Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review

Marco Congedo, Alexandre Barachant, Rajendra Bhatia

▶ **To cite this version:**

Marco Congedo, Alexandre Barachant, Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. Brain-Computer Interfaces, 2017, 4 (3), pp.155-174. 10.1080/2326263X.2017.1297192 . hal-01570120

HAL Id: hal-01570120
https://hal.science/hal-01570120

Submitted on 28 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Riemannian Geometry for EEG-based Brain-Computer Interfaces; a Primer and a Review

Marco Congedo[a]***, Alexandre Barachant**[b], **Rajendra Bhatia**[c]

[a] GIPSA-lab, CNRS, Grenoble Alpes University, Grenoble Institute of Technology, Grenoble, FRANCE; [b] Burke Medical Research Institute, White Plains, NY, USA; [c] Indian Statistical Institute, New Delhi, INDIA

*Corresponding Author.

Marco Congedo received the Ph.D. degree in Biological Psychology with a minor in Statistics from the University of Tennessee, Knoxville, in 2003. From 2003 to 2006 he has been a Post-Doc fellow at the French National Institute for Research in Informatics and Control (INRIA) and at France Telecom R&D, in France. Since 2007 Dr. Congedo is a Research Scientist at the "Centre National de la Recherche Scientifique" (CNRS) in the GIPSA Laboratory, University of Grenoble Alpes and Grenoble Institute of Technology. In 2013 he obtained the HDR degree ("Habilitation à Diriger des Recherches") from Grenoble Alpes University. Dr. Congedo has authored and co-authored over 100 scientific publications.

GIPSA-lab,
11 rue des Mathématiques, Domaine Universitaire - BP 46 - 38402, Grenoble, FRANCE.
Tel : +33-(0)4-76 82 62 52
e-mail : marco.congedo@gmail.com

Alexandre Barachant received his Ph.D. degree in Electrical Engineering in 2012 from the Grenoble Alpes University, Grenoble, France. Between 2012 and 2013 he has been a post-doc fellow of the "Centre National de la Recherche Scientifique" (CNRS) in the GIPSA-lab Laboratory, Grenoble, France. Since November 2014 he is with the Burke Medical Research Institute, New York. Using Riemannian classification methods he has recently won five predictive modeling international BCI data competitions (DecMeg 2014, BCI Challenge 2015, Grasp & Lift EEG Challenge 2015, Decoding Brain Signal 2016 and Biomag 2016).

Burke Medical Research Institute,
785 Mamaroneck Ave, White Plains, NY 10605, USA
Tel: +1 (914) 597-2500
e-mail : alexandre.barachant@gmail.com

Rajendra Bhatia received his Ph.D. in Mathematics from the Indian Statistical Institute, Kolkata, India. He is the author of three well-known books: Perturbation Bounds for Matrix Eigenvalues, Matrix Analysis, and Positive Definite Matrices. His work on the Riemannian geometric mean has stimulated several papers in the past 15 years. He is a Fellow of the Indian National Science Academy, the Indian Academy of Sciences, and the Third World Academy of Sciences. He has been awarded the 2016 Hans Schneider Prize in Linear Algebra.

Indian Statistical Institute.
7, S. J. S. Sansanwal Marg, New Delhi, Delhi 110016, INDIA
Tel : +91-11-41 49 39 43
e-mail: rajenbhatia@gmail.com

Riemannian Geometry for Brain-Computer Interface;

a Primer and a Review

Abstract

Despite its short history, the use of Riemannian geometry in brain-computer interface (BCI) decoding is currently attracting increasing attention, due to an accumulating documentation of its simplicity, accuracy, robustness and transfer learning capabilities, including the winning score obtained in five recent international predictive modeling BCI data competitions. The Riemannian framework is sharp from a mathematical perspective, yet in practice it is simple, both algorithmically and computationally. This allows the conception of online decoding machines suiting real-world operation in adverse conditions. We provide here a review on the use of Riemannian geometry for BCI and a primer on the classification frameworks based on it. While the theoretical research on Riemannian geometry is technical, our aim here is to show the appeal of the framework on an intuitive geometrical ground. In particular, we provide rationale for its robustness and transfer learning capabilities and we elucidate the link between a simple Riemannian classifier and a state-of-the-art spatial filtering approach. We conclude by reporting details on the construction of data points to be manipulated in the Riemannian framework in the context of BCI and by providing links to available open-source Matlab and Python code libraries for designing BCI decoders.

Keywords: Brain-Computer Interface, Riemannian Geometry, Electroencephalography, Classification, Signal Processing, Covariance Matrix, Decoding, Geometric Mean.

1. Introduction

Electroencephalography (EEG) is the oldest and most widespread brain imaging modality, flourishing over an 80 year-old tradition in cognitive and clinical neurophysiology [1]. An EEG-based Brain-Computer Interface (BCI) is a system for translating EEG signals directly into commands for a computerized system. Over the past 15 years the BCI field has grown considerably thanks to substantial granting by the EC in Europe and by the NIH and NSF in the USA, among others, becoming the most prominent applied research area for EEG as a whole [2-5]. Functionally, a BCI is constituted by two agents, the *user* and a *computer* (machine), entering in *reciprocal interaction* by means of two distinct elements: a *decoder*, which is the core of a BCI that translates brain signals into commands and an *interface*, the computerized application, whose ultimate goal is performing actions while giving continuous feedback to the user about its operation. The latest research on BCIs has taken several disjointed but intimately related paths. In order to build a frame for the arguments we advance in this article we provide here a quick overview of these lines of research:

I) The study of the *applicability of BCI technology to clinical population*s (e.g., [6-14]), for instance, in comparison (or combination) to other Assistive Technology (AT) such as eye-tracking, electrooculography and contactors [15, 16].

II) The study of new *electrophysiological markers* that may be used to improve a BCI decoder in addition to the traditional (steady-state) event-related potentials and event-related (de)synchronization, such as *global synchronization* [17] and *phase-lag index variance* [18].

III) The improvement of the *BCI interface* by studying its properties [19-22]. For P300-based BCI these lines of research have led to, inter varia, the introduction of *language models* for letter and word prediction [23-25], *automatic pause detection* [26], the use of faces for flashing symbols [24], the use of random groups or pseudo-random groups flashing instead of row-column flashing [27-30], the use of inter-stimulus intervals randomly drawn from an exponential distribution instead of constant [27], the *dynamic stopping* of flashing sequences [25, 31], etc. For SSVEP-based BCI improvements of the interface include the use of precise tagging of the flickering so as to use phase information (e.g., [32]) and the use of smart flickering sequences such as code modulation [33], multi-phase cycle coding [34], etc.

IV) The study of *physiological*, *anatomical*, *cognitive* and *emotional factors* that may lead to improvements of the performance and usability, such as the motivation and empathy of the user in P300 spellers [35], gray matter volume in non-primary somatosensory and motor areas in a BCI based on motor-imagery [36], etc.

V) The *development of EEG sensing technology*, including portability and autonomy of the amplification unit (e.g., http://openbci.com), the performance, comfort and usability of the electrode-cap (e.g., [37-39]), new type of sensors (e.g., [40]), on-line EEG signal quality monitoring [41, 42], and all ergonomic matters such as *accessibility*, *usability* and *acceptance* of the technology [43].

VI) The study of BCI-related academic subjects such as the *transfer rate metrics*, that is, how the actual online accuracy of a BCI should be evaluated [44].

VII) *Multi-subject BCIs*, that is, BCI systems controlled by several users, in proximity one to the other or remotely connected [45-48]. Besides allowing remote social operation, this functioning has potential to achieve 100% accuracy on single trials by combining the data of several users.

VIII) BCI Hybridization, that is, the combination of several BCI modalities on the same interface to increase the accuracy, the ergonomy, usability and the bit rate (e.g., [49-51]).

IX) The increase of *robustness* for a BCI. This line of research is central to the present article, so we describe it in some details in the next section. *Poor usability* and *lack of robustness* are major limits of state-of-the-art BCIs. This can been found stated in all review and position papers (e.g., [52, 53]), as well as in roadmaps drawn by integrative projects such as *Future BNCI*[1] and *BNCI Horizon 2020*[2]. In summary, it has been stated that: "efforts to commercialize research findings have been tepid, hampered by a general lack of robustness when translating technologies to uncontrolled environments beyond the research laboratory [54]. The usability and robustness of a BCI heavily depend both on the *application* and on the

[1] http://bmiconference.org/wp-content/uploads/files/Future_BNCI_Roadmap.pdf
[2] http://bnci-horizon-2020.eu/roadmap

decoder. The former is a traditional subject of study in the field of human-machine interaction (HMI). In this article, we address the latter.

