ROBUST GEOMETRIC ESTIMATION OF COVARIANCE MATRICES FOR MI-EEG CLASSIFICATION

TAKASHI UEHARA, TOSHIHISA TANAKA, SIMONE FIORI*

Abstract. The estimation of covariance matrices is of prime importance to analyze the distribution of multivariate signals. In motor imagery based brain-computer interfaces (MI-BCI), covariance matrices play a central role in the extraction of features from recorded electroencephalograms (EEGs), therefore, estimating covariances properly is crucial for EEG classification. The present paper focuses on averaging sample covariance matrices (SCMs) for the selection of the reference matrix in tangent space mapping (TSM) based MI-BCI. Tangent space mapping is a powerful method of feature extraction and strongly depends on the selection of reference matrix considered as a covariance matrix. In general, the observed signals may include outliers, therefore, the geometric mean of SCMs which is taken as reference matrix in previous works may not be the best choice. In order to deal with the effects of outliers, we propose the use of robust estimators. In particular, the geometric medians and trimmed averages (defined on the basis of several metrics) are proposed, discussed and tested against each others. The main idea behind trimmed averages is to eliminate those data that exhibit the largest distance from the average covariance calculated on the basis of all available data. We conducted experiments with several types of SCM averages including the conventional methods and the proposed methods. The results of the experiment show that, while the geometric medians show little differences from conventional methods in terms of classification accuracy in the classification of electroencephalographic recordings, the trimmed averages show a significant improvement for all subjects. These results suggest that our proposed averaging method affords estimating covariance matrices of EEG recordings effectively.

Key words. Motor imagery; electroencephalogram; brain-computer interfacing; spatial covariance matrices; automated classification; Riemannian manifold; symmetric positive-definite matrices.

1. Introduction. Brain-computer interfacing (BCI) is a challenging application of signal processing, machine learning, and neuroscience [13]. A brain-computer interface captures brain activities associated with mental tasks and external stimuli to enable non-muscular communication between the brain and computers. Motor imagery (MI) is one of the promising events for BCI to elicit typical brain activities in sensorimotor cortex measured with electroencephalography (EEG). Since MI includes the information to classify which part of the body the subject imagines to move, motor-imagery-based BCI (MI-BCI) is expected to be utilized in medical fields such as rehabilitation.

A well-known approach to extract features form observations of MI is spatial filtering named common spatial pattern (CSP) method [32, 36]. A common spatial pattern is a spatial weight vector whose elements are corresponding to each electrode in a multichannel EEG. Such spatial filter performs linear combination of the EEG signals. The coefficients of such filter are determined so that the variance of the signal extracted by the spatial filter differs between two opposing tasks (e.g., left and right hand movement imagery). The feature vector is obtained as a log-variance of filtered signals.

As an entirely different approach, a method using the covariance structure of the data has been proposed [4]. Since the covariance matrices belong to a Riemannian manifold, we can manipulate such matrices by taking into account the curved nature of the manifold [14,17,31]. The method named tangent space mapping (TSM) extracts

*T. Uehara and T. Tanaka are with the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology (TUAT), 2-24-16, Nakacho, Koganei-shi, Tokyo 184-8588, Japan. S. Fiori is with the Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy

features to classify them by using Riemannian geometry. All sample covariance matrices (SCMs) are mapped onto the linear tangent space at their geometric mean. Since the tangent space is linear, many popular and effective algorithms such as principal component analysis and support vector machine can be applied for classification.

The performance of feature extraction methods such as CSP and TSM depends on the covariance matrices estimated from the observed signals. The simplest approach for estimation is computing the uniform mean of SCMs from each trial. However, the observed EEG normally includes outlier trials strongly affected by the artifacts such as eye blink and muscle movement. Therefore, the uniform mean of SCMs may not be a reliable estimator since the mean is sensitive to outliers [19, 39]. In order to deal with the effects of outliers, robust estimators of covariance matrices have been proposed [30, 37, 39, 40, 42]. However, most of them neglect the covariance structure, namely SCMs are symmetric positive-definite (SPD) and the space of SPD matrices has special metrics.

In this study, we focus on TSM-based feature extraction for MI-BCI because TSM shows higher performances than CSP in classification accuracy [4, 5, 29]. However, TSM itself does not consider the effect of the outliers. We propose to use more robust estimators of covariance matrices by averaging SCMs by several metrics. Particularly, the geometric median and trimmed averaging are proposed and discussed. The geometric median is one of the multivariate extensions of the median (in the scalar case) which is widely used as a robust estimator of centrality instead of the mean. The trimmed average for SCMs removes those samples that lay too far from the average of all matrices and compute an average of the remaining samples. Note that both of them consider the distance between the pair of SCMs so that the effect of outliers for averaging is suppressed.

The experimental evaluation on real EEG data illustrates a comparison between CSP and TSM, the performances of the geometric median for SCMs, and the performance of the trimmed averages for SCMs. We confirm that TSM shows significantly higher classification accuracy than CSP, the geometric median shows little improvement, and that trimmed averaging achieves a significant improvement over the conventional averaging.

The present paper is organized as follows. Section 2 overviews the fundamental features of MI-BCI and the basic methods of feature extraction including CSP and TSM. Section 3 summarizes recent theoretical works for estimating mean covariance matrices. We revise existing robust averaging methods for spatial covariance matrices, and propose new methods, in Section 4. Experimental results are described and discussed in Section 5. Section 6 concludes the paper.

2. Motor-Imagery-Based Brain-Computer Interfacing. The current section recalls from the specialized literature some aspects of motor-imagery-based brain-computer interfacing. In particular, we first present a short overview of MI-BCI in Section 2.1 and then we discuss the feature extraction methods that are commonly used in motor imagery analysis in Section 2.2.

2.1. Overview of MI-BCI. BCI devices are interfaces directly connecting a brain to the external world without muscle movements of any body parts of a human subject. Any BCI using imagination of muscle movements is an interface where a user inputs a command by imaging movement (motor imagery) of body parts such as hands, feet, and tongue [13]. As an example, a task of the motor imagery of right or left hand is to be classified from brain signals. A schematic of a fundamental structure of MI-BCI type is shown in Figure 2.1.

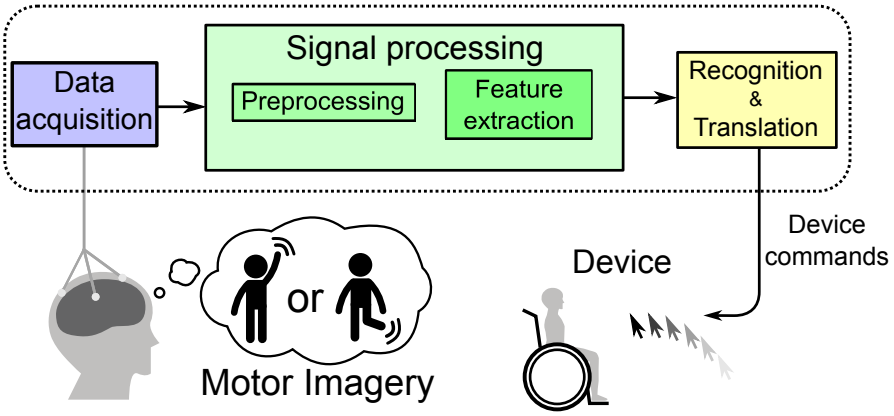

FIG. 2.1. *An exemplary structure of motor-imagery-based brain-computer interface.*

The tasks of the motor imagery decrease the energy in certain frequency bands called mu (8–15 Hz) and beta (10–30 Hz) bands in EEG signals observed by the electrodes located on (sensory) motor cortices. The decrease of the spectral energy is called *event-related desynchronization* (ERD) [38]. Moreover, it is known that the physical location in the brain where an ERD is obserced depends on the body part of which a subject imagines the movement [22], therefore, we can estimate which body part the subject imagines the movement of from EEG signals by detecting the location where ERD happens. An ERD is induced by motor-imagery tasks performed by healthy subjects as well as by paralyzed patients [26].

One of the merits of MI-BCI against the BCIs using perception or flicker stimuli is that an instrument for displaying stimuli is unnecessary. Moreover, it has been reported recently that the detection of the motor-imagery tasks and a feedback are useful for rehabilitation for patients who suffer from motor disorders caused by brain damages [1, 35]. One of the rehabilitation procedures for the recovery of the motor functions is to make a subject perform movements of a disabled body part by a cue and give a visual or physical feedback [35]. The rehabilitation takes advantage of the plasticity of the brain. The plasticity of the brain is an intrinsic property that enables the nervous system to adapt to the environmental pressures, physiologic changes, and experiences [34]. The coincident events of the intention of the movement that the subject performs and the corresponding feedback to the subject are supposed to promote the plasticity of the brain. Because of the plasticity of the brain, the other parts of the brain with no damage take over the functions disabled by the damages. In the rehabilitation, to generate the feedback coincidentally with intention is considered to be significant. In the general procedure of the rehabilitation illustrated above, the cues control the generation of the intention and of the feedback. Using the MI-BCIs for the rehabilitation, the feedback generation is controlled by BCI. MI-BCI enable the rehabilitation system to detect the intention of the movements and generate the feedback at an appropriate time. Some researches have suggested that the rehabilitation by the MI-BCIs can promote the plasticity of the brain more efficiently than the conventional system based on the cue [1, 35].

