Multiclass Brain–Computer Interface Classification by Riemannian Geometry

Alexandre Barachant*, Stéphane Bonnet, Marco Congedo, and Christian Jutten

***Abstract*—This paper presents a new classification framework for brain–computer interface (BCI) based on motor imagery. This framework involves the concept of Riemannian geometry in the manifold of covariance matrices. The main idea is to use spatial covariance matrices as EEG signal descriptors and to rely on Riemannian geometry to directly classify these matrices using the topology of the manifold of symmetric and positive definite (SPD) matrices. This framework allows to extract the spatial information contained in EEG signals without using spatial filtering. Two methods are proposed and compared with a reference method [multiclass Common Spatial Pattern (CSP) and Linear Discriminant Analysis (LDA)] on the multiclass dataset IIa from the BCI Competition IV. The first method, named minimum distance to Riemannian mean (MDRM), is an implementation of the minimum distance to mean (MDM) classification algorithm using Riemannian distance and Riemannian mean. This simple method shows comparable results with the reference method. The second method, named tangent space LDA (TSLDA), maps the covariance matrices onto the Riemannian tangent space where matrices can be vectorized and treated as Euclidean objects. Then, a variable selection procedure is applied in order to decrease dimensionality and a classification by LDA is performed. This latter method outperforms the reference method increasing the mean classification accuracy from 65.1% to 70.2%.**

***Index Terms*—Brain computer interfaces, classification algorithms, covariance matrix, electroencephalography, information geometry.**

I. Introduction

A brain–computer interface (BCI) aims at translating brain signals into commands. On one hand, it will be useful for people with severe motor impairments in order to restore communication and movement, on the other it could be a new interface for healthy people, e.g., in gaming applications.

In the literature several electrophysiological sources of BCI control have been investigated. This paper focuses on motor imagery (MI), although the method we propose herein may be applied to other types of BCIs. In MI-based EEG-BCI, the standard operation mode consists of a cue-based (or synchronous) calibration stage followed by an asynchronous operation mode [1]. The first stage is often devoted to build subject-dependent frequency and spatial filters discriminating between EEG datasets corresponding to two different classes of MI [2]. Such spatial filters perform linear combination of the EEG signals in order to create new signals with maximal variance in one condition and minimal variance in the other condition [2]. Once these spatial filters have been designed, the (log-)variance of the spatially-filtered signals are used as features by a supervised classification algorithm. Linear discriminant analysis (LDA) is often used to perform this processing [3].

In this work, we propose an entirely different approach by exploiting directly the *covariance structure of the data* as the feature of interest. Since the covariance matrices contain the spatial information embedded in EEG signal, this approach aims to merge the spatial filtering and the classification procedure into one unique step. This idea has already been proposed in [4] where the spatial filters and the classifier are trained together in an optimization process. However, the particular structure of covariance matrices, which belong to the Riemannian manifold of the symmetric and positive definite (SPD) matrices, has to be treated carefully. In this context, the Riemannian geometry provides a rich framework to manipulate these matrices. In addition, the properties of the manifold of SPD matrices allow to have explicit formulas for the main operations in the Riemannian manifold, which lead to an easier implementation of algorithms.

A similar approach has also been recently developed in the context of EEG for sleep states detection [5] that focused on the exploitation of the frequential information through the estimation of the power spectral covariance matrices. Working with covariance matrices in their manifold has proven useful in other fields such as radar image processing [6], diffusion tensor imaging [7], and image processing [8]. This methodology has also been used for pedestrian detection in image processing [8].

The first idea of our paper is to manipulate EEG spatial covariance matrices in their native space and to make use of the Riemannian distance between them in this space. Spatial filtering is no more necessary, since the spatial information is embedded in the spatial covariance matrices. Therefore, all distance-based classification algorithms can be extended by applying this new Riemannian distance.

The second idea we propose is to use the concept of tangent space to apply more sophisticated classification algorithms. The operation named *tangent space mapping* sends the covariance matrices, belonging to a manifold, into an Euclidean space where they can be treated as vectors. This mapping operation

Manuscript received March 9, 2011; revised May 12, 2011; accepted June 16, 2011. Date of publication October 14, 2011; date of current version March 21, 2012. The work of M. Congedo and C. Jutten was supported in part by the ANR (French National Research Agency) under Project OpenViBE2 and Project RoBIK. *Asterisk indicates corresponding author.*

*A. Barachant is with the CEA-LETI, MINATEC Campus, 38054 Grenoble, France (e-mail: alexandre.barachant@gmail.com).

S. Bonnet is with the CEA-LETI, MINATEC Campus, 38054 Grenoble, France (e-mail: stephane.bonnet@cea.fr).

M. Congedo and C. Jutten are with the GIPSA-lab, CNRS, Grenoble University, 38402 Saint Martin d'Hères, France (e-mail: marco.congedo@gipsa-lab.grenoble-inp.fr; christian.jutten@gipsa-lab.grenoble-inp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2011.2172210

allows the use of state-of-the-art classifiers within the Riemannian framework.

The proposed methods are developed based on an earlier work in [9]. Here we address the multiclass case and the high dimensionality of the tangent space.

This article is organized as follows: First we discuss about the use of the covariance matrices in motor imagery-based BCI. Then, in Section III we introduce the basic concepts of Riemannian geometry. Section IV presents the two proposed classification methods, and Section V is dedicated to the related results on a BCI dataset. Finally, we conclude on the proposed approach in Section VI.

II. On the Use of Covariance Matrices in BCI

By definition, motor imagery consists in the imagination of a limb movement like hand or foot. The topographical representation and band power change of such mental tasks are well-known since different body parts are represented in different area of the motor cortex (roughly, right hand under C_3 electrode, left hand under C_4, foot under C_z, etc.). The associated frequential phenomena are known as event-related desynchronisation (ERD) and event-related synchronisation (ERS) [10].

The EEG recording of the brain activity is usually cut into trials, i.e., short-time windows. During the calibration operation mode, the trials are supervised, in the sense that the MI mental task or *class* is known. In the functional operation mode, single-trial classification must be performed to recognize the different mental tasks.