1.1 BCI Decoders of Second Generation

The decoder is traditionally conceived as formed by three independent modules: *pre-processing*, *feature extraction* and *classification* [55]. One source of dispersion in the field is that each one of the three main BCI modalities, namely, based on motor/mental-imagery (MI), event-related potentials (ERP) and steady-state evoked potentials (SSEP) is currently treated with dedicated pre-processing, signal processing and classification modules. The decoding strategy itself is also fragmented; we can divide existing paradigms in two categories: those that follow a *hard machine learning* approach, and those that use *signal processing* to increase the signal-to-noise ratio followed by a simple classification algorithm. Some algorithms of the hard machine-learning kind generalize fairly well across sessions and across subjects, but require a substantial amount of training data. Furthermore, they are often computationally intensive. The opposite happens for the spatial filtering kind, where bad generalization capabilities are compensated by a fast training and lower computational cost. The major factors jeopardizing the operation of a BCI decoder in real-world operation are the *inter-subject physiological variability* and the great variability of real-world *environmental conditions* (e.g., [56]). Three inter-connected lines of research are trying to overcome these limits, addressing jointly the improvement of BCI usability and robustness:

1. The ability to use a BCI without calibration or after a short calibration has been recognized in the community as a priority for improving BCI usability. Toward this objective effort is put in the conception, analysis and testing of *generic model classifiers* and/or *domain adaptation methods*, allowing the so-called *transfer learning*, whereas data from other sessions and/or other subjects are used to initialize a BCI so as to start using it without calibration and also to increase the performance of low-performance users [25, 30, 57-63]. While generic initializations by transfer learning are easier to implement, we name a *smart initialization* an initialization seeking optimality for a given user, topic on which little has been done so far.

2. The *continuous (on-line) adaptation of the classifier*, which complements the (smart) initialization of point 1., in that the adaptation ensures that optimal performance is achieved

regardless of how good the initialization is. It also allows keeping optimal performance by adapting to mental and environmental changes during the session, enabling continuous adjustments ("pursuing"), thus ensuring reliability and robustness throughout the session and across-sessions [30, 62, 64-66].

3. The conception, development and maintenance of world-wide massive *databases* (e.g., [54]). Such a resource is necessary to boost research by allowing massive testing of decoding strategies. It also enables the systematic study of the source of variation in individual EEG patterns and their relation to BCI capabilities and individual attainable performances. Finally, it yields the (smart) initialization of a BCI (by subject-specific transfer learning), which is necessary to use effectively a BCI without calibration. It is very recent the inception of a large public database of BCI data thanks to the BNCI Horizon 2020 European Coordination project[3].

In consideration of current limitations of BCI decoders it has been stated that "the field would benefit from a new paradigm in research development that focuses on robust algorithm development" [54]. It has soon been realized that the path for reaching the sought usability and robustness does not pass by the enhancement of the system complexity, instead "we need to make a balance between technological advancement and practical use in a real-world situation" [67]. Along these lines, it has been recommended to start regarding the pre-processing, feature extraction and classification not as isolated processes, but as a whole [67, 68]. Taking into accounts all the aforementioned current lines of research *I-IX*, we suggest here a list of requirements a modern BCI decoder should possess:

a) It should be *accurate* in general, as compared to established state-of-the-art approaches.
b) It should be *reliable*, that is, it should maintain as much constant as possible its functions and accuracy in routine circumstances, as well as in hostile or unexpected circumstances.
c) It should perform generally well as *initialized with generic parameter*, even for a naive user, that is, it should possess good generalization abilities both cross-subject and cross-session.

[3] http://bnci-horizon-2020.eu/database/data-sets

d) It should *learn fast* the individual characteristics and then maintains optimality, *adapting* fast to the mental state of the user and to environmental changes.
e) It should be *universal*, that is, applicable to all BCI paradigms (hence to hybrid systems).
f) It should be *algorithmically simple*, so as to be robust and usable in unsupervised on-line operation.
g) It should be *computationally efficient*, so as to work on small electronic devices in line with the current trend in portability of micro-electronic devices.
h) It should be able to readily generalize to the *multi-user* setting, in line with the current trend on the social nature of current electronic technology.

The aim of this article is to describe a simple BCI decoder paradigm that possesses all these requirements. In our conception the key for the sought usability robustness is obtained in "real-life" situations by both good generalization and fast learning capabilities. The approach we delineate represents a true paradigmatic shift and has been possible thanks to recent advances in *Riemannian geometry*, resulting from converging advances in differential geometry, operator theory, matrix analysis, probability, quantum physics and numerical analysis [69]. The use of Riemannian geometry for classification and detection of empirical data is relatively new, but has spread rapidly driven by practical problems in a number of diverse application fields including radar data processing, image processing, computer vision, shape analysis, medical imaging (especially diffusion magnetic resonance imaging and, indeed, BCI), sensor networks, elasticity, mechanics, optimization and machine learning. Thanks to its properties it behaves well in a fully adaptive mode of operation and generalizes straightforwardly to the multi-user setting. A schematic representation of a BCI mode of operation incorporating such a decoding technology is presented in fig. 1. In this article we explain the key features of a Riemannian decoder, without assuming any knowledge on differential geometry. We then review existing literature on Riemannian BCI decoders. BCI-specific EEG data pre-processing is detailed in Appendix I. In Appendix II we point to available open-source code libraries, including Matlab and Python libraries.

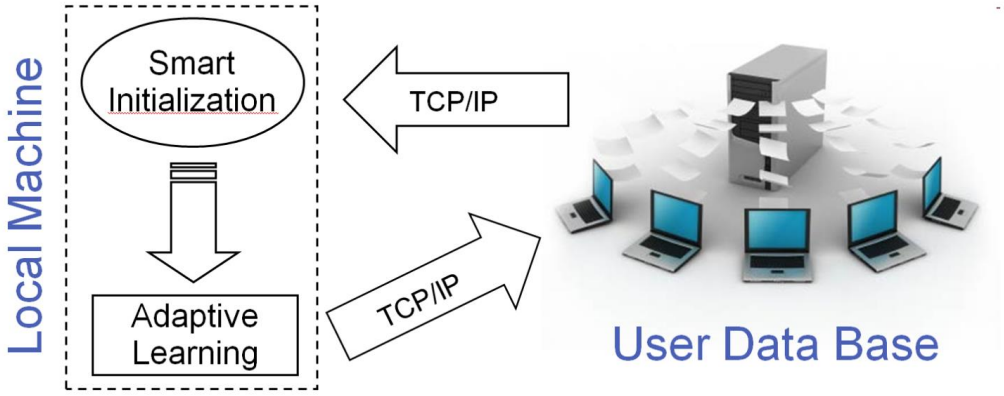

Figure 1: Concept for a possible new generation of BCIs. *At start-up a BCI queries a database to obtain an initialization, possibly sending minimal EEG data of the user so as to receive back a smart initialization that fits appropriately even a naive user. The BCI is operational straightaway, albeit possibly suboptimal at the very beginning. While being used the BCI adapts to the user and sends back the data to the database, along user information, so as to enrich the database and to allow smarter initializations in future sessions of the same user as well as other users. Multiple subjects may use at the same time the same BCI, in which case the decoding machine of the BCI may be located on the server, so to exploit the multitude of data to increase performance.*

2. A Primer on the Riemannian Classification Framework

> « It often happens in analytic geometry that complex algebraic relations translate simple geometric properties, thus … using the geometric language we can express algebraic relations clearly, concisely and in a way that is accessible to intuition… often then a geometric relation can be discovered more easily as compared to its algebraic equivalent, thus the language of geometry provides both an expressive exposition and an effective research tool. » (Levi-Civita, 1925 [70], p.9, translated from Italian by author M.C.)