2.2. Feature extraction from MI data. The acquired brain signals are mixed with components associated to various brain activities and noise. The acquired EEG signals are recorded by means of M electrodes (it is said that the recording possesses

M channels), as illustrated in the Figure 2.2, where $\mathbf{x}_n$ denotes a multichannel readout at the discrete time n ranging from 1 to N, which denotes the total number of samples per trial. In order to estimate the MI task the subjects performed, we remove the noise

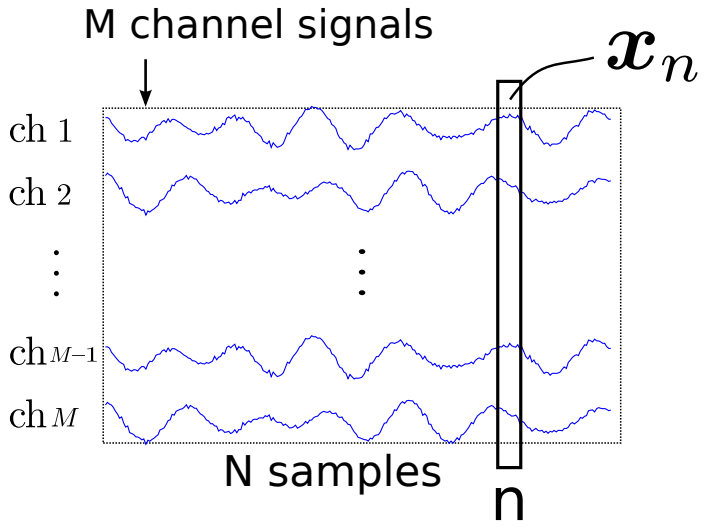

FIG. 2.2. *Illustration of an EEG recording during a motor-imagery task by means of multiple electrodes.*

and extract the feature components that are associated with the task. As recalled in Subsection 2.1, the features of the MI tasks are observed as the energy changes of certain frequency bands. Moreover, spatial patterns changes depend on the kind of MI tasks performed by a subject [38]. For example, it is said that the MI of a left hand evokes energy changes in the right side of a motor cortex. The MI of a right hand movement evokes energy changes in the left side of a motor cortex. These differences in the spatial patterns are caused by the fact that different brain regions work for each muscle. By analyzing the spatial patterns of ERD, we can associate an observed EEG signal with the MI tasks which a subject performs. In the following subsections, we revise two different ways of extracting spatial features from motor-imagery data, namely, *common spatial pattern* and *tangent space mapping*.

2.2.1. Common spatial pattern. A common spatial pattern is a spatial filter described by the relationship

$$y_n = \sum_{m=1}^{M} v_m x_{m,n}, \tag{2.1}$$

where $x_{m,n}$ denotes a sample observed in the mth channel at time n, v_m is a weight coefficient for the signal observed in the mth channel, y_n is the spatial-filtered signal at the time n, as illustrated in the Figure 2.3. The coefficients of a CSP are organized into a vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_M]^T \in \mathbb{R}^M$.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{M\times N}$ denote an observed multichannel signal corresponding to a MI trial ($\boldsymbol{x}_n$ is the nth column of the matrix $\boldsymbol{X}$). An experiment in motor imagery is made up of several different trials, each trial being a recording session where a subject is asked to imagine a motor task. Each trial produces a data-matrix $\boldsymbol{X}$. Normally, a subject is asked to perform the same motor imagery task

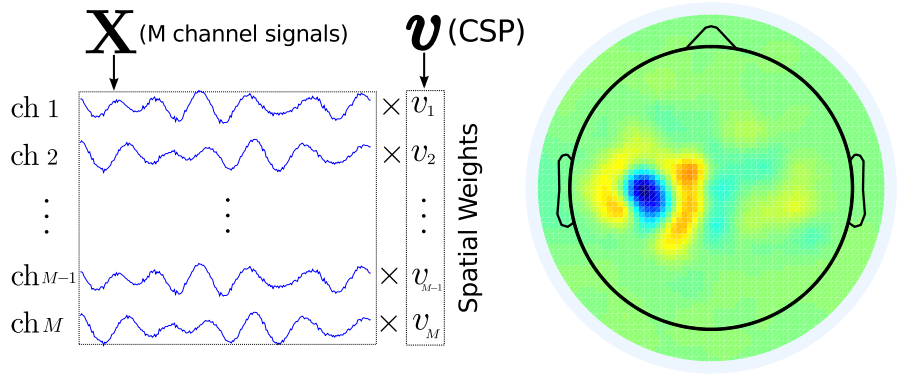

FIG. 2.3. *(left) Illustration of a CSP in terms of spatial weights. The observed signals multiplied by the CSP coefficients give the extracted signals. (right) Illustration of a CSP on a scalp.*

more than once, so as to acquire a 'library' of data matrices pertaining to each of the considered motor-imagery tasks. The set of data-matrices pertaining to the same MI task is termed *class* and is denoted by $\mathcal{C}$.

A spatial filter $\boldsymbol{v} \in \mathbb{R}^M$ is calculated in such a way that the variance of a signal extracted by the linear combination of the subsignals in the data-matrix $\boldsymbol{X}$ is minimized within a MI class [32, 36]. The empirical average of the observed signals is given by

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n, \tag{2.2}$$

The empirical variance of the signals extracted from $\boldsymbol{X}$ by the CSP is given by

$$\sigma^2(\boldsymbol{X}, \boldsymbol{\mu}) = \frac{1}{N}\sum_{n=1}^{N} [\boldsymbol{v}^T(\boldsymbol{x}_n - \boldsymbol{\mu})]^2. \tag{2.3}$$

We assume a two-class MI experiment and we denote the two corresponding sets of the learning data as $\mathcal{C}_1$ and $\mathcal{C}_2$ ($\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$). The CSP method seeks the weight vector $\boldsymbol{v} \in \mathbb{R}^M$ that minimizes the intra-class variance under the normalization of samples. More specifically, CSP is formulated as the following optimization problem [32, 36];

$$\begin{aligned} \min_{\boldsymbol{v}} \quad & \mathbb{E}_{\boldsymbol{X}\in\mathcal{C}_c}\left[\sigma^2(\boldsymbol{X}, \boldsymbol{v})\right], \\ \text{subject to} \quad & \sum_{d=1,2} \mathbb{E}_{\boldsymbol{X}\in\mathcal{C}_d}\left[\sigma^2(\boldsymbol{X}, \boldsymbol{v})\right] = 1, \end{aligned} \tag{2.4}$$

where $\mathbb{E}_{\boldsymbol{X}\in\mathcal{C}_d}[\cdot]$ denotes the expectation over $\mathcal{C}_d$ and c is a fixed class label (1 or 2 in the present example). The optimization problem (2.4) may be rewritten as

$$\begin{aligned} \min_{\boldsymbol{v}} \quad & \boldsymbol{v}^T\boldsymbol{\Sigma}_c\boldsymbol{v}, \\ \text{subject to} \quad & \boldsymbol{v}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{v} = 1, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\Sigma}_d$, $d = 1, 2$, are defined as

$$\boldsymbol{\Sigma}_d = \mathbb{E}_{\boldsymbol{X}\in\mathcal{C}_d}\left[\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T(\boldsymbol{x}_n - \boldsymbol{\mu})\right]. \tag{2.6}$$

Furthermore, the constrained optimization problem (2.5) may be reformulated as the unconstrained minimization of a Rayleigh quotient, namely as

$$\min_{\boldsymbol{v}} \quad \frac{\boldsymbol{v}^T \boldsymbol{\Sigma}_c \boldsymbol{v}}{\boldsymbol{v}^T (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \boldsymbol{v}}. \tag{2.7}$$

Many other methods for feature extraction such as principal component analysis (PCA), canonical correlation analysis (CCA) [23], and linear discriminant analysis (LDA) [8] are formulated in a similar manner.

The solution of (2.7) is given by the generalized eigenvector corresponding to the smallest generalized eigenvalue of the generalized eigenvalue problem described as

$$\boldsymbol{\Sigma}_c \boldsymbol{v} = \lambda (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \boldsymbol{v}. \tag{2.8}$$

Although the solution of (2.7) is given by the eigenvector corresponding to the smallest eigenvalue in (2.8), the remaining eigenvectors might be used for classification purposes [9]. A total of M eigenvectors can be obtained by solving the matrix-pencil problem (2.8): Let us denote by $\hat{\boldsymbol{v}}_i$ the eigenvector corresponding to the ith smallest eigenvalue. In CSP, $2r$ eigenvectors are used for classification of unlabeled data, with r is an integer that controls the number of features used in classification. The feature vector $\boldsymbol{f} \in \mathbb{R}^{2r}$ used to classify an unlabaled data-matrix $\boldsymbol{X}$ is defined as

$$\boldsymbol{f} = [\sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_1), \ldots, \sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_r), \ldots, \sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_{M-r+1}), \ldots, \sigma^2(\boldsymbol{X}, \hat{\boldsymbol{v}}_M)]. \tag{2.9}$$

For classification purposes, log-variances of spatial filtered signal, i.e., the logarithm of the entries of the vector $\boldsymbol{f}$, are used as inputs to a classifier such as linear discriminant analyzer.