Spatial covariance matrices are used in this context. For instance, the popular common spatial pattern (CSP) algorithm is entirely based on their estimation, from which spatial filters are derived to enhance class separability [2].

In the following, let $\mathbf{x}_t \in \mathbb{R}^n$ denote the EEG signal vector at a specific time point t, with n denoting the number of recording channels. The spatial covariance matrix is formally defined by $\mathbf{\Sigma} = E\{(\mathbf{x}_t - E\{\mathbf{x}_t\})(\mathbf{x}_t - E\{\mathbf{x}_t\})^T\}$, where superscript ^{T} denotes vector or matrix transposition and $E\{.\}$ the expected value.

In BCI we consider short-time segments of EEG signal, or trials, in the form of a matrix $\mathbf{X}_i = [\mathbf{x}_{t+T_i} \dots \mathbf{x}_{t+T_i+T_s-1}] \in \mathbb{R}^{n \times T_s}$ which corresponds to the i-th trial of imagined movement started at time $t = T_i$. Here T_s denotes the number of sampled time points in each trial. Here and hereafter, we suppose that each channel measurement has been previously time-centered by a high-pass filtering operation. For the i-th trial, the spatial covariance matrix is estimated using the *Sample Covariance Matrix* (SCM) $\mathbf{P}_i \in \mathbb{R}^{n \times n}$ such as

$$\mathbf{P}_i = \frac{1}{T_s - 1} \mathbf{X}_i \mathbf{X}_i^T. \tag{1}$$

The SCM is known to be an unbiased estimator of the covariance matrix $\mathbf{\Sigma}$ provided that the number of observations T_s is much larger than the number of variables n [11].

III. Riemannian Geometry

In this section,we briefly define the concepts and tools of Riemannian geometry that are needed in the proposed approach. The main point is that spatial SCM belongs to a particular manifold and classification can be directly achieved in this Riemannian space, as will be shown in Section IV.

A. Notations

Denote by $S(n) = \{\mathbf{S} \in M(n), \mathbf{S}^T = \mathbf{S}\}$ the space of all $n \times n$ symmetric matrices in the space of square real matrices $M(n)$ and $P(n) = \{\mathbf{P} \in S(n),\ \mathbf{u}^T \mathbf{P} \mathbf{u} > 0,\ \forall\, \mathbf{u} \in \mathbb{R}^n\}$ the set of all $n \times n$ symmetric positive-definite (SPD) matrices. A SPD matrix is always diagonalizable with strictly real positive eigenvalues. Finally, denote by $Gl(n)$ the set of all $n \times n$ invertible matrices in $M(n)$.

The Frobenius norm of a matrix is given by: $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T) = \sum |A_{ij}|^2$ where $\mathrm{Tr}(.)$ denotes the trace operator.

For a vector $\mathbf{a}$, the $\mathcal{L}_2$ norm is denoted by $\|\mathbf{a}\|_2$.

For SPD matrices in $P(n)$, the exponential matrix of $\mathbf{P}$ is obtained using the eigenvalue decomposition of $\mathbf{P}$:

$$\mathbf{P} = \mathbf{U}\ \mathrm{diag}\left(\sigma_1, \dots, \sigma_n\right) \mathbf{U}^T$$

where $\sigma_1 > \sigma_2 > \dots \sigma_n > 0$ are the eigenvalues and $\mathbf{U}$ the matrix of eigenvectors of $\mathbf{P}$. It reads:

$$\exp(\mathbf{P}) = \mathbf{U}\ \mathrm{diag}\left(\exp(\sigma_1), \dots, \exp(\sigma_n)\right)\ \mathbf{U}^T.$$

The inverse operation is the logarithm of a SPD matrix:

$$\log(\mathbf{P}) = \mathbf{U}\ \mathrm{diag}\left(\log(\sigma_1), \dots, \log(\sigma_n)\right)\ \mathbf{U}^T.$$

We also have the following properties:

1) $\forall\, \mathbf{P} \in P(n),\ \det(\mathbf{P}) > 0$
2) $\forall\, \mathbf{P} \in P(n),\ \mathbf{P}^{-1} \in P(n)$
3) $\forall\, (\mathbf{P}_1, \mathbf{P}_2) \in P(n)^2,\ \mathbf{P}_1 \mathbf{P}_2 \in P(n)$
4) $\forall\, \mathbf{P} \in P(n),\ \log(\mathbf{P}) \in S(n)$
5) $\forall\, \mathbf{S} \in S(n),\ \exp(\mathbf{S}) \in P(n)$

Finally, notation $\mathbf{P}^{1/2}$ defines a symmetric matrix $\mathbf{A}$ that fulfills the relation $\mathbf{A}\mathbf{A} = \mathbf{P}$.

B. Riemannian Natural Metric

The space of SPD matrices $P(n)$ is a differentiable Riemannian manifold $\mathcal{M}$ [12]. The derivatives at a matrix $\mathbf{P}$ on the manifold lies in a vector space $T_{\mathbf{P}}$, which is the *tangent space* at that point. The tangent space is lying in the space $S(n)$. The manifold and the tangent space are $m = n(n+1)/2$ dimensional.

Each tangent space has an inner product $\langle , \rangle_{\mathbf{P}}$ that varies smoothly from point to point over the manifold. The *natural metric* on the manifold of SPD matrices is defined by the local inner product:

$$\langle \mathbf{S}_1, \mathbf{S}_2 \rangle_{\mathbf{P}} = \mathrm{Tr}\left(\mathbf{S}_1 \mathbf{P}^{-1} \mathbf{S}_2 \mathbf{P}^{-1}\right). \tag{2}$$

The inner product induces a norm for the tangent vectors on the tangent space, such that, $\|\mathbf{S}\|_{\mathbf{P}}^2 = \langle \mathbf{S}, \mathbf{S} \rangle_{\mathbf{P}} = \mathrm{Tr}\left(\mathbf{S}\mathbf{P}^{-1}\mathbf{S}\mathbf{P}^{-1}\right)$. We note that, at identity matrix, such norm simplifies into the Frobenius norm, i.e., $\langle \mathbf{S}, \mathbf{S} \rangle_{\mathbf{I}} = \|\mathbf{S}\|_F^2$.