A symmetric positive definite (SPD) matrix, hereafter referred to as a *positive matrix*, can be thought of as a generalization to multiple dimensions of squared real numbers. Particularly, whereas the variance of a random variable in one dimension is a positive number (a sum of squares), the variance of a *N*-dimensional random variable is a positive matrix, usually named the *covariance matrix*. A positive matrix has a long list of important algebraic and analytic properties, which can be found in the monograph of Bhatia [69]. Our goal in this section is to introduce the basic Riemannian approach for classifying EEG data and to explain its rationale. We will do it by showing that it is a generalization to the multi-dimensional case of the well-

known straightforward decoding approach based on a single-channel signal variance thresholding. Throughout this document *N* indicates the number of available EEG sensors.

Consider first the case of EEG signal recorded at one electrode (*N*=1), say, at the vertex (Cz). Such a signal forms a single time series, which we will indicate as $x(t)$. This suffices, for instance, to detect the beta-rebound phenomenon occurring after the feet have been moved for a few seconds and as such it has been used, for example, in the "tie-fighter" OpenViBE demonstrator [71] and in [72]. The beta-rebound is a temporary increase in signal energy, often named "power", in the beta frequency range (e.g., 16-24 Hz). Let us indicate by $\boldsymbol{x}_k$ a time-window composed of *T* samples in which the beta-rebound is to be detected, where *k* is the index of the time-window under consideration (e.g., a trial) and where the bold character $\boldsymbol{x}_k$ denotes a vector (hereafter matrices as well will be indicated by bold characters). For data filtered in an appropriate band-pass region, which has the effect to retain the energy only in that region and to nullify the mean of the signal, the energy of the signal in time-window $\boldsymbol{x}_k$ can be estimated by the signal *variance* $\mathrm{Var}(\boldsymbol{x}_k)$, often indicated by greek letter sigma σ_k^2. Since the signal is centered, the variance is estimated simply averaging the squares of the elements of $\boldsymbol{x}_k$. The variance can be monitored on-line by means of sliding windows, yielding continuous estimations σ_k^2, for k=1,2,… A beta-rebound is detected whenever the signal variance for the current time-window exceeds a *threshold* (Fig. 2). Equivalently, one can estimate a representative value of the variance at rest, that is, a "mean" of the variance at rest, name it $\bar{\sigma}_0^2$ and a mean of the variance during the beta-rebound, name it $\bar{\sigma}_+^2$; the beta-rebound is then detected whenever the current observed value σ_k^2 is closer to the beta-rebound mean as compared to the rest mean, i.e., whenever $\delta(\sigma_+^2, \sigma_k^2) < \delta(\sigma_0^2, \sigma_k^2)$. Here $\delta(\cdot,\cdot)$ denotes an appropriate *distance function* between the two scalar arguments (Fig. 2). Such a classifier can be found in the literature under the name *Minimum Distance to Mean* (MDM), *Nearest Centroid* or *Mean-of-Class Prototype* [73]. It is a particular instance of the weighted nearest neighbor classifier, which is the *simplest* classification approach one can think of[4]. Its interest is that it works in the same way for any number of classes and that it can be extended appropriately to any dimension, as we will see.

[4] In fact, on can think of the mean as a weighted function of the training examples.

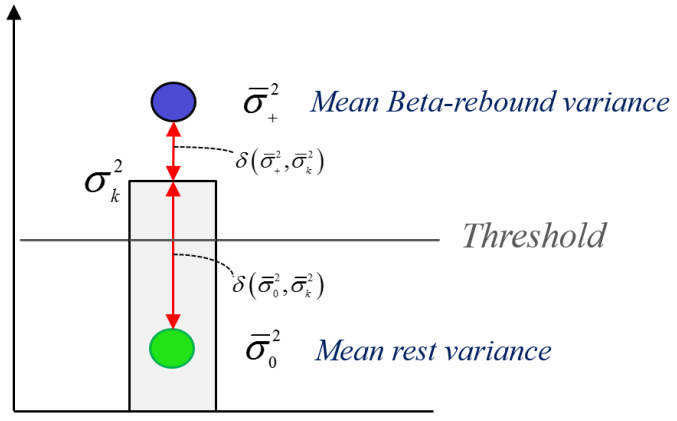

Figure 2: The Minimum Distance to Mean (MDM) classifier in the two-class mono-dimensional case. *Given the observed value of the variance σ_k^2 the current trial is assigned to class "Beta-rebound" if its value exceed the threshold, to the "rest" class otherwise. Equivalently, one may estimate a mean for each class and assign the current trial to the closest mean, according to a distance function here indicated by the arrows. This latter procedure generalizes straightforwardly to the multi-dimensional case and to any number of classes.*

In order to implement the MDM we need an appropriate *distance* function (metric) and the corresponding *mean* function. In mathematics, a *metric* (or distance) is a function that defines a distance between each pair of elements of a set, with the following properties:

- it is non-negative
- it is equal to zero only if the two elements are equal
- it is symmetric
- it obeys the triangle inequality.

A set endowed with a metric is called a *metric space*. For example, we define here the metric space ($\mathcal{S}$, d) as the set of positive real numbers $\mathcal{S} = (0, \infty)$ endowed with metric d. According to a principle of Maurice Fréchet, each metric in such metric space leads to the concept of *mean* as it follows:

Fréchet's variational approach **(mono-dimensional case)**

Let ($\mathcal{S}$, d) be the metric space of positive real numbers endowed with metric d and $\{c_1,\ldots, c_K\}$ be a set of K points in it. The *mean* of the set $\{c_1,\ldots, c_K\}$ is a point x minimizing the *dispersion* $\frac{1}{K}\sum_k d^2(x, c_k)$.

The determination of the mean in this way is a *least-squares* problem, actually the reason why Adrien-Marie Legendre developed it for his astronomical studies:

> « Therefore we see that, in a way, the least-squares method let us know the center around which the experimental results take position, so as to deviate from it as little as possible » (Legendre, 1806 [74], p.75 (appendix), translated from French by author M.C.)

The most familiar metric on $\mathcal{S}$ is the "usual" Euclidean distance. If we take the Euclidean distance

$$d_E(a,b) = |a-b|, \tag{1}$$

the corresponding Euclidean mean of a set of points $\{c_1,\ldots,c_K\}$ is the point m that solves the minimization $\arg\min_m \frac{1}{K}\sum_k d_E^2(m,c_k) = \arg\min_m \frac{1}{K}\sum_k |m-c_k|^2$. This turns out to be the usual *arithmetic mean* $m = \frac{1}{K}\sum_k c_k$. Thus, the Euclidean mean is the point minimizing the *sample variance*, the dispersion of the set around the mean according to the Euclidean distance. For BCI applications, an MDM classifier based on the Euclidean distance and associated mean gives poor results already in the mono-dimensional case. The practice in the BCI field is to consider the log-transformed variance instead (e.g., see [75]). This results in the *log-Euclidean distance*, also called *hyperbolic* or *geometric* distance on $\mathcal{S}$, defined as

$$d_G(a,b) = |\log a - \log b| = \left|\log \tfrac{a}{b}\right|. \tag{2}$$

This switch to the logarithmic scale is common to many areas. Some examples are the Richter scale for earthquake intensity, the decibel scale for sound, the octave scale in music, the pH value in chemistry and f-stops in camera light exposure. In statistics, if a and b are two sample variances, their ratio is the usual F Snedecor's statistic for testing equality of two variances and the log of their ratio has the general Fisher's z distribution [76][5]. Contrarily to the Euclidean distance (1), the geometric distance (2) enjoys the *scale invariance*

[5] It is worthwhile to note here that the logarithmic function is a one-to-one function from the half-line $\mathcal{S}=(0,\infty)$ to the whole real line $\mathbb{R}=(-\infty,\infty)$. Its inverse map is the exponential map. Whereas the space $\mathcal{S}$ has a boundary point 0, which does not belong to $\mathcal{S}$, the real line does not have such a point. In the language of metric spaces, $\mathcal{S}$ with the metric (2) is a *complete metric space*. This guarantees the existence of certain limits and minima/maxima. Loosely speaking, with respect to the Euclidean distance (1) the point *0* is at a finite distance *a* from any point *a* of $\mathcal{S}$, whereas with respect to the geometric distance (2) *0* is infinitely distant from *a*. Also,

$$d_G(xa, xb) = d_G(a, b), \text{ for all } a, b, x > 0 \tag{3}$$

and the *invariance under inversion*

$$d_G\left(a^{-1}, b^{-1}\right) = d_G(a, b) \text{ , for all } a, b > 0. \tag{4}$$

The Fréchet mean of the K points $\{c_1,\ldots,c_K\}$ corresponding to the geometric distance is the point g that solves the minimization problem $\arg\min_g \frac{1}{K}\sum_k d_G^2(g, c_k) = \frac{1}{K}\sum_k \left|\log g - \log c_k\right|^2$.