2.3. Tangent-space mapping of spatial covariance matrices. Spatial covariance matrices are often utilized to extract features from MI data since they encode spatial information embedded in the recorded signals. Let us denote by $\boldsymbol{X}_i \in \mathbb{R}^{M \times N}$ a data-matrix consisting of a M-channel EEG signal with N time samples corresponding to the ith trial of a MI mental task, where the index i ranges from 1 to the total number of trials, denoted by I. Each row of the data-matrix $\boldsymbol{X}_i$ gets pre-processed (i.e., the mean of the $\boldsymbol{X}_i$'s row gets zeroed and band-pass filtered). For the i-th trial, the spatial covariance matrix is estimated as a sample covariance matrix

$$\boldsymbol{P}_i = \frac{1}{N-1} \boldsymbol{X}_i \boldsymbol{X}_i^T \in \mathbb{R}^{M \times M}, \tag{2.10}$$

as illustrated in the Figure 2.4.

Tangent-space mapping of covariance matrices is a novel approach to extract features from MI recordings [4]. Barachant *et al.* proposed a nonlinear transformation of spatial covariance matrices and a framework that directly classifies the covariance matrices of the observed signals instead of spatially filtered signals' variances.

Mathematically speaking, a spatial covariance matrix is an instance of *symmetric, positive-definite* matrix. We denote the set of all $M \times M$ real symmetric matrices as $S(M)$ and the set of all $M \times M$ symmetric positive-definite matrices as $\mathrm{P}(M) = \left\{ \boldsymbol{P} \in \mathrm{S}(M) \mid \boldsymbol{u}^T \boldsymbol{P} \boldsymbol{u} > 0, \forall \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{u} \neq \boldsymbol{0} \right\}$.

In general, a d-dimensional manifold $\mathcal{M}$ is a topological space that can be locally regarded as d-dimensional Euclidean space $\mathbb{R}^d$ in the neighborhood of any point $p \in \mathcal{M}$ [21]. Especially, if the differential of $\mathcal{M}$ is globally defined at any point p, $\mathcal{M}$ is called

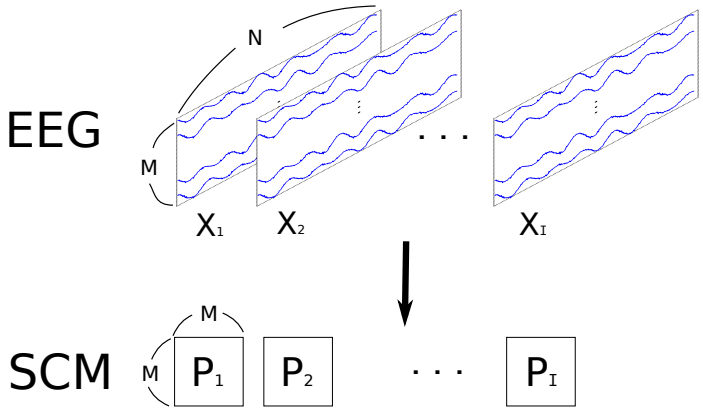

FIG. 2.4. *Empirical spatial covariance matrices of multichannel EEG recordings.*

differentiable manifold. For example, the sphere in the three-dimensional Euclidean space is a 2-dimensional differentiable manifold because the sphere is differentiable at any point and the neighborhood of any point in the sphere can be approximated as as patch on a two-dimensional Euclidean space.

In a differentiable manifold $\mathcal{M}$, tangent vectors at $p \in \mathcal{M}$ can be calculated and the set of such tangent vectors span an Euclidean space called the tangent space $T_p\mathcal{M}$. A Riemannian manifold is a manifold that is endowed with a specific metric, that is an inner product of tangent vectors $\langle \cdot, \cdot \rangle_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ defined in $T_p\mathcal{M}$ at any point $p \in \mathcal{M}$. The inner product of two tangent vectors induces a distance between two different points on the Riemannian manifold.

The set of SPD matrices form a Riemannian manifold. The tangent space of $\mathrm{P}(M)$ at $\boldsymbol{P} \in \mathrm{P}(M)$ denoted as $T_{\boldsymbol{P}}\mathrm{P}(M)$ is isomorphic to the space of real symmetric matrices $\mathrm{S}(M)$. The tangent space locally approximates the space $\mathrm{P}(M)$. The tangent space mapping of spatial covariance matrices is illustrated in the Figure 2.5. Since the

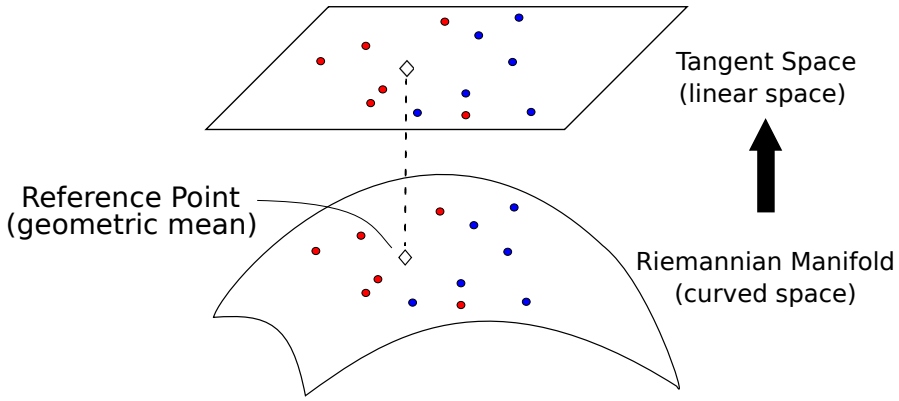

FIG. 2.5. *Illustration of the tangent space mapping of covariance matrices.*

tangent space is a linear space, efficient classification algorithms, such as LDA, can be applied if all the spatial covariance matrices are mapped onto the tangent space [4].

Each spatial covariance matrix corresponding to a given MI trial $\boldsymbol{P}_i \in \mathbb{R}^{M \times M}$ is mapped into the tangent space located at the geometric mean $\boldsymbol{P}_{\mathcal{G}}$ of the whole set of

trials [18] and vectorized by

$$\boldsymbol{s}_i = \text{upper}\left(\log\left(\boldsymbol{P}_{\mathcal{G}}^{-\frac{1}{2}}\boldsymbol{P}_i\boldsymbol{P}_{\mathcal{G}}^{-\frac{1}{2}}\right)\right), \tag{2.11}$$

where the upper$(\cdot)$ operator vectorizes the upper triangular part of a symmetric matrix by applying the weight $\sqrt{2}$ only to the off-diagonal entries and $\log(\cdot)$ represents the principal matrix logarithm.

The dimensionality $M(M+1)/2$ of a reference tangent space may be so large that it would be beneficial to reduce it for classification purposes. Using principal component analysis, the vectors in the vectorized-tangent-matrices $\boldsymbol{Q} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_I] \in \mathbb{R}^{m\times I}$ get orthogonalized and the number of dimensions gets reduced to obtain $\boldsymbol{Q}_0 = \boldsymbol{U}^T\boldsymbol{Q} \in \mathbb{R}^{l\times I}$, where $\boldsymbol{U}$ is the matrix of the l left singular vectors of $\boldsymbol{Q}$. The new basis vectors are sorted in ascending order of p-values given by the one-way ANOVA. Linear discriminant analysis is applied to classify the columns of $\boldsymbol{Q}_0$ after the variable selections.

3. Spatial Covariance Matrices in MI-BCI. As recalled in Subsection 2.2, spatial covariance matrices play crucial roles in feature extraction from MI recording, therefore, estimating spatial covariance matrices is a crucial problem in MI-BCI. For example, the spatial filters in the CSP method are designed by the intra-class covariance matrix defined in (2.6). The TSM method also strongly depends on the spatial covariance matrices of the whole set of observed MI signals as shown in (2.11). Note that both the empirical unbiased estimators and the maximum-likelihood estimator of covariance matrices are sensitive to outliers. Particularly, while estimating covariances from only one trial, other robust covariance estimators and regularization techniques can be applied to reduce the influence of outliers [27]. However, we focus on other approaches for the robust estimation of covariances.

There exist many approaches to estimate an average covariance matrix out of a set of empirical spatial covariance matrices of EEG recordings [30, 37, 39, 40, 42].

Covariance shrinkage is a popular estimatotion methods [6, 20, 27, 30]. The simplest covariance shrinkage approach consists in taking a convex combination of an empirical average $\bar{\boldsymbol{P}}$ and a scaled identity matrix, namely

$$\boldsymbol{P}_{\text{sh}} = (1-\mu)\bar{\boldsymbol{P}} + \mu\frac{\text{tr}(\bar{\boldsymbol{P}})}{M}\boldsymbol{I}_M, \tag{3.1}$$

where $\mu \in [0,\ 1]$ is a regularization parameter and $\boldsymbol{I}_M$ denotes a M-dimensional identity matrix. Covariance shrinkage performs well when the number N of observations is smaller than the number of variables M. Therefore, such an approach does not seem suitable in EEG-BCI applications because in the BCI paradigm the recorded signals are usually well conditioned, namely, the number N is larger than the number M.