Working in the manifold $\mathcal{M}$ of the SPD matrices has several advantages. Due to the properties of SPD matrices, there exists explicit formulae for the major concepts of Riemannian geometry. Therefore, they can be applied easily in the context of signal processing.

C. Riemannian Geodesic Distance

Let $\mathbf{\Gamma}(t) : [0,1] \rightarrow P(n)$ be any (differentiable) path from $\Gamma(0) = \mathbf{P}_1$ to $\Gamma(1) = \mathbf{P}_2$. The length of $\mathbf{\Gamma}(t)$ is given by

$$L(\mathbf{\Gamma}(t)) = \int_0^1 \|\dot{\mathbf{\Gamma}}(t)\|_{\mathbf{\Gamma}(t)} dt \tag{3}$$

with the norm defined previously. The minimum length curve connecting two points on the manifold is called *the geodesic*, and the Riemannian distance between the two points is given by the length of this curve. The natural metric (2) induces the geodesic distance [13]

$$\delta_R(\mathbf{P}_1, \mathbf{P}_2) = \|\log\left(\mathbf{P}_1^{-1}\mathbf{P}_2\right)\|_F = \left[\sum_{i=1}^{n} \log^2 \lambda_i\right]^{1/2} \tag{4}$$

where $\lambda_i, i = 1 \ldots n$ are the real eigenvalues of $\mathbf{P}_1^{-1}\mathbf{P}_2$. The main properties of the Riemannian geodesic distance are as follows:

1) $\delta_R(\mathbf{P}_2, \mathbf{P}_1) = \delta_R(\mathbf{P}_1, \mathbf{P}_2)$
2) $\delta_R(\mathbf{P}_1^{-1}, \mathbf{P}_2^{-1}) = \delta_R(\mathbf{P}_1, \mathbf{P}_2)$
3) $\delta_R(\mathbf{W}^T\mathbf{P}_1\mathbf{W}, \mathbf{W}^T\mathbf{P}_2\mathbf{W}) = \delta_R(\mathbf{P}_1, \mathbf{P}_2)\ \forall\ \mathbf{W} \in Gl(n)$.

The third property is very important, since it implies that the space of SPD matrices is invariant by projection. It allows us to manipulate such space with tools like PCA without incidence on the distance.

D. Exponential Map

For each point $\mathbf{P} \in P(n)$, we can thus define a tangent space composed by the set of tangent vectors at $\mathbf{P}$. Each tangent vector $\mathbf{S}_i$ can be seen as the derivative at $t = 0$ of the geodesic $\mathbf{\Gamma}_i(t)$ between $\mathbf{P}$ and the exponential mapping $\mathbf{P}_i = \mathrm{Exp}_{\mathbf{P}}(\mathbf{S}_i)$, defined as

$$\mathrm{Exp}_{\mathbf{P}}(\mathbf{S}_i) = \mathbf{P}_i = \mathbf{P}^{\frac{1}{2}} \exp\left(\mathbf{P}^{-\frac{1}{2}}\mathbf{S}_i\mathbf{P}^{-\frac{1}{2}}\right)\mathbf{P}^{\frac{1}{2}}. \tag{5}$$

The inverse mapping is given by the logarithmic mapping defined as

$$\mathrm{Log}_{\mathbf{P}}(\mathbf{P}_i) = \mathbf{S}_i = \mathbf{P}^{\frac{1}{2}} \log\left(\mathbf{P}^{-\frac{1}{2}}\mathbf{P}_i\mathbf{P}^{-\frac{1}{2}}\right)\mathbf{P}^{\frac{1}{2}}. \tag{6}$$

This procedure is described geometrically in Fig. 1.

Equivalent definitions of the Riemannian distance are

$$\begin{aligned}\delta_R(\mathbf{P}, \mathbf{P}_i) &= \|\mathrm{Log}_{\mathbf{P}}(\mathbf{P}_i)\|_{\mathbf{P}} = \|\mathbf{S}_i\|_{\mathbf{P}} \\ &= \|\mathrm{upper}(\mathbf{P}^{-\frac{1}{2}}\mathrm{Log}_{\mathbf{P}}(\mathbf{P}_i)\mathbf{P}^{-\frac{1}{2}})\|_2 = \|\mathbf{s}_i\|_2\end{aligned} \tag{7}$$

where the $\mathrm{upper}(.)$ operator consists in keeping the upper triangular part of a symmetric matrix and vectorizing it by applying unity weight for diagonal elements and $\sqrt{2}$ weight for out-of-diagonal elements [8]. Here $\mathbf{s}_i$ is the m-dimensional vector $\mathrm{upper}\left(\mathbf{P}^{-\frac{1}{2}}\mathrm{Log}_{\mathbf{P}}(\mathbf{P}_i)\mathbf{P}^{-\frac{1}{2}}\right)$ of the normalized tangent space.

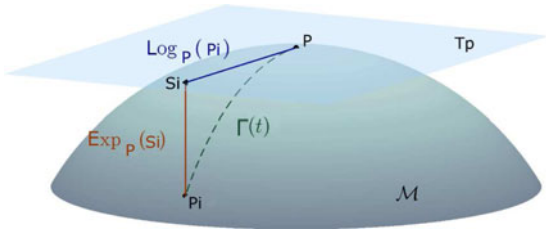

Fig. 1. Tangent space at point $\mathbf{P}$, $\mathbf{S}_i$ a tangent vector at $\mathbf{P}$ and $\mathbf{\Gamma}_i(t)$ the geodesic between $\mathbf{P}$ and $\mathbf{P}_i$.

Under some conditions over $\mathbf{P}$ and the $\mathbf{P}_i$ [8], (7) leads to an approximation in terms of distance between the tangent space and the Riemannian manifold, such as

$$\forall i, j \quad \delta_R(\mathbf{P}_i, \mathbf{P}_j) \approx \|\mathbf{s}_i - \mathbf{s}_j\|_2. \tag{8}$$

In order to verify these conditions the $\mathbf{P}_i$ must be locally distributed into the manifold, i.e., located into a small part of the manifold and $\mathbf{P}$ must be the mean of the $\mathbf{P}_i$.