This turns out to be another famous Pythagorean mean, the *geometric mean*:

$$g = \sqrt[K]{c_1 \cdot c_2 \cdot \ldots \cdot c_K} = \exp\left(\frac{1}{K}\sum_k \log c_k\right). \tag{5}$$

Inherited from the geometric distance, the geometric mean is an appropriate descriptor of the central tendency (expected value) for the variance, while, inherited from the Euclidean distance, the arithmetic mean is not. In fact, the distribution of the variance is a chi-squared. As such, it is symmetric only asymptotically with sampling size and it is well-known that the arithmetic mean is a good central tendency descriptor only for symmetric distributions. This is illustrated in the left column of Fig 3; while for Gaussian distributions both the arithmetic and the geometric mean are good descriptor of the central tendency, for chi-squared distributions the geometric mean is a better descriptor.

---

when b is a reference, say b=1, the square of the geometric distance goes to infinity as a goes toward either ∞ or 0. This is useful in modeling phenomena where a very small signal is as unlikely as a very large one and provides us with an appropriate measure to compare two variances.

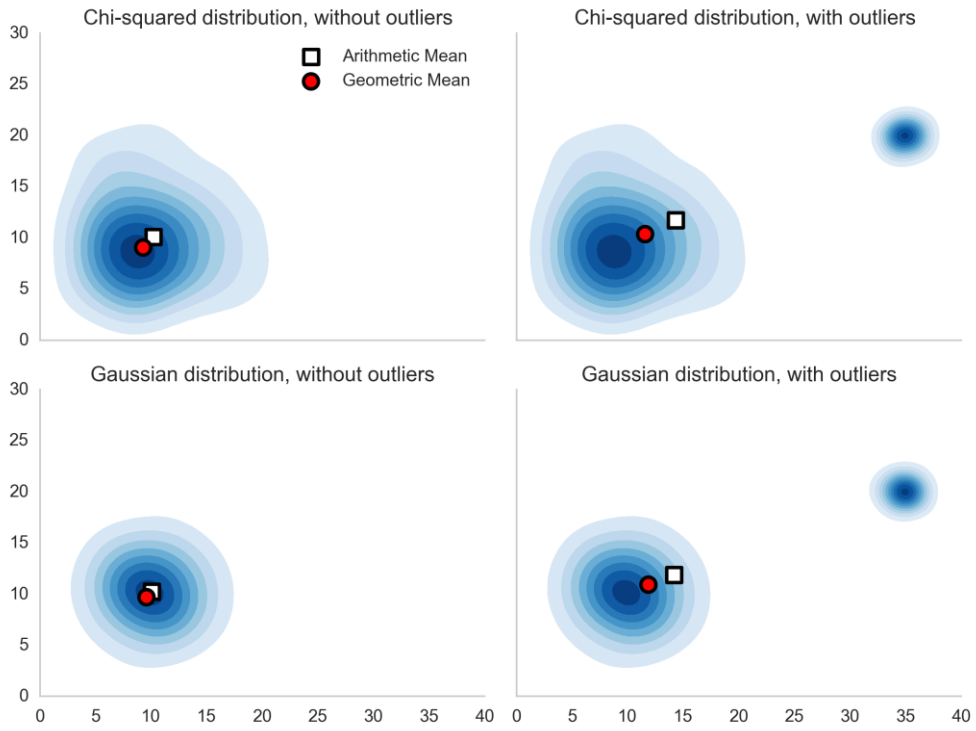

Figure 3: Arithmetic and geometric means. *Empirical isodensities of Chi-Square (10 degrees of freedom, top row) and Gaussian (bottom row) distributions, without (left column) and with outliers (right column). See text for explanations.*

Let us now see how *the Riemannian distance on the manifold of positive matrices can be understood as a straightforward generalization to any dimension of the mono-dimensional geometric distance (2) and how we can derive the associated mean*. Again, we will consider a practical example. Consider the case N=2. Let $x_1(t)$ and $x_2(t)$ be two EEG time series (random variables) recorded at electrodes C3 and C4 as a function of time. Such a setting has been used, for instance, in the pioneering work on BCI [77] for classifying left hand and right hand motor imagery trials. As before, let $\boldsymbol{x}_{1k}$ and $\boldsymbol{x}_{2k}$ be the k^{th} time-window under analysis. Their covariance matrix $\boldsymbol{C}_k$ is

$$\boldsymbol{C}_k = \begin{pmatrix} \mathrm{Var}(\boldsymbol{x}_{1k}) & \mathrm{Cov}(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k}) \\ \mathrm{Cov}(\boldsymbol{x}_{2k}, \boldsymbol{x}_{1k}) & \mathrm{Var}(\boldsymbol{x}_{2k}) \end{pmatrix}. \tag{6}$$

As compared to the mono-dimensional setting, we are considering here not only the variance of the time-series, which are the diagonal elements of $\boldsymbol{C}_k$, but also their covariance, which are the off-diagonal elements. Since $\mathrm{Cov}(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k}) = \mathrm{Cov}(\boldsymbol{x}_{2k}, \boldsymbol{x}_{1k})$, the matrix $\boldsymbol{C}_k$ is symmetric, thus

it is determined by *(N(N+1))/2* of its elements, three in this case. Therefore, we can represent covariance matrices $\boldsymbol{C}_k$ as *data points* in a 3D space with coordinates along the axes Var($\boldsymbol{x}_1$), Var($\boldsymbol{x}_2$) and Cov($\boldsymbol{x}_1$, $\boldsymbol{x}_2$). Because of the Cauchy-Schwarz inequality (or, equivalently, because of positivity of $\boldsymbol{C}_k$), $|\mathrm{Cov}(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k})|^2 \leq \mathrm{Var}(\boldsymbol{x}_{1k})\ \mathrm{Var}(\boldsymbol{x}_{2k})$, any data point is confined in the interior of a symmetric cone (Fig 4). In electrophysiological terms, realization $\boldsymbol{C}_k$ moves along the three coordinates when either the energy (variance) at any of the two electrodes changes or when the phase synchronization and/or amplitude co-modulation between the signal captured at the two electrode changes. The more two points move away from each other along these coordinates, the more they will occupy separated regions in the cone (Fig 4). For any higher dimension *N*, the cone will become an hyper-cone, but everything we say applies in the same way.

Now we turn to the question of *equipping the cone of positive matrices with a suitable distance akin to the geometric distance on* $\mathcal{S}$. The space of *N*x*N* symmetric matrices is a linear space of dimension *N*(*N*+1)/2. It has a natural *inner product* on it given by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{AB})$ and associated Euclidean norm $\|\boldsymbol{A}\|_2$, defined by $\|\boldsymbol{A}\|_2^2 = \mathrm{tr}\,\boldsymbol{A}^2 = \sum_n \lambda_n(\boldsymbol{A})^2$, where $\lambda_n(\boldsymbol{A})$ are the *N* eigenvalues of $\boldsymbol{A}$ and tr is the trace operator. If the matrix $\boldsymbol{A}$ has elements a_{ij}, then $\|\boldsymbol{A}\|_2^2 = \sum_{i,j=1}^{N} |a_{ij}|^2$. So, $\|\boldsymbol{A}\|_2$ is a natural extension to matrices of the Euclidean norm on vectors. The cone of positive matrices is a subset of symmetric matrices and naturally inherits this Euclidean norm. This might be satisfactory for some problems, but has several shortcoming in the BCI context, even in the simplest mono-dimensional case. Fortunately, another norm and associated distance coming from Riemannian geometry turns out to be just the right one. The set $\mathcal{S}_{++}(N)$ of *N*x*N* positive matrices is a differentiable manifold. By that we mean that every small neighborhood around a point $\boldsymbol{P}$ in it "looks like" the Euclidean space of symmetric matrices, of which it is an open set. Think of "flattening" a small region of a surface (Fig. 5). The space of all symmetric matrices at any base point on the manifold is called the *tangent space*. Riemannian geometry begins by equipping with an inner product each tangent space, in a way that the resulting metric varies smoothly from point to point[6]. In this case, the inner product at a given point $\boldsymbol{P}$ is

[6] Sometimes the inner product on the tangent space is referred to as the "metric". In this work we name metric the distance function on the manifold deriving from the definition of inner product on the tangent space.