One of the robust ways to estimate covariances is to down-weight noisy trials in the averaging of sample spatial covariance matrices. To this aim, Samek *et al.* [37] proposed the concept of minimizing the β-divergence that, in the scalar case, reads

$$D_\beta(p(x)\ \|\ q(x)) = \int\left[\frac{1}{\beta}\left\{p^\beta(x) - q^\beta(x)\right\}p(x)\ - \frac{1}{\beta+1}\left\{p^{\beta+1}(x) - q^{\beta+1}(x)\right\}\right]dx, \tag{3.2}$$

where $p(x)$ is the empirical data distribution and $q(x)$ is the model probability distribution with parameter θ, and where $\beta > 0$. Minimizing the divergence leads to a robust estimate of the parameter θ.

In the present case, the data and the estimated parameter can be regarded as spatial covariance matrices. Note that the covariance matrices are going to get weighted, therefore the input data may be taken as scatter matrices rather than covariance matrices, which differ from each other by a scaling coefficient only. For the matrix-type data model, we can use a Wishart distribution

$$\mathcal{W}(\boldsymbol{S};\Sigma,N)=\frac{1}{2^{\frac{NM}{2}}|\Sigma|^{\frac{N}{2}}\Gamma_M\left(\frac{N}{2}\right)}|\boldsymbol{S}|^{\frac{N-M-1}{2}}\exp\left\{\mathrm{tr}\left(\frac{1}{2}\Sigma^{-1}\boldsymbol{S}\right)\right\}, \tag{3.3}$$

where $\boldsymbol{S}\in\mathbb{R}^{M\times M}$ is a scatter matrix and Γ_M denotes the multivariate 'gamma function' defined as

$$\Gamma_M\left(\frac{N}{2}\right)=\pi^{M(M-1)/4}\prod_{j=1}^{M}\Gamma\left(\frac{N}{2}+\frac{1-j}{2}\right). \tag{3.4}$$

In the above expressions, $\Gamma(\cdot)$ denotes a gamma function and the notation $|\boldsymbol{S}|$ denotes the determinant of the matrix $\boldsymbol{S}$.

The estimated covariance matrix that minimizes a β-divergence between Wishart distributions can be obtained by means of the iterative procedure

$$\Sigma^{(k+1)}=\frac{\sum_{i=1}^{I}\psi_\beta(\boldsymbol{S}_i;\Sigma^{(k)},N)\boldsymbol{S}_i}{N\sum_{i=1}^{I}\psi_\beta(\boldsymbol{S}_i;\Sigma^{(k)},N)-\gamma|\Sigma^{(k)}|^{(N-M-1)\beta/2}}, \tag{3.5}$$

where I denotes again the number of trials that the robust average is being taken out from, while

$$\psi_\beta(\boldsymbol{S};\Sigma,N)=|\boldsymbol{S}|^{(N-M-1)\beta/2}\exp\left\{\mathrm{tr}\left(\frac{1}{2}\Sigma^{-1}\boldsymbol{S}\right)\right\}, \tag{3.6}$$

$\boldsymbol{S}_i=\boldsymbol{X}_i\boldsymbol{X}_i^T$, and

$$\gamma=\frac{I\beta(M+1)}{2^{\frac{NM}{2}}\Gamma_M\left(\frac{N}{2}\right)(\beta+1)}\left(\frac{2}{\beta+1}\right)^{\frac{NM(\beta+1)}{2}-\frac{M(M+1)\beta}{2}}\Gamma_M\left(\frac{N(\beta+1)}{2}-\frac{(M+1)\beta}{2}\right). \tag{3.7}$$

The initial parameter-matrix $\Sigma^{(0)}$ may be chosen as the identity matrix $\boldsymbol{I}$ or as one of the $\boldsymbol{S}_i$'s at random.

Note that ψ_β works as a factor down-weighting the influence of outlier trials and for $\beta=0$ the solution gives

$$\Sigma=\frac{1}{IN}\sum_{i=1}^{I}\boldsymbol{S}_i=\frac{1}{I}\sum_{i=1}^{I}\boldsymbol{P}_i, \tag{3.8}$$

which is equivalent to the standard maximum-likelihood estimator of the (or to the arithmetic) mean.

In order to estimate an average covariance matrix out of a set of spatial covariance matrices, methods like the ones above, treating covariance matrices as elements of the

Euclidean space $\mathbb{R}^{M\times M}$, have been extensively invoked in the framework of brain-computer interfacing. However, such approaches are affected by a side effect called *swelling effect* [14, 43]. The value of the determinant of a covariance is a measure of the dispersion of the observed signal, so that it stands as one of the most relevant properties of covariances matrices as much as positive-definiteness. Since the space of SPD matrices is a convex set, the arithmetic average of a set of SPD matrices lays in this space. However, the linear average of SPD matrices, that are characterized by the same determinant, possesses a lager determinant than that of the original matrices. Such a swelling is apparently an artifact due to the averaging method.

It is well known that the space of symmetric, positive-definite matrices is a Riemannian manifold (that can also be regarded as a Lie group, upon defining a group structure [2]) that may be endowed with specific particular metrics. Taking into account the geometry of such a curved space leads to a geometrically sound definition of average of spatial covariance matrices, which possesses the additional advantage of preserving the property of the determinant of matrices [14]. An graphical rendering of a two-class data set is given in the Figure 3.1. Each dot in the Figure denotes an empirical spatial covariance matrix $\boldsymbol{P}_i$.

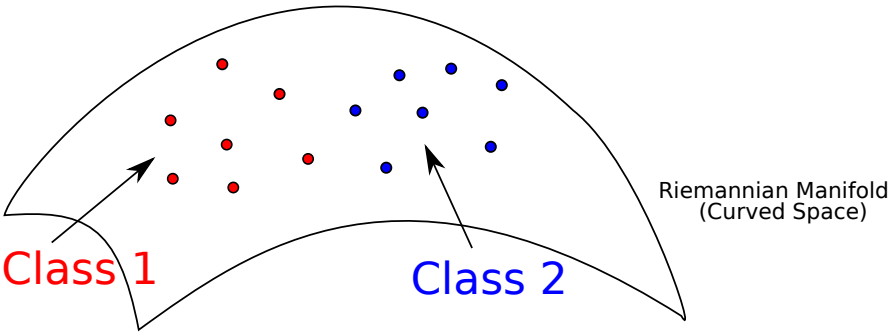

FIG. 3.1. *Illustration of SCMs from a two-class experiment in MI. Each dot denotes an empirical spatial covariance matrix. The samples lay over a curved surface that represents the manifold of SPD matrices* $\mathrm{P}(M)$.

4. Robust Averaging of EEG-MI Spatial Covariance Matrices. The aim of the present section is to describe new methods to compute the robust mean of spatial covariance matrices in the context of motor imagery. We suggest to use different metrics to achieve averaging over the space of SPD matrices and also propose a new averaging method based on data trimming. The Subection 4.1 summarizes the geometry of the manifold of SPD matrices (namely, the Riemannian metrics and related distance functions that are used in the context of averaging). The Subsection 4.2 explains geometric averaging by criterion function minimization on the basis of a number of possible metrics. Subsection 4.3 proposes a new averaging method for SPD matrices termed *trimmed averaging*.

4.1. The geometry of the SPD manifold. A SPD matrix is always diagonalizable with strictly real positive eigenvalues. The principal matrix logarithm of a symmetric, positive-definite matrix $\boldsymbol{P} \in \mathrm{P}(M)$ and the matrix exponential of a symmetric matrix $\boldsymbol{S} \in \mathrm{S}(M)$ is obtained using the eigenvalue decomposition of $\boldsymbol{P}$ and $\boldsymbol{S}$,

namely, by

$$\log(\boldsymbol{P}) = \boldsymbol{U}\ \mathrm{diag}(\log(\sigma_1), \ldots, \log(\sigma_n))\ \boldsymbol{U}^T, \tag{4.1}$$

$$\exp(\boldsymbol{S}) = \boldsymbol{V}\ \mathrm{diag}(\exp(\phi_1), \ldots, \exp(\phi_n))\ \boldsymbol{V}^T. \tag{4.2}$$

where the σ_i's and the ϕ_i's denote the eigenvalues and $\boldsymbol{U}$ and $\boldsymbol{V}$ denote the matrices of the eigenvectors of $\boldsymbol{P}$ and $\boldsymbol{S}$, respectively.

Let $\boldsymbol{A}, \boldsymbol{B} \in T_{\boldsymbol{P}}\mathrm{P}(M)$ be tangent vectors at $\boldsymbol{P}$ on $\mathrm{P}(M)$. The natural metric on the space of SPD matrices is named *affine invariant Riemannian metric* [11,12] and is defined as

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\boldsymbol{P}} = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{P}^{-1}\right). \tag{4.3}$$

Such inner product induces the norm for the tangent vectors:

$$\|\boldsymbol{A}\|_{\boldsymbol{P}} = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}^{-1}\right)^{1/2} = \|\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\|_{\mathrm{F}}, \tag{4.4}$$

where the symbol $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Note that if the reference point $\boldsymbol{P}$ is taken as the identity $\boldsymbol{I}$, the above metrics coincide to the Euclidean inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}} = \mathrm{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right)$ and the Frobenius norm $\|\boldsymbol{A}\|_F = \mathrm{tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{1/2}$, respectively.