E. Mean of SPD Matrices

1) Euclidean Mean: Using the Euclidean distance on $\mathcal{M}(n)$, $\delta_E(\mathbf{P}_1, \mathbf{P}_2) = \|\mathbf{P}_1 - \mathbf{P}_2\|_F$, it is possible to define the Euclidean mean of $I \geq 1$ SPD matrices by

$$\mathfrak{A}(\mathbf{P}_1, \ldots, \mathbf{P}_I) = \arg\min_{\mathbf{P} \in P(n)} \sum_{i=1}^{I} \delta_E^2(\mathbf{P}, \mathbf{P}_i) = \frac{1}{I}\sum_{i=1}^{I} \mathbf{P}_i. \tag{9}$$

Such mean is also referred to as the *arithmetic* mean in the literature [13] and it has been used ubiquitously so far.

2) Riemannian Mean: Using Riemannian geodesic distance (4) the Riemannian mean of $I \geq 1$ SPD matrices is given by

$$\mathfrak{G}(\mathbf{P}_1, \ldots, \mathbf{P}_I) = \arg\min_{\mathbf{P} \in P(n)} \sum_{i=1}^{I} \delta_R^2(\mathbf{P}, \mathbf{P}_i). \tag{10}$$

This mean is also referred to as the *geometric* mean. Indeed, if we consider 1×1 SPD matrices $\{x_i > 0\}_{1 \leq i \leq I}$, this definition gives $\mathfrak{G}(x_1, \ldots, x_I) = \arg\min_{x>0} \sum_{i=1}^{I} \log^2(x_i/x) = \sqrt[I]{x_1 \ldots x_I}$.

For a manifold of nonpositive sectional curvature like $P(n)$, such local minimum exists and is unique [14]. However, there is no closed-form expression to compute the mean and optimisation algorithms must be employed. An efficient iterative algorithm to compute the Riemannian mean of I SPD matrices is given in [7].

IV. MI-based BCI Classification Using Riemannian Framework

During the calibration operation mode, a set of annotated trials is obtained for each class of motor imagery. Suppose, we have a set of trials obtained for the k-th condition $\{\mathbf{X}_i, i \in \mathfrak{I}^{(k)}\}$ with $\mathfrak{I}^{(k)}$ the set of indices of the trials corresponding to the k-th condition. The SCM of each trial can be computed using (1). Since SCMs are SPD matrices, they do belong to the manifold $\mathcal{M}$. Let $\mathbf{P}$ denote the SCM of the unlabel trial $\mathbf{X}$.

It is possible, using the results in Section III, to derive several efficient classification algorithms to estimate the unknown label of trial $\mathbf{X}$.

A. Classification in the Riemannian Manifold

1) Minimum Distance to Riemannian Mean (MDRM): The intraclass covariance matrices for each condition $\mathbf{P}_{\mathfrak{G}}^{(k)}$, where $k = [1 : K]$ denotes the class indices, can be computed using the results in Section III-E. A very simple *supervised* classification algorithm consists in computing the Riemannian distance between the unknown SCM and every intraclass covariance matrix $\mathbf{P}_{\mathfrak{G}}^{(k)}$. The condition with minimum distance is affected to the unknown trial $\mathbf{X}$. For each new trial $\mathbf{X}$, this procedure amounts to estimating K intraclass covariance matrices and then compute K eigenvalue decompositions of (4).

Algorithm 1 Minimum Distance to Riemannian Mean

Input: a set of trials $\mathbf{X}_i$ of K different known classes.
Input: $\mathbf{X}$ an EEG trial of unknown class.
Input : $\Im^{(k)}$ the set of indices of the trials corresponding to the k-th condition.
Output: $\hat{k}$ the estimated class of test trial $\mathbf{X}$.

```
Compute SCMs of X_i to obtain P_i, (1).
Compute SCM of X to obtain P, (1).
for k = 1 to K do
    P_G^(k) = G(P_i, i ∈ I^(k)), (10).
end for
k̂ = arg min_k δ_R(P, P_G^(k)), (4).
return k̂
```

B. Classification in the Riemannian Tangent Space

Many popular and efficient classification algorithms [15] (LDA, SVM, Neural Network) are based on projections into hyperplanes. As such, they cannot be implemented directly in Riemannian manifold. The implementation of more sophisticated classification algorithms can be readily obtained with the use of the tangent space located at the geometric mean of the whole set of trials: $\mathbf{P}_{\mathfrak{G}} = \mathfrak{G}(\mathbf{P}_i,\ i = 1 \dots I)$. Each SCM $\mathbf{P}_i$ is then mapped into this tangent space, to yield the set of $m = n(n+1)/2$ dimensional vectors:

$$\mathbf{s}_i = \text{upper}\left(\mathbf{P}_{\mathfrak{G}}^{-\frac{1}{2}} \text{Log}_{\mathbf{P}_{\mathfrak{G}}}(\mathbf{P}_i)\, \mathbf{P}_{\mathfrak{G}}^{-\frac{1}{2}}\right). \tag{11}$$

Algorithm 2 Tangent Space Mapping

Input: a set of I SPD matrices $\mathbf{P}_i \in P(n)$
Output: a set of I vectors $\mathbf{s}_i$.

```
P_G = G(P_i, i = 1...I) {Compute Riemannian mean of the whole set, (10). }
for i = 1 to I do
    s_i = upper( P_G^{-1/2} Log_{P_G}(P_i) P_G^{-1/2} )
end for
return s_i
```

1) Variable Selection: Since the tangent space is a $m = n(n+1)/2$ dimensional space, the number of dimensions may now exceed the number of trials of each class. Regularised classification algorithms are usually needed in this case to tackle this problem [3], [16].

An alternative solution that we have explored is to consider a variable selection procedure to decrease the space dimensionality. We use a one-way ANOVA to select the most discriminant variables. To account for any correlation between variables, the vectors $\mathbf{s}_i$ are first orthogonalized using a singular value decomposition (SVD)

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{S} = [\mathbf{s}_1 \dots \mathbf{s}_I] \in \mathbb{R}^{m \times I}$, $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{I \times I}$ two orthogonal matrices and $\mathbf{\Lambda} \in \mathbb{R}^{m \times I}$ a diagonal matrix composed by the singular values of $\mathbf{S}$. The tangent space $\mathbf{S}$ is projected using the orthogonal basis $\mathbf{U}$

$$\mathbf{S}_o = \mathbf{U}^T\mathbf{S}$$

to obtain the orthogonalized tangent space $\mathbf{S}_o$ where all variables are uncorrelated. This operation is known as principal component analysis (PCA).