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\boldsymbol{P}} = \mathrm{tr}(\boldsymbol{P}^{-1}\boldsymbol{A}\,\boldsymbol{P}^{-1}\boldsymbol{B}). \tag{7}$$

The associated norm is then $\|\boldsymbol{A}\|_{2,\boldsymbol{P}}$, given by $\|\boldsymbol{A}\|_{2,\boldsymbol{P}}^2 = \|\boldsymbol{P}^{-1}\boldsymbol{A}\|_2^2 = \|\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{P}^{-\frac{1}{2}}\|_2^2$. When $\boldsymbol{P}=\boldsymbol{I}$, the identity matrix, this reduces to the norm $\|\boldsymbol{A}\|_2$ introduced earlier[7].

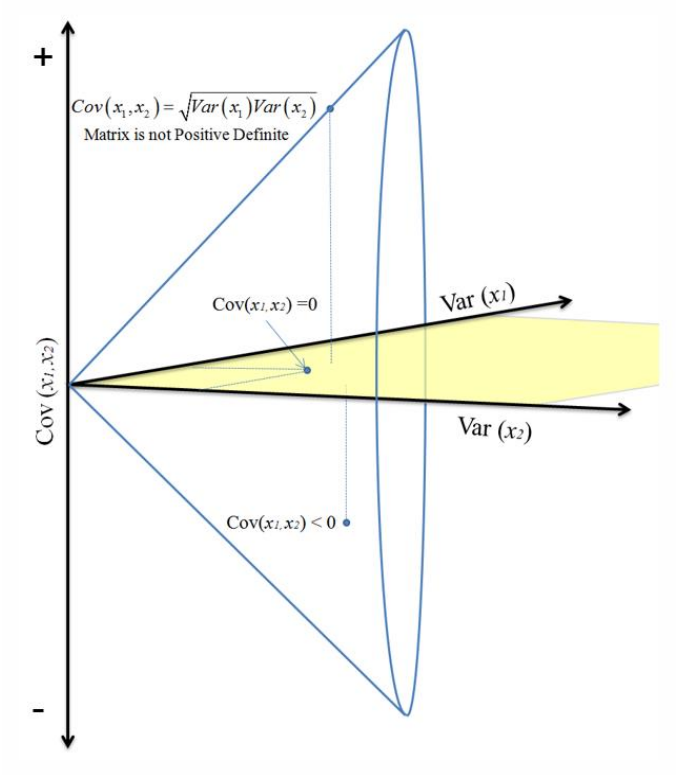

Figure 4: Symmetric Convex Cone of SPD matrices. *Any symmetric positive definite matrix lie in the interior of an open cone because of the Cauchy-Schwarz inequality. When the point touches the border of the cone the inequality becomes an equality and the matrix is no more positive definite.*

[7] Note that this Riemannian inner product is a matrix version of the Fisher information metric in probability; for two N-dimensional vectors $\boldsymbol{a}=(a_1,\ldots, a_N)$ and $\boldsymbol{b}=(b_1,\ldots, b_N)$ and a positive vector $\boldsymbol{p}=(p_1,\ldots, p_N)$, the Fisher metric is defined as $\sum_n \left(a_n b_n / p_n^2\right)$. For this reason metric (7) is sometimes called the *Fisher metric* (or Fisher-Rao metric).

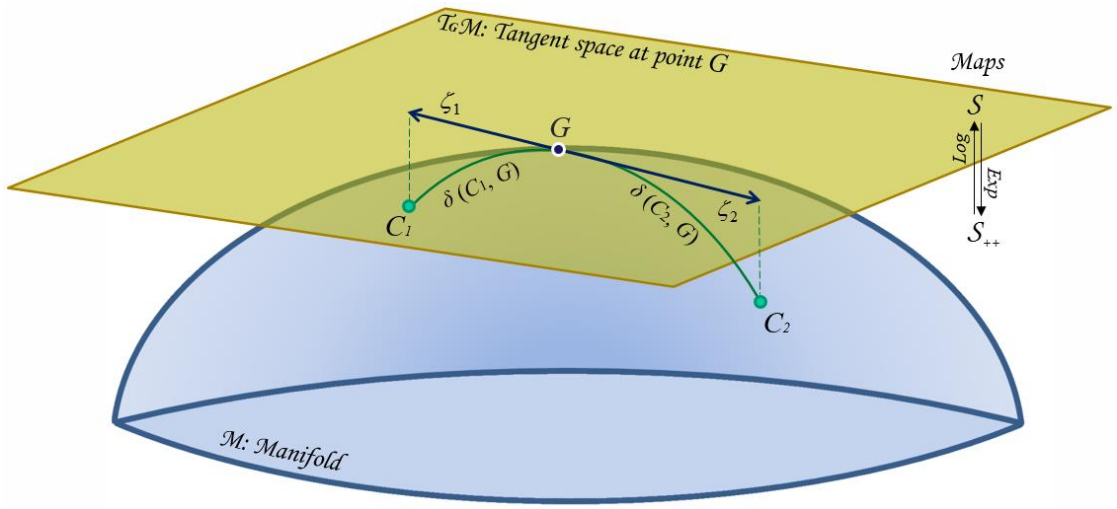

Figure 5. *Schematic representation of the symmetric positive definite matrix manifold, the geometric mean* $\boldsymbol{G}$ *of two points and the tangent space at* $\boldsymbol{G}$*. Consider two points (e.g., two covariance matrices)* $\boldsymbol{C}_1$ *and* $\boldsymbol{C}_2$ *on* $\mathcal{M}$*. The geometric mean of these points is the midpoint on the geodesic connecting* $\boldsymbol{C}_1$ *and* $\boldsymbol{C}_2$*, i.e., it minimizes the sum of the two squared distances* $\delta^2(\boldsymbol{C}_1, \boldsymbol{G})+\delta^2(\boldsymbol{C}_2, \boldsymbol{G})$*. Now construct the tangent space* $\mathcal{T}_G\mathcal{M}$ *at* $\boldsymbol{G}$*. There exists one and only one tangent vector* $\boldsymbol{\zeta}_1$ *(respectively* $\boldsymbol{\zeta}_2$*) corresponding to the geodesic departing from* $\boldsymbol{G}$ *and arriving at* $\boldsymbol{C}_1$ *(respectively* $\boldsymbol{C}_2$*) on the manifold; The map from the tangent space (symmetric matrices* $\mathcal{S}$*) to the manifold (symmetric positive definite matrices* $\mathcal{S}_{++}$*) is an exponential map. The inverse map from the manifold to the tangent space is a logarithmic map (see [69] for details).*

The next step is to compute the length of any curve in the space $\mathcal{S}_{++}(N)$ using inner product (7) on the tangent space. This can be done using calculus. Now, given any two points $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ in $\mathcal{S}_{++}(N)$ there could be several curves passing through them. If there is a *unique one of minimal length*, it is called a *geodesic*[8]. Fortunately, in the manifold of positive matrices equipped with inner product (7) on the tangent space *a geodesic exists for any two points* $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. The *length of the geodesic* from $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ (or vice versa) gives the *Riemannian distance* (Fig. 5). Simple calculations show that this distance has closed form solution given by [69, 78-83]

$$\delta_G(\boldsymbol{C}_1, \boldsymbol{C}_2) = \left\| \mathrm{Log}\left(\boldsymbol{C}_1^{-\frac{1}{2}} \boldsymbol{C}_2 \boldsymbol{C}_1^{-\frac{1}{2}}\right) \right\|_F = \sqrt{\sum_{n=1}^{N} \log^2 \lambda_n}\,, \tag{8}$$

[8] It is useful here to look what happens on a sphere. Great circles are geodesic curves. If two points on the sphere are not antipodal, then there is a geodesic passing through them. If the points are antipodal, then there are infinitely many great circles passing through them.

where λ_n are the N eigenvalues of matrix $\boldsymbol{C}_1^{-½}\boldsymbol{C}_2\boldsymbol{C}_1^{-½}$ or, equivalently, matrix $\boldsymbol{C}_1^{-1}\boldsymbol{C}_2$ and where in both expressions the indices 1 and 2 can be permuted, showing that this distance is symmetric[9]. It is clear that in the very special case N=1, this reduces to the geometric distance (2) between positive real numbers[10].