The exponential map $\mathrm{Exp}_{\boldsymbol{P}} : T_{\boldsymbol{P}}\mathrm{P}(M) \to \mathrm{P}(M)$ maps a tangent vector to the manifold, while the logarithmic map $\mathrm{Log}_{\boldsymbol{P}} : \mathrm{P}(M) \to T_{\boldsymbol{P}}\mathrm{P}(M)$ performs the inverse operation. The formulation of such maps is:

$$\mathrm{Exp}_{\boldsymbol{P}}(\boldsymbol{S}) = \boldsymbol{P}^{1/2}\exp(\boldsymbol{P}^{-1/2}\boldsymbol{S}\boldsymbol{P}^{-1/2})\boldsymbol{P}^{1/2}, \tag{4.5}$$

$$\mathrm{Log}_{\boldsymbol{P}}(\boldsymbol{Q}) = \boldsymbol{P}^{1/2}\log(\boldsymbol{P}^{-1/2}\boldsymbol{Q}\boldsymbol{P}^{-1/2})\boldsymbol{P}^{1/2}, \tag{4.6}$$

for $\boldsymbol{P}, \boldsymbol{Q} \in \mathrm{P}(M)$ and $\boldsymbol{S} \in \mathrm{S}(M)$.

The natural metric on the manifold of SPD matrices defined by the local inner product (4.3) induces the Riemannian distance between two SPD matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ [31]:

$$d_R(\boldsymbol{P}, \boldsymbol{Q}) = \|\mathrm{Log}_{\boldsymbol{P}}(\boldsymbol{Q})\|_{\boldsymbol{P}} = \|\log\left(\boldsymbol{P}^{-1/2}\boldsymbol{Q}\boldsymbol{P}^{-1/2}\right)\|_{\mathrm{F}}. \tag{4.7}$$

4.2. Averaging of SCMs Using Various Metrics. On the basis of the above metrics it is possible to design closed-form or iterative algorithms to compute the mean out of a set of spatial covariance matrices. The averages based on other metrics also achieved successful performances in a number of applications [2,28]. We will introduce alternative averages, aimed at robustness, to achieve a more effective tangent space mapping and, consequently, better performing classifications schemes.

The mean is the most common estimator of centrality. There exist several types of means of positive scalars, such as the arithmetic, the geometric and the harmonic mean, which may be generalized to arbitrarily complex, metrizable spaces, by means of an appropriate formulation of averaging in terms of a least-squares problem [15]. For a set of SPD matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I$, the mean of these matrices is expressed as

$$\boldsymbol{P}_{\mathrm{MEAN}} = \underset{\boldsymbol{P} \in \mathrm{P}(M)}{\arg\min} \sum_{i=1}^{I} d^2(\boldsymbol{P}, \boldsymbol{P}_i) \tag{4.8}$$

where $d(\cdot, \cdot)$ denotes a distance function in $\mathrm{P}(M)$. Note that the use of different distance functions in (4.8) results in different types of means.

By using the Euclidean distance

$$d_E(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{F}}, \tag{4.9}$$

the solution of (4.8) results in the *arithmetic mean*. The arithmetic mean admits the closed-form expression

$$\boldsymbol{P}_{\mathcal{A}} = \frac{1}{I} \sum_{i=1}^{I} \boldsymbol{P}_i. \tag{4.10}$$

The Riemannian geometric mean of SPD matrices is one of the averages extended from the geometric mean in the set of positive numbers. By making use of the Riemannian geodesic distance

$$d_R(\boldsymbol{A}, \boldsymbol{B}) = \| \log \left(\boldsymbol{A}^{-1/2} \boldsymbol{B} \boldsymbol{A}^{-1/2} \right) \|_{\mathrm{F}}, \tag{4.11}$$

the solution of (4.8) results in the Riemannian geometric mean. The minimum is uniquely determined on the manifold of SPD matrices [24]. This mean is referred to as the Karcher mean. There is no closed-form solution for the Riemannian geometric mean, but an efficient iterative algorithm to compute such mean was given in [17,31] as:

$$\boldsymbol{S}^{(k)} = \frac{1}{I} \sum_{i=1}^{I} \mathrm{Log}_{\boldsymbol{P}_{\mathcal{R}}^{(k)}}(\boldsymbol{P}_i), \tag{4.12}$$

$$\boldsymbol{P}_{\mathcal{R}}^{(k+1)} = \mathrm{Exp}_{\boldsymbol{P}_{\mathcal{R}}^{(k)}}(\boldsymbol{S}^{(k)}), \tag{4.13}$$

where the initial guess $P_{\mathcal{R}}^{(0)}$ is set to the identity matrix $\boldsymbol{I}$.

The Log-Euclidean mean is another extension of the geometric mean of positive numbers. By using a Log-Euclidean distance

$$d_L(\boldsymbol{A}, \boldsymbol{B}) = \| \log(\boldsymbol{A}) - \log(\boldsymbol{B}) \|_F, \tag{4.14}$$

the solution of the minimization problem (4.8) results in the Log-Euclidean mean. The closed-form solution of such optimization problem is given by [17]:

$$\boldsymbol{P}_{\mathcal{L}} = \exp \left(\frac{1}{I} \sum_{i=1}^{I} \log \boldsymbol{P}_i \right). \tag{4.15}$$

It is worth mentioning that such an averaging rule is a special instance of a more general class of averaging schemes, termed *Kolmogoroff-Nagumo averaging formulas*, that were recently surveyed in [16] (although it was not applied to SPD matrices in [16]).

The harmonic mean is one of Pythagorean means, together with the arithmetic and the geometric mean. Since the harmonic mean is always the smallest of the three Pythagorean means, it can suppress the effects of the large outlier. The harmonic mean is used for several applications, e.g., computing the centrality of graph in graph theory and in network analysis [10]. By way of the metric

$$d_H(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\|_F \tag{4.16}$$

in (4.8), the harmonic mean of SPD matrices is induced. The solution of the optimization problem can be expressed as

$$\boldsymbol{P}_{\mathcal{H}} = \left(\frac{1}{I} \sum_{i=1}^{I} \boldsymbol{P}_i^{-1} \right)^{-1}. \tag{4.17}$$

The above expression states that the inverse of $\boldsymbol{P}_{\mathcal{H}}$ is regarded as the arithmetic mean of inverses of the $\boldsymbol{P}_i$'s.

Although not directly related to the optimization problem (4.8), we think that it is appropriate to cite here the notion of *resolvent mean* of SPD matrices, because it exhibit similar properties to the Riemannian geometric mean [7]. The resolvent of a symmetric, positive-definite matrix A is defined as

$$J_{\boldsymbol{A}} = (\boldsymbol{A} + \boldsymbol{I})^{-1}, \tag{4.18}$$

and the resolvent mean is given as

$$\boldsymbol{P}_{\mathcal{R}}(\mu) = \left(\sum_{i=1}^{I} (\boldsymbol{P}_i + \mu^{-1}\boldsymbol{I})^{-1} \right)^{-1} - \mu^{-1}\boldsymbol{I}, \tag{4.19}$$

where $\mu > 0$. Such averaging scheme is motivated from the fact that when $\mu = 1$, it holds that

$$J_{\boldsymbol{P}_{\mathcal{R}}(\mu=1)} = \frac{1}{I} \sum_{i=1}^{I} J_{\boldsymbol{P}_i}, \tag{4.20}$$

that is, the resolvent of $\boldsymbol{P}_{\mathcal{R}}(\mu = 1)$ is the arithmetic mean of resolvents of the matrices $\boldsymbol{P}_i$. The resolvent mean is viewed as a parameterized harmonic mean [25] because there hold a harmonic-resolvent-arithmetic mean inequality and limits

$$\mathcal{H} \preceq \mathcal{R}_{\mu} \preceq \mathcal{A}, \tag{4.21}$$

$$\mathcal{R}_{\mu} \to \mathcal{A} \quad \text{when} \quad \mu \to 0^{+}, \tag{4.22}$$

$$\mathcal{R}_{\mu} \to \mathcal{H} \quad \text{when} \quad \mu \to +\infty. \tag{4.23}$$

The Riemannian geometric mean also satisfies the inequality (4.21), therefore, the resolvent mean and the Riemannian geometric mean seem to enjoy similar properties.