Then, for each variable of $\mathbf{S}_o$, a one-way ANOVA is applied. All variables are ranked according to their p-values and a weighted false discovery rate (FDR) [17] is applied to select automatically the minimal number of variables. The singular values are used as weight for the FDR procedure in order to give priority to variables that best represent the structure of the data.

2) Tangent Space Linear Discriminant Analysis (TSLDA): After the variable selection, we can apply any kind of classical classification algorithms since the tangent space is Euclidean. In this paper, the *classify* function of Matlab (The MathWorks, inc, Natick MA) will be used. This function performs the linear discriminant analysis of Fisher between each pair of class. Then, the $\frac{K(K-1)}{2}$ classifiers are combined to obtain the classification results.

V. Results

A. Description of Data

We analyze the dataset IIa from the BCI competition IV [18]. It contains EEG data from nine subjects who perform four kinds of motor imagery (right hand, left hand, foot, and tongue imagined movements). EEG signals are recorded using 22 electrodes. For each subject, a training set and a test set are available. In this paper, both sets are concatenated and performances are assessed by means of a cross-validation procedure. A total of 576 trials per subject are available (144 trials per class). The same preprocessing step is applied on the whole dataset. The EEG signals are bandpass filtered by a 5-th order Butterworth filter in the $8 - 30$ Hz frequency band. The time interval is restricted to the time segment comprised between 0.5 and 2.5 s after the cue instructing the user to perform the mental task.

B. Results

1) Minimum Distance to Riemannian Mean: Let us observe the results of the MDRM (Algorithm 1) algorithm. For illustration purpose, we consider the case of the subject S1, who achieves fair performances. In order to show

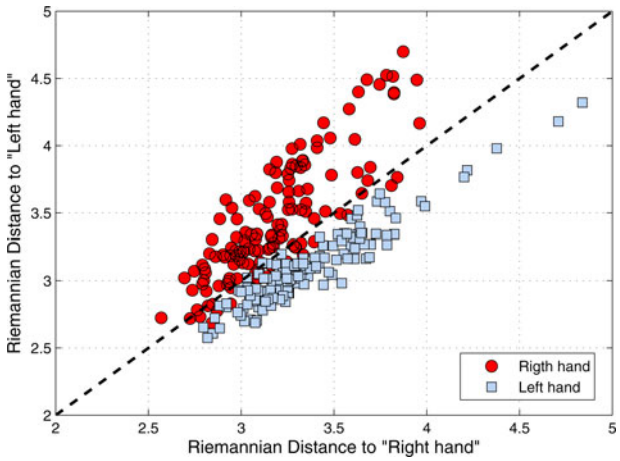

Fig. 2. Riemannian distance to Riemannian mean for two class-related mean covariance matrices—case of right hand versus left hand.

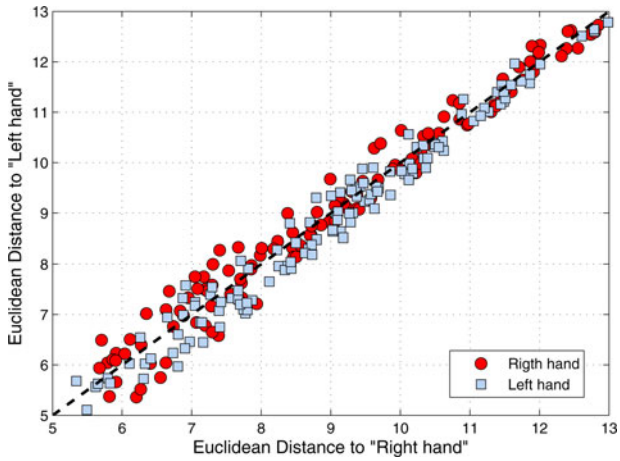

Fig. 3. Euclidean distance to Euclidean mean for two class-related mean covariance matrices—case of right hand versus left hand.

interpretable figures, we reduce the dataset to a two-class case, considering the data from the right hand and left hand MI realisations.

Fig. 2 shows the distance of each trial from the two class-related mean covariance matrices $\mathbf{P}_{\mathfrak{G}}^{(1)}$ and $\mathbf{P}_{\mathfrak{G}}^{(2)}$. On the abscissa, we have the distance from the right hand mean, i.e., $\delta_R(\mathbf{P}_{\mathfrak{G}}^{(1)}, \mathbf{P}_i)$. On the ordinate, we have the distance from the left hand mean, i.e., $\delta_R(\mathbf{P}_{\mathfrak{G}}^{(2)}, \mathbf{P}_i)$. The dashed line represent the decision border whereupon $\delta_R(\mathbf{P}_{\mathfrak{G}}^{(1)}, \mathbf{P}_i) = \delta_R(\mathbf{P}_{\mathfrak{G}}^{(2)}, \mathbf{P}_i)$. In this instance, only a small percentage of trials are misclassified and the use of Riemannian distance to mean is effective.

To understand the benefit of the Riemannian framework, the same process is applied using both Euclidean distance and Euclidean mean, as expressed in (9). Fig. 3 illustrates the obtained results. The trials are distributed around the dashed line regardless of their class membership. This fact implies that the relevant information about class membership is not accessible using the Euclidean distance and the associated mean.

2) Tangent Space: We need to check the approximation, in terms of distance between the Riemannian manifold and the tangent space, as stated by (8). It can be empirically verified looking at the normalized pairwise errors given by

$$\epsilon_{ij} = \frac{|\delta_R(\mathbf{P}_i, \mathbf{P}_j) - \|\mathbf{s}_i - \mathbf{s}_j\|_2|}{\delta_R(\mathbf{P}_i, \mathbf{P}_j)}. \tag{12}$$

The average and standard deviation of ϵ_{ij} across the 9 subjects are 2% and 0.6%, respectively. We conclude that approximation (8) is verified in these data.