Definition of distance (8) as the length of a geodesic ensures that it has all the properties of a distance function. It has several nice additional properties, some of which are important for its use in BCI. We mention just two of them: given an invertible matrix $\boldsymbol{X}$, the matrix $\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T$ is called a *congruence* or a *conjugation* (here superscript T denotes the transpose). A matrix $\boldsymbol{A}$ is positive if and only if $\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T$ is positive. The distance (8) is invariant under any congruence, i.e.,

$$\delta_G\left(\boldsymbol{X}\boldsymbol{C}_1\boldsymbol{X}^T, \boldsymbol{X}\boldsymbol{C}_2\boldsymbol{X}^T\right) = \delta_G\left(\boldsymbol{C}_1, \boldsymbol{C}_2\right) . \tag{9}$$

It is also invariant under inversion, i.e.,

$$\delta_G\left(\boldsymbol{C}_1^{-1}, \boldsymbol{C}_2^{-1}\right) = \delta_G\left(\boldsymbol{C}_1, \boldsymbol{C}_2\right). \tag{10}$$

These two properties are straightforward extensions to matrices of (3) and (4), respectively, and are instrumental for the robustness of Riemannian BCI decoders, as we will see.

Having defined the Riemannian distance for the general (matrices) case, we can obtain the corresponding *Riemannian mean*:

[9] In order to avoid confusion in (8) and hereafter we denote Log(·) and log(·) the matrix and scalar logarithm, respectively. One should be careful with quantities involved in equation (8): while $\boldsymbol{C}_1^{-1}\boldsymbol{C}_2$ and $\boldsymbol{C}_1^{-½}\boldsymbol{C}_2\boldsymbol{C}_1^{-½}$ are similar, hence have the same eigenvalues, they do not have the same eigenvectors. The fact that indices 1 and 2 can be permuted in equation (8) is due to the fact that the eigenvalues of $\boldsymbol{C}_1^{-1}\boldsymbol{C}_2$ are the inverse of those of $\boldsymbol{C}_2^{-1}\boldsymbol{C}_1$ (since these two matrices are the inverse of each other) and for any positive number λ it holds $\log^2 \lambda = \log^2 \lambda^{-1}$.

[10] This is so if we take d_G to mean the extreme right hand side expression in (2). The middle expression could suggest a simpler generalization of (2) allowing matrix distance function $\left\|\operatorname{Log}\boldsymbol{C}_1 - \operatorname{Log}\boldsymbol{C}_2\right\|_2$. This is called the Log-Euclidean distance and is different from Riemannian distance δ_G unless $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ commute in multiplication. The Log-Euclidean distance and associated mean $\operatorname{Exp}\left(\frac{1}{K}\sum_k \operatorname{Log}(\boldsymbol{C}_k)\right)$ are useful in several problems (e.g., [84]), however they fail to have the crucial congruence-invariance property. This will be encountered next.

Fréchet's variational approach (general case): Let $(\mathcal{S}_{++}(N), \delta)$ be the metric space of positive matrices endowed with metric δ and $\{\boldsymbol{C}_1,\ldots, \boldsymbol{C}_K\}$ be a set of K points in it. If there exists a unique point $\boldsymbol{X}$ for which the dispersion $\frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\delta^2\left(\boldsymbol{C}_k, \boldsymbol{X}\right)$ is minimal, then $\boldsymbol{X}$ is called the mean of the points $\{\boldsymbol{C}_1,\ldots, \boldsymbol{C}_K\}$.

Note the condition that the minimizer of the dispersion exists and that it is unique. In fact, it may not be the case. For example, on the surface of a sphere with its usual distance, any two antipodal points have infinitely many "means", due to the fact that a geodesic does not exist in this case. As an example take the points as the two poles: all points on the equator are equally reasonable candidates for a "mean".

Again, the arithmetic mean $\boldsymbol{M} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{C}_k$ is the minimizer of the dispersion according to the Euclidean distance, i.e., $\arg\min_{\boldsymbol{M}} \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\delta_E^2\left(\boldsymbol{C}_k, \boldsymbol{M}\right) = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\left\|\boldsymbol{C}_k - \boldsymbol{M}\right\|_F^2$. We consider the corresponding problem with respect to the distance (8), yielding $\arg\min_{\boldsymbol{G}} \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\delta_G^2\left(\boldsymbol{C}_k, \boldsymbol{G}\right)$. The Riemannian metric δ_G has several nice properties ensuring that $\boldsymbol{G}$ thus defined exists and is unique. This was first established by Élie Cartan, and $\boldsymbol{G}$ is variously called the Cartan, Karcher, Fréchet, Riemannian or *geometric mean* as well as the *center of mass* of $\{\boldsymbol{C}_1,\ldots,\boldsymbol{C}_K\}$[11].

The invariance property (9) and (10) of the Riemannian distance bestow upon geometric mean $\boldsymbol{G}$ the *congruence invariance*

$$\boldsymbol{G}\left(\boldsymbol{X}\boldsymbol{C}_1\boldsymbol{X}^T, \cdots, \boldsymbol{X}\boldsymbol{C}_2\boldsymbol{X}^T\right) = \boldsymbol{X}\boldsymbol{G}\left(\boldsymbol{C}_1, \cdots, \boldsymbol{C}_2\right)\boldsymbol{X}^T, \tag{11}$$

for any invertible $\boldsymbol{X}$, and the *self-duality*

[11] The geometric mean of two positive matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ has been studied since the 1970's by electrical engineers, physicists and matrix analysts. A closed-form expression for it is known. More recent is the realization that this mean is simply the midpoint of the Riemannian geodesic connecting $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ (Fig. 5). The geometric mean of several positive matrices $\{\boldsymbol{C}_1,\ldots,\boldsymbol{C}_K\}$, though known to geometers since the work of Cartan, has entered the discourse on matrix theory and its applications more recently. It has been an object of intense study in the past 15 years and several new facts and applications have been discovered.

$$\boldsymbol{G}\left(\boldsymbol{C}_1^{-1},\cdots,\boldsymbol{C}_2^{-1}\right)=\left(\boldsymbol{G}\left(\boldsymbol{C}_1,\cdots,\boldsymbol{C}_2\right)\right)^{-1}, \tag{12}$$

both of which are very useful in applications. The arithmetic mean possesses the congruence invariance but not the self-duality[12]. Another important property is the *determinant identity*: the determinant of the geometric mean $\boldsymbol{G}(\boldsymbol{C}_{1,\ldots,}\boldsymbol{C}_K)$ is equal to the geometric mean of the determinants of $\boldsymbol{C}_{1,\ldots,}\boldsymbol{C}_K$. This is in stark contrast to the arithmetic mean, where the determinant of $½(\boldsymbol{C}_1+\boldsymbol{C}_2)$ can be larger than both the determinant of $\boldsymbol{C}_1$ and of $\boldsymbol{C}_2$[13]. When K>2, no closed-form expression for the geometric mean is known. A useful characterization of the geometric mean is as the unique solution of non-linear matrix equation

$$\tfrac{1}{K}\sum_k \mathrm{Log}\left(\boldsymbol{G}^{-½}\boldsymbol{C}_k\boldsymbol{G}^{-½}\right)=\boldsymbol{0}. \tag{13}$$

A steepest descent iterative algorithm has been proposed [81, 87]. A fixed point approximation with faster convergence rate and favorable convergence properties can be found in [88]. Appendix II points to programming code resources for estimating the geometric mean.

In summary, the Riemannian metric on the manifold of positive matrices provides us with an appropriate definition of distance and mean functions, regardless of the dimension *N*. Thanks to these two functions we can apply the MDM classifier in the same way for any dimension of the data (i.e., any number of electrodes) and any number of classes. Based on this level of generality, we will now answer the natural question: why such a simple decoding approach works well in practice?

