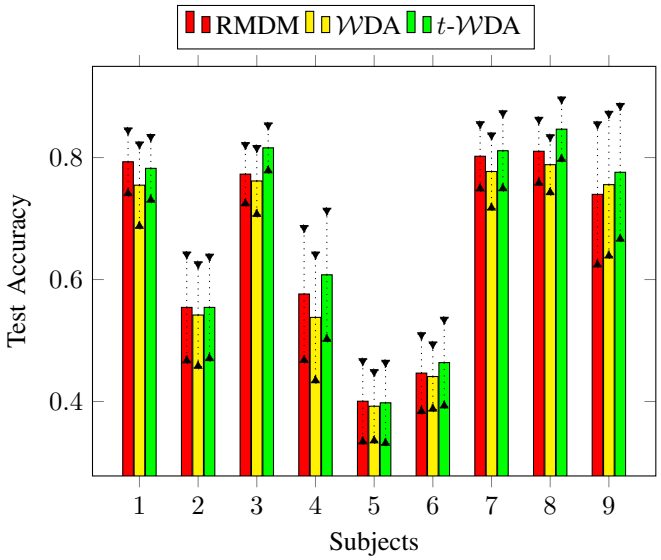

Fig. 2. Comparison of classifiers on the MI dataset. Standard deviations are plotted in dotted lines. The degree of freedom for t-$\mathcal{W}$DA is $\nu = 10$.

for the five remaining subjects. Averaging over all the subjects, t-$\mathcal{W}$DA has a 1.78% gain over RMDM and a 3.39% gain over $\mathcal{W}$DA. Again, t-$\mathcal{W}$DA appears to be the most powerful classifier on this MI dataset compared to RMDM and $\mathcal{W}$DA.

VI. Conclusion and Perspectives

In this paper, we propose an original classification method that exploits the t-Wishart distribution over the set of matrices $\mathcal{S}_p^{++}$. The classifier is obtained by leveraging the Discriminant Analysis framework. On two standard EEG BCI datasets, our proposed method outperforms best state-of-the-art MDM methods.

This work also yields several perspectives. As of now, the degree of freedom of the t-Wishart distribution is a hyper-parameter. Instead, it should be estimated during the training step and different degrees of freedom for the different classes should be considered. Other Elliptical Wishart distributions could also be investigated.

References

[1] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[2] 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," *Journal of Neural Engineering*, vol. 15, no. 3, 2018.

[3] S. Chevallier, E. Kalunga, Q. Barthélemy, and E. Monacelli, "Review of Riemannian distances and divergences, applied to SSVEP-based BCI," *Neuroinformatics*, vol. 19, 01 2021.

[4] 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, 2012.

[5] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[6] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

[7] G. Z. Thompson, R. Maitra, W. Q. Meeker, and A. F. Bastawros, "Classification with the matrix-variate t distribution," *Journal of Computational and Graphical Statistics*, vol. 29, no. 3, pp. 668–674, 2020.

[8] K.-T. Fang and T. W. Anderson, *Statistical inference in elliptically contoured and related distributions*. Allerton Press, 1990.

[9] F. J. Caro-Lopera, G. González-Farías, and N. Balakrishnan, "On generalized Wishart distributions-I: Likelihood ratio test for homogeneity of covariance matrices," *Sankhya A*, vol. 76, no. 2, pp. 179–194, 2014.

[10] J. Wishart, "The generalized product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20A, no. 1-2, pp. 32–52, 12 1928.

[11] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Transactions on signal processing*, vol. 60, no. 11, pp. 5597–5625, 2012.

[12] L. T. Skovgaard, "A Riemannian geometry of the multivariate normal model," *Scandinavian journal of statistics*, pp. 211–223, 1984.

[13] S. Kullback, "Information theory and statistics," *New York: Dover*, 1968.

[14] Z. Chebbi and M. Moakher, "Means of hermitian positive-definite matrices based on the log-determinant α-divergence function," *Linear Algebra and its Applications*, vol. 436, no. 7, pp. 1872–1889, 2012.

[15] I. Ayadi, F. Bouchard, and F. Pascal, "Elliptical Wishart distribution: maximum likelihood estimator from information geometry," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[16] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton University Press, 2008.

[17] B. Jeuris, R. Vandebril, and B. Vandereycken, "A survey and comparison of contemporary algorithms for computing the matrix geometric mean," *Electronic Transactions on Numerical Analysis*, vol. 39, pp. 379–402, 2012.

[18] B. Ghojogh and M. Crowley, "Linear and quadratic discriminant analysis: Tutorial," 2019. [Online]. Available: https://arxiv.org/abs/1906.02590

[19] S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton, "Riemannian gaussian distributions on the space of symmetric positive definite matrices," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2153–2170, 2017.

[20] 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*, p. 55, 2012.