Then we apply Algorithm 2 to map the data in the Tangent space. It is interesting to observe the distribution of each class across trials. Fig. 4 illustrates these distributions for the most discriminant variable, selected using an ANOVA over the m variables of the tangent space. Clearly, all classes are not separated equally well. We can distinguish two groups. On one side, the trials corresponding to both hands and on the other side the trials corresponding to the foot and tongue. This phenomenon is due to physiological reasons, as it was discussed in [19]. Also, for this variable, the trials corresponding to the left hand are almost fully separated from those corresponding to the tongue movement. Obviously, increasing the number of variables will bring supplementary information useful to treat the other cases.

The CSP algorithm is widely used in BCI. If we observe the distributions of features corresponding to the best CSP spatial filter (Fig. 5) and selected using the methods described in [20], we found the same pattern of distributions as observed in Fig. 4. This is not surprising, according to our work in [21]; we have shown that the CSP feature space can be viewed as a filtered and approximated representation of the Riemannian manifold. In a similar way, the tangent space is also an approximated representation of the Riemannian manifold as explained in III-D.

In this example the feature corresponding to the best CSP filter is able to better discriminate the different classes as compared to the best variable in tangent space. However, we need to keep in mind that CSP is a supervised technique and the spatial filters are build using the knowledge of class label in order to maximize the discrimination of the different classes. On the other hand, the tangent space mapping is a fully unsupervised operation. In this sense, the variables obtained with the tangent space mapping are not subject to over-fitting. In this article, we use a supervised variable selection in order to reveal the variables of interest. However, this step is not mandatory and depends on the used classification algorithms.

The number of electrodes is $n = 22$, thus the tangent space is of dimension $m = 253$, which is not much lower than the number of trials (576). Albeit possible it is not efficient to use standard classification algorithms like LDA. This is due to the difficulty to compute an unbiased estimation of the features covariance matrix. For this reason, we perform a selection of discriminant variables according to the procedure described in Section IV-B1.

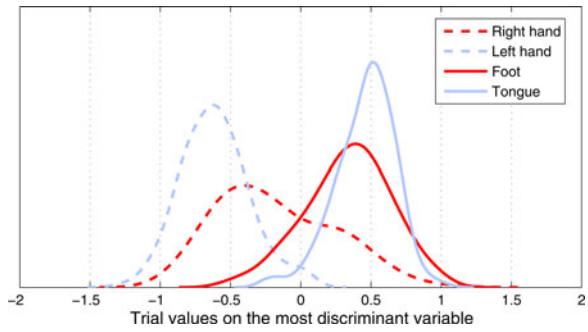

Fig. 4. Class distribution across trials for the most discriminant variable in tangent space.

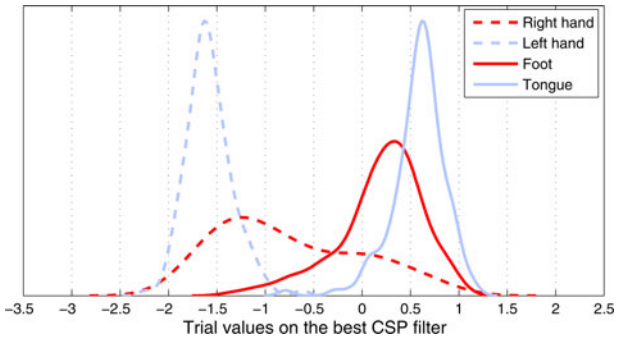

Fig. 5. Class distribution across trials for the best CSP spatial filter.

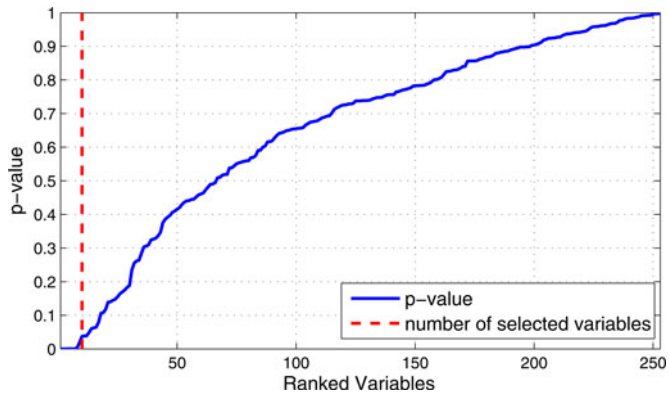

Fig. 6. Evolution of the p-value (straight line) and number of selected variables (dashed line).

After the orthogonalization of the tangent space, the ANOVA sorts the variables in terms of discriminability between the classes. Fig. 6 shows the p-value evolution along the sorted variables. One can observe a quick increase of the p-value, suggesting that only a small subset of variables is useful. The minimal number of variables is set using the weighted FDR algorithm with $q = 0.05$ (in the FDR procedure q is the expected proportion of false rejections with respect to all rejections). For this user, the number of selected variables is 10.

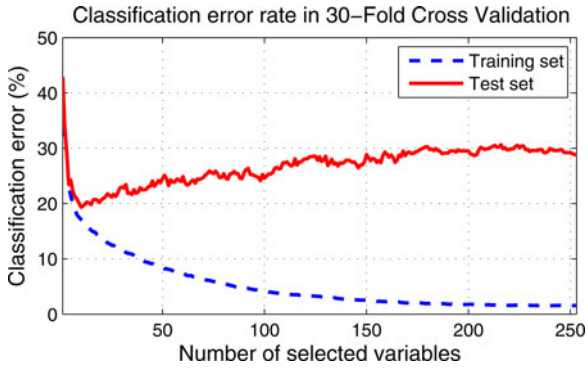

Fig. 7. Classification error versus number of selected variables.