[12] The dual of the arithmetic mean is another famous Pythagorean mean, the *harmonic mean*, both for positive real numbers and positive matrices. For scalars, this is the Fréchet mean corresponding to the distance function $d_H\left(a,b\right)=\left|a^{-1}-b^{-1}\right|$ on $(0,\ \infty)$. Only the geometric mean is its own dual, as per (12). Another interesting property of the geometric mean is that the geometric mean of the arithmetic and the harmonic mean is the geometric mean of the set.

[13] This is meaningful, for instance, in interpolating points along geodesics. Let $\boldsymbol{C}$ be the covariance matrix of a multivariate *N*-dimensional Gaussian process, with eigenvectors $\boldsymbol{u}_n$ and corresponding eigenvalues λ_n. The determinant of $\boldsymbol{C}$ is the *generalized variance* of the process [85, 86]. The square root of this determinant is proportional to the volume of an ellipsoid whose axes point in the direction of $\boldsymbol{u}_n$, and whose length is proportional to the square root of λ_n. Therefore, we require that the determinant of a point $\boldsymbol{C}$ interpolating between $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ be continuous and monotonic along the geodesic joining $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. Imagine an ellipsoid whose shape and volume smoothly changes from that of $\boldsymbol{C}_1$ to that of $\boldsymbol{C}_2$ (see for example [84]).

3. Why the Riemannian Metric?

3.1 Equivalence between Sensor Space and Source Space

Much research effort in the BCI field have been injected in feature extraction methods based on spatial filters and source separation. The aim of these methods is to decompose the sensor measurement in a *signal part* plus a *noise part* and feed a classifier only with features extracted on the signal part. The principal reason why manipulating data points by Riemannian geometry is effective is that the Riemannian manipulations in the sensor space are equivalent to those that can be done in the source space of the same dimension. To see this, consider first the nature of spatial filters.

Spatial Filters: Given the N-dimensional observed EEG measurement vector $\boldsymbol{x}(t)$, a spatial filter is a PxN matrix $\boldsymbol{B}$, with $0<P\leq N$, realizing linear combinations of samples $\boldsymbol{x}(t)$ possessing a desired optimal property. Any linear spatial filter can be written as

$$\boldsymbol{y}(t)=\boldsymbol{B}\boldsymbol{x}(t). \tag{14}$$

Spatial filters differ from one another depending on how matrix $\boldsymbol{B}$ is derived. Borrowing from standard machine learning techniques, the extracted components $\boldsymbol{y}(t)$ are usually forced by construction of the filter to be uncorrelated and their number, P, is usually chosen smaller than the number of electrodes, N, wherein the discarded N-P components explain EEG energy not related to the task, i.e., the noise suppressed by the filter.

A specific class of spatial filters is the family of blind-source separation methods [89]. While the components $\boldsymbol{y}(t)$ of a spatial filter do not need to have any physiological meaning, source separation components are estimations of the *waveform of actual brain sources* generating the observed EEG scalp measurement. It is universally accepted that the observed EEG is well approximated by a linear mixture of brain dipolar sources [90], so that we usually employ as EEG data generation model

$$\boldsymbol{x}(t)=\boldsymbol{A}\boldsymbol{s}(t), \tag{15}$$

where ***s(t)*** is a vector holding the unknown source processes and ***A*** is the *mixing matrix*, assumed here invertible, with its left-inverse ***B*** named the *demixing matrix*. The mixing matrix depends on the dipole position and orientation in the brain, the physical properties of the head and on the position of the electrodes on the scalp. Once estimated ***B***, the source process are estimated by (14), out of the usual scaling and permutation ambiguities [90, 91].

Now, let $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ be the covariance matrix of the unknown source process for any two trials. From (15), the two corresponding sensor covariance matrices are $\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T$ and $\boldsymbol{C}_j = \boldsymbol{A}\boldsymbol{S}_j\boldsymbol{A}^T$. Because of the congruence invariance property of the Riemannian distance (9) we have this insightful result:

$$\text{If } P=N \text{ then } \delta_G\left(\boldsymbol{S}_i, \boldsymbol{S}_j\right) = \delta_G\left(\boldsymbol{C}_i, \boldsymbol{C}_j\right), \tag{16}$$

i.e., if we estimate a square spatial filter the distance in the sensor space ($\boldsymbol{C}_i$ and $\boldsymbol{C}_j$) is equivalent to the distance in the source space ($\boldsymbol{S}_i$ and $\boldsymbol{S}_j$), it does not matter how optimal for classification the extracted components are. We say that EEG spatial mixing is an isometry in the Riemannian space. From a classification point of view, in a given space dimension the congruence invariance allows to keep information of the feature space *for whatever change of coordinates*. If we take $P<N$, meaning that we estimate less components than available sensors, we can still find a projection in a source sub-space enhancing the separation of the classes, that is, we can still improve the classification achieved by the MDM as applied in the sensor space. We will see in section 4 that in practice the improvement is worth the effort of estimating a spatial filter only if the number of available electrodes is large. This happens because the Riemannian distance is robust with respect to noisy components and if N is small the number of noisy components is small as well.

3.2 Robustness of Geometric Mean

As it is often reported, one of the major challenges of EEG-based BCI is the fact that EEG data are contaminated by several sources of artefacts, including biological, environmental and instrumental artefacts. These contaminations are better controlled in a laboratory as compared to real-world situations. An obvious reason why the Riemannian metric proves advantageous is the robustness of the geometric mean to *outliers*. This is illustrated in Fig. 3; when outliers

are present, the geometric mean deviates less from the center of the distribution as compared to the arithmetic mean, both in the case of data distributed as Gaussian (symmetric) and Chi-Squared (asymmetric). The distortion of the arithmetic mean is more pronounced for the Chi-Squared data. Such robustness of the geometric mean is directly inherited from the geometric distance function (see section 2).

3.3 Generalization Capabilities

The invariance by congruence of the Riemannian distance (9) is also instrumental to make the whole Riemannian framework more robust to modifications of the EEG source spatial distributions typically observed across sessions and across subjects. The true mixing matrix $\boldsymbol{A}$ in (15) is highly specific to a single subject and changes also for different sessions of the same subject because of unavoidable different position and impedance of the electrodes.

Consider cross-session learning first. Let again $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ be the covariance matrix of the unknown source process for any two trials and $\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T$, $\boldsymbol{C}_j = \boldsymbol{A}\boldsymbol{S}_j\boldsymbol{A}^T$ be the two corresponding sensor covariance matrices. Now consider two trials in another session with the same source process covariance matrices $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ and let $\tilde{\boldsymbol{A}}$ be the mixing matrix of the new session. The sensor covariance matrices of the new session are $\boldsymbol{Q}_i = \tilde{\boldsymbol{A}}\boldsymbol{S}_i\tilde{\boldsymbol{A}}^T$ and $\boldsymbol{Q}_j = \tilde{\boldsymbol{A}}\boldsymbol{S}_j\tilde{\boldsymbol{A}}^T$. So, in the two sessions we observe different covariance matrices albeit the source covariance is identical ($\boldsymbol{S}_i$ and $\boldsymbol{S}_j$). We have this remarkable result: the distance between the two observed covariance matrices is the same in the two sessions, i.e.,

$$\delta_G\left(\boldsymbol{C}_i, \boldsymbol{C}_j\right) = \delta_G\left(\boldsymbol{Q}_i, \boldsymbol{Q}_j\right). \tag{17}$$

The proof is straightforward: conjugating both $\boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T$ and $\boldsymbol{A}\boldsymbol{S}_j\boldsymbol{A}^T$ by $\boldsymbol{A}^{-1}$ (a congruence) we obtain $\delta_G\left(\boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T, \boldsymbol{A}\boldsymbol{S}_j\boldsymbol{A}^T\right) = \delta_G\left(\boldsymbol{S}_i, \boldsymbol{S}_j\right)$ and conjugating both $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ by $\tilde{\boldsymbol{A}}$ we obtain the desired result (17). Note that isometry (17) holds it does not matter how far $\tilde{\boldsymbol{A}}$ is from $\boldsymbol{A}$. In contrast, the more $\tilde{\boldsymbol{A}}$ differs from $\boldsymbol{A}$, the more the performance of a spatial filter applied across-session will degrade. An empirical validation of this property is illustrated Fig. 6.