In general, the mean is a very popular estimator of centrality, although it is known not to be robust with respect to outliers. On the other hand, the median is one of the most common robust estimators of centrality. In the multivariate/structured case, the median has several types of extensions from the scalar case. The geometric median, also referred to as spatial median, is one of the types of multivariate medians. Similarly to the means, the geometric medians can be particularized with different metrics to obtain different averages. The geometric median is defined through the following optimization problem:

$$\boldsymbol{P}_{\text{MED}} = \underset{\boldsymbol{P} \in \text{S}(M)}{\arg\min} \sum_{i=1}^{I} d(\boldsymbol{P}, \boldsymbol{P}_i). \tag{4.24}$$

Comparing this definition (4.24) with the definition of mean (4.8), it is readily appreciated that the geometric mean is the minimum of the summed *squared* distances, while the geometric median is the minimum of the summed distances. The

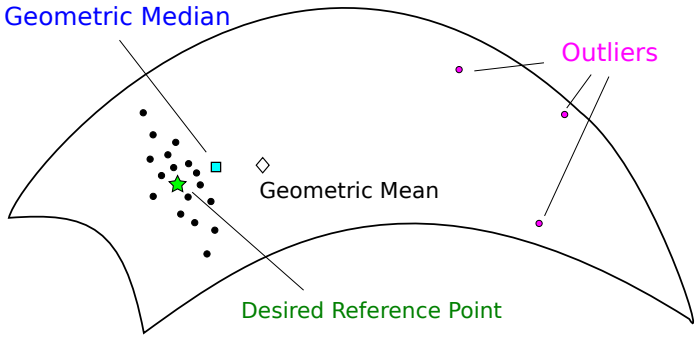

FIG. 4.1. *Illustration of the geometric median on a Riemannian manifold. The geometric mean is more sensitive to the outliers than the geometric median.*

geometric median de-emphasizes the samples that lay farther from the center of the distribution, hence resulting in a more robust estimation of centrality, as illustrated in the Figure 4.1.

The *Euclidean geometric median* is a geometric median in the space $\mathrm{P}(M)$ considered as an Euclidean space endowed with an Euclidean metric. The solution of the minimization problem is given by the following iterative algorithm named *Weiszfeld algorithm* [41]:

$$\boldsymbol{P}_{\mathcal{A}^*}^{(k+1)} = \left(\sum_{i+1}^{I} \frac{\boldsymbol{P}_i}{d_E(\boldsymbol{P}_{\mathcal{A}^*}^{(k)}, \boldsymbol{P}_i)}\right)\left(\sum_{i+1}^{I} \frac{1}{d_E(\boldsymbol{P}_{\mathcal{A}^*}^{(k)}, \boldsymbol{P}_i)}\right)^{-1}, \tag{4.25}$$

where $\boldsymbol{P}_{\mathcal{A}^*}^{(k)}$ denotes the value of the Euclidean geometric median of the set of SPD matrices $\boldsymbol{P}_i$ in the kth iteration. The initial guess $\boldsymbol{P}_{\mathcal{A}^*}^{(0)}$ can be chosen as in such a way that $\boldsymbol{P}_{\mathcal{A}^*}^{(0)} \neq \boldsymbol{P}_i$, e.g., as the identity matrix. This median uses the same metric as the arithmetic mean of SPD matrices and arises as a natural extension of the corresponding type of median of real numbers.

The geometric median in linear spaces is one of the most common robust estimators of centrality, therefore, such geometric median is extended to manifold-valued representations [19]. The geometric median on a Riemannian manifold may be computed by the iterative algorithm:

$$\boldsymbol{V}^{(k)} = \left(\sum_{i=1}^{I} \frac{\mathrm{Log}_{\boldsymbol{P}_{\mathcal{G}}^{(k)}}(\boldsymbol{P}_i)}{d_R(\boldsymbol{P}_{\mathcal{G}}^{(k)}, \boldsymbol{P}_i)}\right)\left(\sum_{i=1}^{I} \frac{1}{d_R(\boldsymbol{P}_{\mathcal{G}}^{(k)}, \boldsymbol{P}_i)}\right)^{-1}, \tag{4.26}$$

$$\boldsymbol{P}_{\mathcal{G}}^{(k+1)} = \mathrm{Exp}_{\boldsymbol{P}_{\mathcal{G}}^{(k)}}\left(\boldsymbol{V}^{(k)}\right), \tag{4.27}$$

This algorithm is analogous to the Weiszfeld algorithm for Euclidean data. The initial guess $\boldsymbol{P}_{\mathcal{G}}^{(0)}$ can be chosen as such as $\boldsymbol{P}_{\mathcal{G}^*}^{(0)} \neq \boldsymbol{P}_i$, e.g., as the identity matrix.

The *Log-Euclidean median* is also computed by a Weiszfeld-like algorithm that embodies the Log-Euclidean metric. We compute this type of median by the iteration

$$\boldsymbol{V}^{(k)} = \left(\sum_{i=1}^{I} \frac{R_e^{-1}(\boldsymbol{P}_{\mathcal{L}^*}^{(k)}, \boldsymbol{P}_i)}{d_R(\boldsymbol{P}_{\mathcal{L}^*}^{(k)}, \boldsymbol{P}_i)}\right)\left(\sum_{i=1}^{I} \frac{1}{d_R(\boldsymbol{P}_{\mathcal{L}^*}^{(k)}, \boldsymbol{P}_i)}\right)^{-1}, \tag{4.28}$$

$$\boldsymbol{P}_{\mathcal{L}^*}^{(k+1)} = R_e(\boldsymbol{P}_{\mathcal{L}^*}^{(k)}, \boldsymbol{V}^{(k)}), \tag{4.29}$$

where R_e and R_e^{-1} denote a retraction map and its inverse (lifting) operator respectively, defined as

$$R_e(\boldsymbol{P}_{\text{ref}}, \boldsymbol{S}) = \exp(\log(\boldsymbol{P}_{\text{ref}}) + \boldsymbol{S}), \tag{4.30}$$

$$R_e^{-1}(\boldsymbol{P}_{\text{ref}}, \boldsymbol{P}) = \log(\boldsymbol{P}) - \log(\boldsymbol{P}_{\text{ref}}) \tag{4.31}$$

where $\boldsymbol{P}_{\text{ref}}, \boldsymbol{P} \in \mathrm{P}(M)$ and $\boldsymbol{S} \in \mathrm{S}(M)$. The initial guess $\boldsymbol{P}_{\mathcal{L}^*}^{(0)}$ can be chosen as in such a way that $\boldsymbol{P}_{\mathcal{L}^*}^{(0)} \neq \boldsymbol{P}_i$ as, e.g., the identity matrix.

4.3. Trimmed averages of spatial covariance matrices. A direct approach to deal with outliers is to eliminate them from a data set. In the case of scalar real-valued data sets, the mean of the remaining data after removing the lowest and highest is called *trimmed mean* (also referred to, in the literature, as *truncated mean* or *Windsor mean*) [3]. A trimmed mean is less susceptible to the effects of extreme samples than is the arithmetic mean.

The trimmed mean of scalar numbers has a parameter α. When $\alpha\%$ lowest and $\alpha\%$ highest data are eliminated, the mean of remaining data is called $\alpha\%$ trimmed mean. (As a consequence, the median is the mean trimmed 100% and the arithmetic mean is the mean trimmed 0%.)

In order to extend the notion of trimmed mean to the case of SPD matrices, it is necessary to define the meaning of "lowest data" and "highest data". Since we can compute the distance between SPD matrices and their averages by using several metrics, we can estimate the outliers according to the distance of the samples from the averages.

We define the *trimmed Riemannian mean* $\boldsymbol{P}_{\hat{\mathcal{G}}}$ and the *trimmed Log-Euclidean mean* $\boldsymbol{P}_{\hat{\mathcal{L}}}$ of a set of spatial covariance matrices as

$$\boldsymbol{P}_{\hat{\mathcal{G}}} = \mathcal{G}\left[\mathrm{Trim}_d^{\mathcal{G}}(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I)\right], \tag{4.32}$$

$$\boldsymbol{P}_{\hat{\mathcal{L}}} = \mathcal{L}\left[\mathrm{Trim}_d^{\mathcal{L}}(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I)\right], \tag{4.33}$$

where $\mathcal{G}[\cdot]$ and $\mathcal{L}[\cdot]$ are a Riemannian geometric mean and a Log-Euclidean mean operator, and $\mathrm{Trim}_d^{\mathcal{G}}(\cdot)$ and $\mathrm{Trim}_d^{\mathcal{L}}(\cdot)$ are the trimming operators discarding $d\%$ matrices that exhibit the largest distance to the mean in terms of Riemannian and Log-Euclidean metrics, respectively.

The Riemannian median and the Log-Euclidean geometric median are more robust estimators than the means. However, more robust statistics will be obtained if we discard the outliers before computing the medians. On the basis of such observation, we define the *trimmed Riemannian median* $\boldsymbol{P}_{\hat{\mathcal{G}}^*}$ and the *trimmed Log-Euclidean median* $\boldsymbol{P}_{\hat{\mathcal{L}}^*}$ as

$$\boldsymbol{P}_{\hat{\mathcal{G}}^*} = \mathcal{G}^*\left[\mathrm{Trim}_d^{\mathcal{G}^*}(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I)\right], \tag{4.34}$$

$$\boldsymbol{P}_{\hat{\mathcal{L}}^*} = \mathcal{L}^*\left[\mathrm{Trim}_d^{\mathcal{L}^*}(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I)\right], \tag{4.35}$$

where $\mathcal{G}^*[\cdot]$ and $\mathcal{L}^*[\cdot]$ denote a Riemannian geometric median and a Log-Euclidean median operators, while $\mathrm{Trim}_d^{\mathcal{G}^*}(\cdot)$ and $\mathrm{Trim}_d^{\mathcal{L}^*}(\cdot)$ are the trimming operators discarding the $d\%$ of matrices that show the largest distance to the median in terms of Riemannian and Log-Euclidean metrics, respectively.

In order to clarify how the above trimmed means/medians work, the mean-based trimmed mean is illustrated in the Figure 4.2.