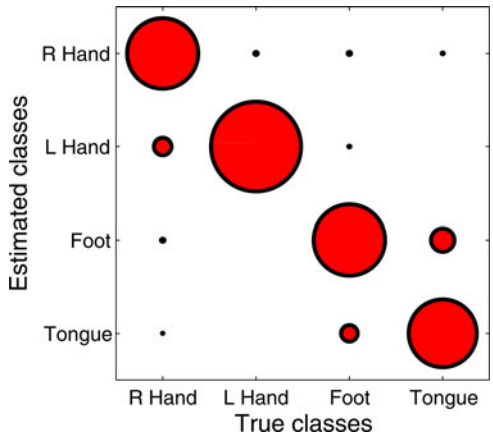

Fig. 8. Confusion matrix in 30-fold cross validation for the subject S1.

3) Classification: Fig. 7 shows the evolution of the classification error rate against the number of selected variables. The classifier is a multiclass implementation of LDA, thus it is very sensitive to over-fitting; the error on the training set decreases to zero while the test-set error rate increases after ten variables.

The minimum classification error rate is given for ten variables. The weighted FDR procedure allows to select the optimal number of variables. Retaining the best ten variables the error rate amounts to 19.5% for this user. A graphical representation of the confusion matrix is given in Fig. 8. As expected, most ambiguities are found between the two pairs of hands and foot versus tongue.

C. Classification Results

We have evaluated the performance of the proposed methods by means of classification accuracy. For all methods, we applied a 30-fold cross-validation procedure [22]. Performance is measured in terms of classification accuracy, thus the higher the score the better the performance. Since there are four classes, the chance level is 25%.

The two proposed Riemannian methods are compared with the classical approach consisting in a spatial filtering by multiclass CSP [1], [20] followed by an LDA classification [2] on

TABLE I
CLASSIFICATION ACCURACY IN 30-FOLD CROSS-VALIDATION

	CSP+LDA	CSP*+LDA (nb. of filters)	MDRM	TSLDA (nb. of var.)
S1	78.3	**81.8** (14.3)	77.8	80.5 (10.6)
S2	44.7	45.1 (6.6)	44.1	**51.3** (13.7)
S3	82.2	83.5 (13.5)	76.8	**87.5** (12.6)
S4	59.1	59 (9.2)	54.9	**59.3** (13.4)
S5	39.7	42.2 (4)	43.8	**45** (15)
S6	50.1	43.3 (3)	47.1	**55.3** (23.3)
S7	81	81.5 (9.1)	72	**82.1** (11.7)
S8	68.5	69.6 (10)	75.2	**84.8** (8.9)
S9	77.4	80 (15.1)	76.6	**86.1** (12.2)
mean	64.6 $\pm$ 16.6	65.1 $\pm$ 17.9 (9.4 $\pm$ 4.3)	63.2 $\pm$ 15.2	**70.2** $\pm$ **17.1** (13.5 $\pm$ 4.1)

TABLE II
MEAN CONFUSION MATRIX FOR CSP+LDA

	R hand	L hand	foot	tongue
R hand	66.8	14.4	11.2	9.2
L hand	17	63.3	9.7	7.1
foot	9.9	12.3	61.5	16.6
tongue	6.2	10	17.7	67.2

TABLE III
MEAN CONFUSION MATRIX FOR CSP*+LDA

	R hand	L hand	foot	tongue
R hand	68.6	14.9	10.5	8.3
L hand	16.5	64.5	11.1	8.5
foot	9.1	11.1	59.8	15.6
tongue	5.9	9.5	18.5	67.7

the feature space composed by the log variance of the spatially filtered signals. The multiclass CSP used in this paper is an implementation of [1] that performed approximate joint diagonalization of all the class-related mean covariance matrices. This algorithm is more efficient than the standard CSP used in a one-versus-rest manner for the multiclass case. The number of CSP spatial filter is set to 8 as proposed in [1], [20]. For a more fair comparison, we also presented the results for the CSP method (denoted CSP*+LDA) whereby the optimal number of filters was selected for each subject according to a weighted FDR criterion that is similar to our method.

No parameters need to be set for both Riemannian methods. For the TSLDA, the procedure is fully automatic since the weighted FDR chooses the optimal number of variables.

The results are shown in Table I. First of all, the subject specific selection of the number of spatial filters did not improve significantly the results of the reference method. However, this selection makes the procedure fully automatic. The MDRM method offers a slightly worse result as compared to the reference methods. This is not surprising since this method does not perform any kind of denoising. More interesting results are given by the TSLDA method. For all subjects except subject S1, this method outperforms the reference methods. It is also interesting to see that the improvement is more important for the difficult cases. The average number of selected variables is small ($\sim$13.5) in comparison to the total number of variables.

A more sophisticated analysis can be done using the confusion matrices given in Tables II–V. It appears that the reference methods have more difficulties to treat the left hand and foot classes. We observe the same behaviour with the MDRM method. On the other hand, the TSLDA approach seems not affected by this problem, displaying around 70% correct classification rate for all classes. In summary, the improvements brought upon by TSLDA are mainly due to a better handling of critical cases, resulting in a 7% improvement classification for both left hand and foot classes.

TABLE IV
MEAN CONFUSION MATRIX FOR MDRM

	R hand	L hand	foot	tongue
R hand	68	21.9	13	10.3
L hand	15.2	56.1	6.1	4.8
foot	9.1	11	61.5	17.1
tongue	7.7	10.9	19.5	67.8

TABLE V
MEAN CONFUSION MATRIX FOR TSLDA

	R hand	L hand	foot	tongue
R hand	**70.7**	15.5	8.5	7.3
L hand	15.5	**68.5**	7	7.1
foot	7.3	8.5	**69.9**	14
tongue	6.4	7.6	14.5	**71.7**

Finally, Table VI gives the results in terms of kappa values, as it was done for the BCI competition IV. The described methods are applied in the conditions of the competition in order to compare performances. The MDRM method achieved a mean performance of 0.52 which ranks this method to the second place of the competition. The TSLDA method achieved a mean performance of 0.567 which is close to the score of the winner. Moreover, the scores are obtained without exploiting the frequential information, contrary to the winner. Higher performances can be reached using an optimisation of frequency filters.