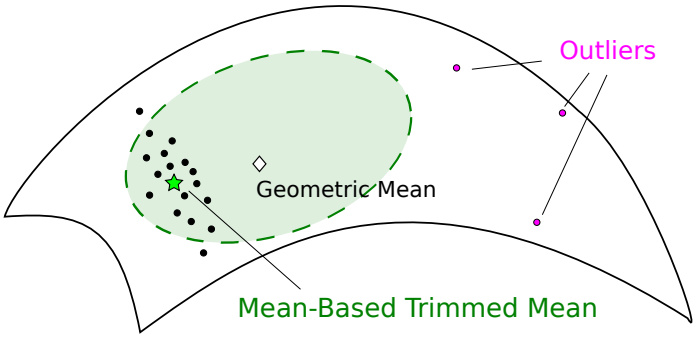

FIG. 4.2. *Illustration of one of the trimmed averages on a Riemannian manifold. The SPD matrices outside of the dashed line are discarded and averaging is taken only over those matrices sitting within the shadowed area.*

5. Experimental results. In order to compare the performances of the discussed algorithms on the averaging of spatial covariance matrices versus the performances of the CSP and the TSLDA method, the following types of algorithms to compute the basis of the reference tangent space were considered:

- **Geometric means**: The arithmetic mean $\mathcal{A}(\boldsymbol{P})$, the Riemannian geometric mean $\mathcal{G}(\boldsymbol{P})$, the Log-Euclidean mean $\mathcal{L}(\boldsymbol{P})$, the harmonic mean $\mathcal{H}(\boldsymbol{P})$ and the resolvent mean $\mathcal{R}(\boldsymbol{P})$;
- **Geometric medians**: The Euclidean geometric median $\mathcal{A}^*(\boldsymbol{P})$, the Riemannian geometric median $\mathcal{G}^*(\boldsymbol{P})$ and the Log-Euclidean geometric median $\mathcal{L}^*(\boldsymbol{P})$;
- **Trimmed averages**: The trimmed Riemannian geometric mean $\hat{\mathcal{G}}(\boldsymbol{P})$, the trimmed Log-Euclidean geometric mean $\hat{\mathcal{L}}(\boldsymbol{P})$, the trimmed Riemannian geometric median $\hat{\mathcal{G}}^*(\boldsymbol{P})$ and the trimmed Log-Euclidean geometric median $\hat{\mathcal{L}}^*(\boldsymbol{P})$.

In addition to such averages, we also used the tangent space at the identity matrix $\boldsymbol{I}$ for comparison purpose.

5.1. Data Description. We used the dataset *JK-HH 1* containing EEG signals recorded during left hand and foot motor-imageries [39].

The EEG signals were recorded from five subjects, denoted as *sa, sb, sc, sd, se*. All the subjects were males, with an average age of 23.2 years with standard deviation of 1.6 years.

During the recording, the subjects performed the motor-imagery tasks being instructed by a visual cue. The cue was given by an arrow on an LCD screen, as shown in the Figure 5.1. The right and down arrows instructed the subjects to perform the motor imagery tasks of the right hand and the foot, respectively. The EEG signals were recorded with Ag/AgCl active electrodes (g.LADYbird, g.LADYbirdGND, and g.GAMMAearclip produced by Guger Technologies) and a power supply (g.GAMMAbox produced by Guger Technologies).

The signal acquisition setup consists in 29 electrodes, which were placed at F3, Fz, F4, FC5, FC3, FC6, FCz, FC2, FC4, FC6, T7, C5, C3, C1, Cz, C2, C4, C6, T8, CP5, CP3, CP1, CPz, CP2, CP4, CP6, P3, Pz, and P4 (the locations refer to the *International 10–10 system* notation [33]). The signals observed from the electrodes were amplified using a bioamplifier (MEG-6116 produced by Nihon Kohden). The amplifier analogically filtered the signals with a passband of $0.5 - 100$ Hz. The sig-

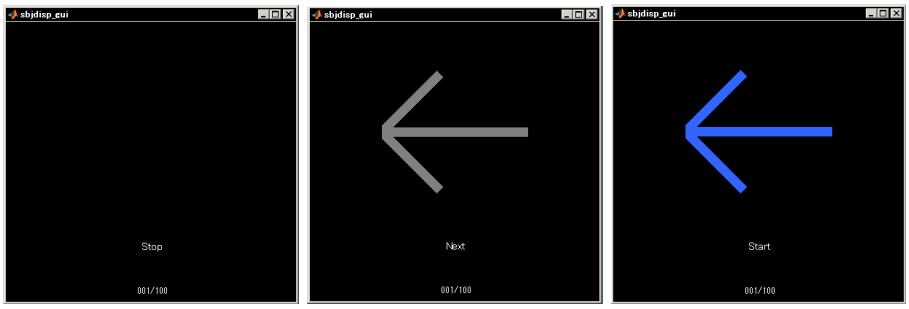

FIG. 5.1. *Illustration of the visual cues to instruct the subjects. The cue appears onscreen in the following order: The left-hand panel means "No task" (0.5 seconds), the central panel means "Notice of the coming task" (0.5 seconds), while the right-hand panel means "Execute the task" (4 seconds). The task to execute is chosen randomly as one of the following (with associated symbol displayed onscreen): "Idle" (symbol displayed onscreen: "+), "Move the left hand" (symbol displayed onscreen: "←"), "Move the right hand" (symbol displayed onscreen: "→"), and "Move both feet" (symbol displayed onscreen: "↓".*

TABLE 5.1
Classification accuracy [%] of CSP using the proposed robust averages in a 10-fold cross-validation

Subject	*sa*	*sb*	*sc*	*sd*	*se*	Mean
Arithmetic mean $\mathcal{A}$	78.50	65.50	57.50	51.00	89.50	68.40
Riemannian geometric mean $\mathcal{G}$	82.50	66.00	61.00	63.00	89.50	72.40
Log-Euclidean mean $\mathcal{L}$	82.50	65.50	62.50	59.50	88.50	71.70
Harmonic mean $\mathcal{H}$	78.50	66.50	58.50	57.50	89.00	70.00
Resolvent mean $\mathcal{R}(a)$	80.00	70.00	58.50	49.50	89.50	69.50
Euclidean geometric median $\mathcal{A}*$	78.00	69.00	57.50	49.00	88.00	68.30
Riemannian geometric median $\mathcal{G}^*$	82.00	63.50	61.00	61.50	89.50	71.50
Log-Euclidean geometric median $\mathcal{L}^*$	79.50	62.50	61.50	62.50	90.00	71.20
Identity matrix	64.00	55.00	47.00	62.50	67.50	59.20

nals through the amplifier were sampled using an A/D converter (AIO-163202F-PE produced by Contec) with a sampling rate of 256 Hz. The converted signals were recorded with the MATLAB's Data Acquisition Toolbox (MathWorks).

As a pre-processing step, we applied to the acquired signals a fourth order Butterworth low-pass filter, with a cutoff frequency of 50 Hz, and downsampled the signals to 128 Hz. In the present experiment, we further applied a band-pass filter whose pass band is 7–30 Hz.

The subjects performed the tasks repeatedly with an interval of around 3 seconds. The dataset for each subject consisted of signals recovered over 100 trials per class, for a total of two classes.

5.2. Results. In order to compare the performances of the tangent-space mapping and that of the common spatial patterns, we first evaluated the performance of averages in the CSP method. The number of the spatial filters is set to 6 ($r = 3$). We used an LDA as a classifier for the log-variance of the filtered signals, and applied a 10-fold cross-validation procedure to compute the resulting classification accuracy.

The table 5.1 shows the results of CSP with various SCMs averages. The arithmetic mean $\mathcal{A}$ is a standard approach to estimate the central matrix out of a set of

TABLE 5.2
Classification accuracy [%] of TSLDA using the proposed robust averages in a 10-fold cross-validation

Subject	*sa*	*sb*	*sc*	*sd*	*se*	Mean
Arithmetic mean $\mathcal{A}$	88.50	78.50	69.50	79.00	93.50	81.80
Riemannian geometric mean $\mathcal{G}$	87.50	80.00	69.00	79.00	92.50	81.60
Log-Euclidean mean $\mathcal{L}$	87.50	77.00	73.50	78.50	92.50	81.80
Harmonic mean $\mathcal{H}$	87.00	79.50	70.50	77.50	93.50	81.60
Resolvent mean $\mathcal{R}(a)$	88.50	80.50	71.00	79.00	94.00	82.60
Euclidean geometric median $\mathcal{A}*$	90.00	76.50	68.00	80.50	94.50	81.90
Riemannian geometric median $\mathcal{G}^*$	88.00	80.50	71.00	81.00	93.50	82.80
Log-Euclidean geometric median $\mathcal{L}^*$	87.50	77.00	71.00	80.00	92.50	81.60
Identity matrix	85.00	77.50	68.50	79.00	93.00	80.60

TABLE 5.3
Classification accuracy [%] of TSLDA using the proposed trimmed averages in a 10-fold cross-validation

Subject	*sa*	*sb*	*sc*	*sd*	*se*	Mean
Trimmed Riemannian mean $\hat{\mathcal{G}}(d)$	91.00	85.00	79.50	84.50	95.50	87.10
Trimmed Log-Euclidean mean $\hat{\mathcal{L}}(d)$	90.50	84.50	78.00	87.50	94.50	87.00
Trimmed Riemannian median $\hat{\mathcal{G}}^*(d)$	91.00	85.00	79.50	84.50	95.50	87.10
Trimmed Log-Euclidean median $\hat{\mathcal{L}}^*(d)$	90.00	85.50	77.50	87.00	95.00	87.00

covariance matrices in the existing contributions. Examining the mean classification accuracy across the five subjects, we see that the accuracy of all reference points, except identity matrix $\boldsymbol{I}$, exceeds that of the reference method $\mathcal{A}$. Overall, the averages $\mathcal{G}, \mathcal{L}, \mathcal{G}^*, \mathcal{L}^*$ show particularly better performances in comparison with the others.