TABLE VI
RESULTS OF THE BCI COMPETITION IN KAPPA VALUES

	mean kappa	S1	S2	S3	S4	S5	S6	S7	S8	S9
1^{st}	**0.57**	0.68	**0.42**	**0.75**	0.48	**0.40**	0.27	**0.77**	**0.75**	0.61
TSLDA	0.567	0.74	0.38	0.72	0.50	0.26	**0.34**	0.69	0.71	**0.76**
MDRM	0.52	**0.75**	0.37	0.66	**0.53**	0.29	0.27	0.56	0.58	0.68
2^{nd}	0.52	0.69	0.34	0.71	0.44	0.16	0.21	0.66	0.73	0.69
3^{rd}	0.31	0.38	0.18	0.48	0.33	0.07	0.14	0.29	0.49	0.44

VI. CONCLUSION

This article presents a *new framework* to classify single trials in BCI applications based on motor imagery. It relies on the covariances matrices as descriptors of EEG signal. Two methods based on Riemannian geometry have been proposed to classify directly the covariance matrices. The first method, named MDRM, is an implementation of the minimum distance to mean (MDM) algorithm using Riemannian distance and Riemannian mean instead of the classical Euclidean equivalent. This algorithm is simple, effective, and offers results close to those obtained with the reference method (consisting in a multiclass CSP followed by an LDA). The second method, named TSLDA, is based on a powerful Riemannian concept: the tangent space. The covariance matrices are mapped onto a higher dimensional space where they can be vectorized and treated as Euclidean objects. In this tangent space, a variable selection procedure is applied in order to decrease dimensionality and a classification by LDA is performed. Significant better results have been achieved with the TSLDA method compared to the reference one. This improvement is mainly due to a better handling of difficult (noisy) cases where there is a strong ambiguity between the class memberships.

Both methods do not need spatial filtering anymore. In the case of the MDRM classification method, the only necessary signal processing steps are a temporal filtering, an estimation of the mean covariance matrices and a comparison of Riemannian distances. No parameters need to be set and the obtained results appear satisfactory. In addition, the tangent space mapping is an unsupervised operation that extracts spatial information with a quality comparable to the state-of-the-art CSP. This result may lead to the development of efficient unsupervised training algorithms in BCI applications.

Yet, there are several issues that need to be investigated in details. On one hand, spatial filtering by CSP can be viewed as a way to reduce the dimensionality of the problem and to give a criterion based on eigenvalues in order to select the best components. On the other hand, the tangent space mapping allows gathering more spatial information by making a higher dimensional space. This high dimensionality of the tangent space could bring difficulties for some algorithms, causing overfitting or bias in statistical estimations. For this reason, the unconditional replacement of spatial filtering by the tangent space mapping is not recommended in all cases. However, several solutions are available to address this problem. In this paper, we have chosen to use a variable selection procedure but a regularization strategy may be efficient as well.

Future works should try to reduce the computational cost of these methods. Since a computation of a Riemannian distance involves a matrix inversion and a matrix diagonalization, the computational time dramatically increases with the number of electrodes. This effect could be reduced using an optimized implementation of eigenvalue decomposition or by using other metrics approximating the Riemannian metric [23].

Finally, it is difficult to interpret the results physiologically, unlike the observation of the spatial patterns in CSP, which gives information about the brain areas involved in the MI. Even if selected variables in the tangent space correspond to geodesics in the Riemann manifold, these geodesics cannot be trivially used to visualize physiological principles, i.e., electrode positions or weights. We will investigate this point in a future work.

Beyond these few limitations the proposed framework appears very promising. This work opens the door to a new family of BCI algorithms wherein investigations have just started.

REFERENCES

[1] C. Gouy-Pailler, M. Congedo, C. Brunner, C. Jutten, and G. Pfurtscheller, "Nonstationary brain source separation for multiclass motor imagery," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 2, pp. 469–478, Feb. 2010.

[2] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. Müller, "Optimizing spatial filters for robust EEG Single-Trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[3] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. Müller, "Single-trial analysis and classification of ERP components—A tutorial," *NeuroImage*, vol. 56. no. 2, pp. 814–825, May 2011.

[4] J. Farquhar, "A linear feature space for simultaneous learning of spatio-spectral filters in BCI," *Neural Netw.*, vol. 22, no. 9, pp. 1278–1285, 2009.

[5] Y. Li, K. Wong, and H. deBruin, "EEG signal classification based on a riemannian distance measure," in *Proc. IEEE Toronto Int. Conf. Sci. Technol. Humanity (TIC-STH)*, 2009, pp. 268–273.

[6] F. Barbaresco, "Innovative tools for radar signal processing based on cartan's geometry of SPD matrices & information geometry," in *Proc. IEEE Radar Conf.*, 2008, pp. 1–6.

[7] P. T. Fletcher and S. Joshi, "Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors," in *Proc. Comput. Vis. Math. Methods Med. Biomed. Image Anal.*, 2004, pp. 87–98.

[8] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[9] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," presented at the Conf. Latent Variable Anal. (LVA/ICA), St Malo, France, 2010.

[10] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[11] S. Smith, "Covariance, subspace, and intrinsic Cramèr-Rao bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, May 2005.

[12] W. Förstner and B. Moonen, "A metric for covariance matrices," Tech. Rep. 1999.6, Dept. Geodesy Geoinformat., Stuttgart Univ., Germany, 1999, pp. 113–128.

[13] M. Moakher, "A differential geometric approach to the geometric mean of symmetric Positive-Definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 3, pp. 735–747, 2005.
[14] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Commun. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–541, 1977.
[15] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 4, pp. R1–R13, 2007.
[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, Feb. 2010.
[17] Y. Benjamini and Y. Hochberg, "Multiple hypotheses testing with weights," *Scand. J. Statist.*, vol. 24, no. 3, pp. 407–418, 1997.
[18] R. Leeb, C. Brunner, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI Competition 2008 - Graz data set B," Graz University of Technology, Austria.
[19] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, May 2006.
[20] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.
[21] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Common spatial pattern revisited by riemannian geometry," in *Proc. 2010 IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, pp. 472–476.
[22] B. Graimann, B. Allison, and G. Pfurtscheller, *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction*. New York: Springer-Verlag, Nov. 2010.
[23] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric Positive-Definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, Jan. 2007.

Authors' photographs and biographies not available at the time of publication.