The table 5.2 shows the classification accuracy of the TSLDA method with SCM means and geometric medians, excluding the trimmed averages. We applied a 10-fold cross-validation procedure to compute the classification accuracy. In comparison to the result of CSP, TSLDA shows clearly higher classification results for all averaging methods. The difference in accuracy between all the considered averages is little (including in the case of the identity matrix). This result suggests that the distances between such averages are small on the Riemannnian manifold of symmetric, positive-definite matrices.

Likewise, the Table 5.3 shows the results of applying TSLDA using any of the trimmed averages as the reference points for the tangent spaces. This illustrated classification accuracies show how the trimmed averages exhibit significantly better performances than non-trimmed averages. This result suggests that the trimmed averaging technique can estimate the covariance matrix more properly by eliminating the trials that are considered as outliers.

6. Conclusions. Generally, the EEG data include outliers, therefore, the geometric mean is not always the optimal estimator. In this paper, we proposed to use more robust estimators than the geometric mean, in order to improve the classification performances of tangent-spaces linear-discriminant analysis. To cope with such a problem, we proposed using the geometric median and the geometric trimmed averages.

In the experimental evaluation, while the geometric median showed little improvement over the geometric mean, the trimmed averaging techniques exhibited clear improvement.

REFERENCES

[1] K. K. Ang and C. Guan, *Brain-computer interface in stroke rehabilitation*, Journal of computing science and engineering, 7 (2013), pp. 139–146.

[2] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, *Geometric means in a novel vector space structure on symmetric positivedefinite matrices*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 328–347.

[3] G. Arulmozhi, *Statistics For Management*, Tata McGraw-Hill Education, second edition ed., 2009.

[4] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, *Multiclass brain-computer interface classification by Riemannian geometry*, IEEE Transactions on Biomedical Engineering, 59 (2012), pp. 920–928.

[5] ———, *Classification of covariance matrices using a Riemannian-based kernel for BCI applications*, Neurocomputing, 112 (2013), pp. 172–178.

[6] D. Bartz and K.-R. Müller, *Generalizing analytic shrinkage for arbitrary covariance structures*, Advances in Neural Information Processing Systems, (2013), pp. 1869–1877.

[7] H. H. Bauschke, S. M. Moffat, and X. Wang, *The resolvent average for positive semidefinite matrices*, Linear Algebra and Its Applications, 432 (2010), pp. 1757–1771.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2005.

[9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, *Optimizing spatial filters for robust EEG single-trial analysis*, IEEE Signal Processing Magazine, 25 (2008), pp. 41–56.

[10] P. Boldi and S. Vigna, *Axioms for centrality*, Internet Mathematics, 10 (2014), pp. 222–262.

[11] C. A. Castaño-Moraga, C. Lenglet, R. Deriche, and J. Ruiz-Alzola, *A Riemannian approach to anisotropic filtering of tensor fields*, Signal Processing, 87 (2007), pp. 263–276.

[12] A. Cherian and S. Sra, *Riemannian sparse coding for positive definite matrices*, in Computer Vision – ECCV 2014, vol. 8691, jul 2014, pp. 299–314.

[13] G. Dornhege, J. R. Millán, T. Hinterberger, D. Mcfarland, and K.-R. Müller, *Towards Brain-Computer Interfacing*, The MIT press, 2007.

[14] S. Fiori, *Learning the Fréchet mean over the manifold of symmetric positive-definite matrices*, Cognitive Computation, 1 (2009), pp. 279–291.

[15] ———, *On vector averaging over the unit hyphersphere*, Digital Signal Processing, 19 (2009), pp. 715–725.

[16] ———, *Kolmogoroff-nagumo mean over the affine symplectic group of matrices*, Applied Mathematics and Computation, 266 (2015), pp. 820–837.

[17] S. Fiori and T. Tanaka, *An algorithm to compute averages on matrix Lie groups*, IEEE Transactions on Signal Processing, 57 (2009), pp. 4734–4743.

[18] P. T. Fletcher and S. Joshi, *Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors*, in ECCV Workshops CVAMIA and MMBIA, vol. 3117, Berlin, Heidelberg, 2004, pp. 87–98.

[19] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi, *The geometric median on Riemannian manifolds with application to robust atlas estimation*, NeuroImage, 45 (2009), pp. S143–S152.

[20] J. H. Friedman, *Regularized discriminant analysis*, Journal of the American Statistical Association, 84 (1989), pp. 165–175.

[21] Y. Fu, J. Gao, X. Hong, and D. Tien, *Low rank representation on Riemannian manifold of symmetric positive definite matrices*, in Proceedings of the 2015 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, Philadelphia,

PA, jun 2015, pp. 316–324.
[22] M. Grosse-Wentrup and M. Buss, *Multiclass common spatial patterns and information theoretic feature extraction*, IEEE Transactions on Biomedical Engineering, 55 (2008), pp. 1991–2000.
[23] H. Hotelling, *Relations between two sets of variates*, Biometrika, 28 (1936), pp. 321–377.
[24] H. Karcher, *Riemannian center of mass and mollifier smoothing*, Communications on Pure and Applied Mathematics, 30 (1977), pp. 509–541.
[25] S. Kim, J. Lawson, and Y. Lim, *The matrix geometric mean of parameterized, weighted arithmetic and harmonic means*, Linear Algebra and its Applications, 435 (2011), pp. 2114–2131.
[26] A. Kübler, F. Nijboer, J. Mellinger, T. M. Vaughan, H. Pawelzik, G. Schalk, D. J. McFarland, N. Birbaumer, and J. R. Wolpaw, *Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface*, Neurology, 64 (2005), pp. 1775–1777.
[27] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365–411.
[28] P. Li and Q. Wang, *Local log-Euclidean covariance matrix (L2ECM) for image representation and its applications*, Computer Vision – ECCV 2012, 7574 (2012), pp. 469–482.
[29] A. Llera, V. Gómez, and H. J. Kappen, *Adaptive multiclass classification for brain computer interfaces*, Neural Computation, 26 (2014), pp. 1108–1127.
[30] F. Lotte and C. Guan, *Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms*, IEEE Transactions on Biomedical Engineering, 58 (2011), pp. 355–362.
[31] M. Moakher, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM Journal on Matrix Analysis and Applications, 26 (2005), pp. 735–747.
[32] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, *Designing optimal spatial filters for single-trial EEG classification in a movement task*, Clinical Neurophysiology, 110 (1999), pp. 787–798.
[33] R. Oostenveld and P. Praamstra, *The five percent electrode system for high-resolution EEG and ERP measurements*, Clinical Neurophysiology, 112 (2001), pp. 713–719.
[34] A. Pascual-Leone, A. Amedi, F. Fregni, and L. B. Merabet, *The plastic human brain cortex*, Annual Review of Neuroscience, 28 (2005), pp. 377–401.
[35] G. Pfurtscheller, G. Muller-Putz, R. Scherer, and C. Neuper, *Rehabilitation with brain-computer interface systems*, Computer, 41 (2008), pp. 58–65.
[36] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, *Optimal spatial filtering of single trial EEG during imagined hand movement*, IEEE Transactions on Rehabilitation Engineering, 8 (2000), pp. 441–446.
[37] W. Samek and M. Kawanabe, *Robust common spatial patterns by minimum divergence covariance estimator*, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), may 2014, pp. 2040–2043.
[38] S. Sanei and J. Chambers, *EEG Signal Processing*, John Wiley & Sons Ltd,, West Sussex, England, sep 2007.
[39] N. Tomida, T. Tanaka, S. Ono, M. Yamagishi, and H. Higashi, *Active data selection for motor imagery EEG classification*, IEEE Transactions on Biomedical Engineering, 62 (2015), pp. 458–467.
[40] N. Tomida, M. Yamagishi, I. Yamada, and T. Tanaka, *A reduced rank approach for covariance matrix estimation in EEG signal classification*, in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, aug 2014, pp. 668–671.
[41] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Mathematical Journal, 43 (1937), pp. 355–386.
[42] Y. Xinyi, R. K. Ward, and G. E. Birch, *Robust common spatial patterns for EEG signal preprocessing*, in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, aug 2008, pp. 2087–2090.
[43] F. Yger, F. Lotte, and M. Sugiyama, *Averaging covariance matrices for EEG signal classification based on the CSP: An empirical study*, in 23rd European Signal Processing Conference (EUSIPCO), aug 2015, pp. 2721–2725.