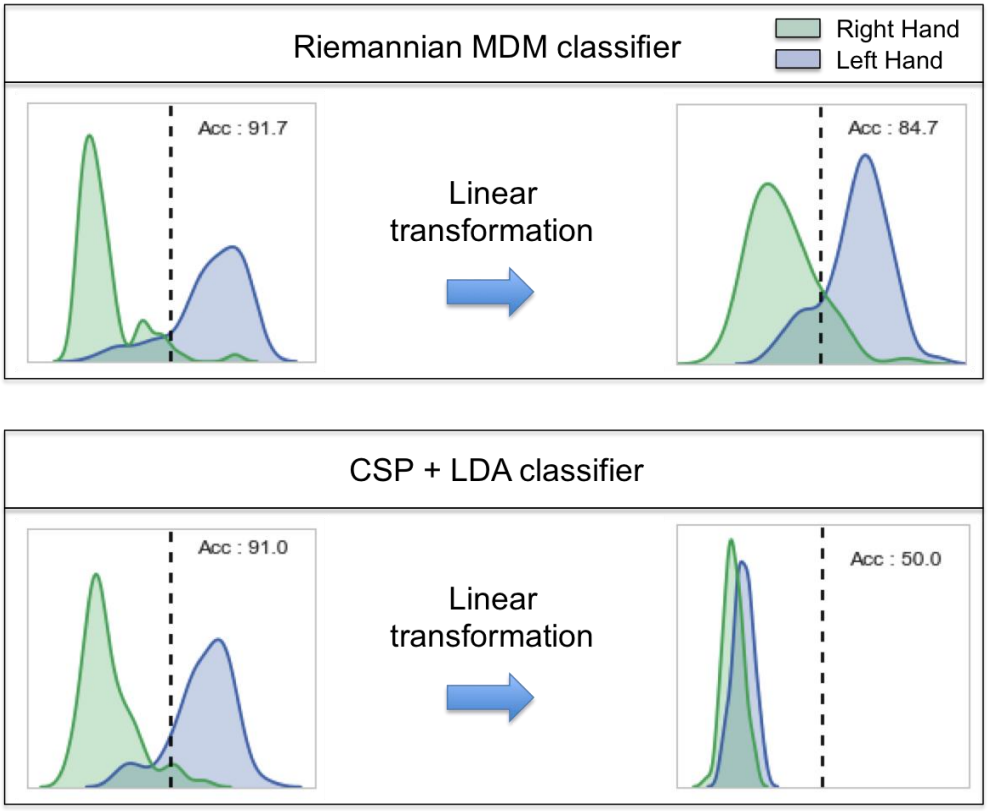

Figure 6: Illustration of Cross-Session Transfer Learning. *Distributions of the classification output for a two-class motor imagery classification problem. After applying a random linear transformation, the Riemannian classifier conserve a relatively good accuracy, while the classification based on the common spatial pattern filter (CSP: see section 4) is unable to capture relevant information and outputs random labels.*

In cross-subject learning the source process of different subjects is different in addition to their mixing matrix. Thus the deterioration of the performance will be higher for both the Riemannian MDM and a classifier based on spatial filtering, however, the former will still be superior thanks to the aforementioned isometry of the mixing process. This has been shown on BCI data in [58] - see figure 8.5 therein.

4. Comparison with the Common Spatial Pattern

In this section we compare the Riemannian MDM algorithm to the state-of-the-art spatial filter known as common spatial pattern (CSP), which has proven flexible, simple and accurate in controlled conditions [55, 93]. The CSP has been introduced in the machine learning arena by Fukunaga and Koontz in 1970 [93] and adopted for EEG data analysis by Koles [94, 95].

In the BCI field it has been introduced in [96]. In the field of pattern recognition it is known also as the Fukunaga-Koontz transform (FKT: see for example [97]). In the signal processing literature one can find thousands of papers solving similar optimization problems, which always, like for the CSP, boil down to a generalized eigenvector-eigenvalue decomposition. Here we show that the CSP is intimately related to the more general framework offered by the Riemannian MDM algorithm. First, consider the rationale of spatial filters and particularly how the CSP filter is constructed.

Let as consider again the general multidimensional case ($N \geq 2$) in a BCI where trials of left hand and right hand motor imagery are to be classified given a number of training trials. Let $\boldsymbol{C}_A$, $\boldsymbol{C}_B$ be estimations of the expected covariance matrix related to the two classes A and B (the order of classes is not important). They may be estimated by the arithmetic average of the training trials covariance matrices, as it is usually done, as well as by their Riemannian geometric mean [98, 99].

The Common Spatial Pattern (CSP): the matrix $\boldsymbol{F}$ holding in the rows the first $P/2$ and last $P/2$ eigenvectors of $N \times N$ matrix $\boldsymbol{C}^{-1}\boldsymbol{C}_A$, whenever for any $\alpha,\beta > 0$ linear combination $\boldsymbol{C} = \alpha\boldsymbol{C}_A + \beta\boldsymbol{C}_B$ is positive, satisfies

$$\begin{cases} \boldsymbol{F}\boldsymbol{C}_A\boldsymbol{F}^T = \boldsymbol{D}_A \\ \boldsymbol{F}\boldsymbol{C}_B\boldsymbol{F}^T = \boldsymbol{D}_B \\ \boldsymbol{F}\boldsymbol{C}\boldsymbol{F}^T = \boldsymbol{D} \end{cases}, \tag{18}$$

where D_A, D_B and D are $P \times P$ diagonal matrices with positive diagonal elements $d_{A,p}$, $d_{B,p}$ and d_p, respectively, with $p=\{1,\ldots,P\}$ and $1<P<N$ ([100], p.28-34; [101], p.160-165). The CSP is usually defined taking $\boldsymbol{C}=\frac{1}{2}\boldsymbol{C}_A+\frac{1}{2}\boldsymbol{C}_B$ and scaling $\boldsymbol{F}$ such that $\boldsymbol{F}\boldsymbol{C}\boldsymbol{F}^T=\boldsymbol{I}$, after which the vectors $\boldsymbol{f}_1,\cdots \boldsymbol{f}_P$ of $\boldsymbol{F}$ are the solution to the P optimization problems

$$\max_{\boldsymbol{f}_1,\ldots,\boldsymbol{f}_P} \frac{d_{A,p}}{d_{B,p}}, \quad w.c. \quad (18) \quad \text{and} \quad d_{A,p} + d_{B,p} = 1, \ \text{for all } p\,. \tag{19}$$

Note that the Riemannian distance between $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$ is a function of the eigenvalues of matrix $\boldsymbol{C}_B^{-1}\boldsymbol{C}_A$ (8), thus the P most extreme eigenvalues forming this distance are associated with the eigenvectors forming the CSP filter $\boldsymbol{F}$. The elements $d_{A,p}$ and $d_{B,p}$ are the variances of the data filtered by the CSP filter as per (14). Each associated eigenvector f_p projects the data covariance in a mono-dimensional space where the ratio of the variance in the two classes is maximized (19), that is, the points are aligned so as to be maximally separated; the first $P/2$ vectors explains the maximum of the variance of class A and the minimum of class B, while the last $P/2$ vectors explains the maximum of the variance of class B and the minimum of class A. Consider now two unlabeled trials with filtered covariance matrix $\boldsymbol{E}_i = \boldsymbol{F}\boldsymbol{C}_i\boldsymbol{F}^T$ and $\boldsymbol{E}_j = \boldsymbol{F}\boldsymbol{C}_j\boldsymbol{F}^T$, with diagonal elements denoted $e_{i,P}$ and $e_{j,p}$, respectively. Matrices $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ are not diagonal, thus their diagonal elements are not their eigenvalues. As a consequence, the Riemannian distance between the two trials $\delta_G\left(\boldsymbol{E}_i, \boldsymbol{E}_j\right)$ cannot be expressed as a function of the diagonal elements $e_{i,p}$ and $e_{i,p}$. In symbols, we have the following inequality:

$$\sum_p \left(\log e_{i,p} - \log e_{j,p}\right)^2 \leq \delta_G^2\left(\boldsymbol{E}_i, \boldsymbol{E}_j\right), \tag{20}$$

with equality only if $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ are diagonal. In (20) the left-hand side is the sum of the P geometric distances between the features $e_{i,p}$ and $e_{j,p}$ extracted by the CSP filter (the CSP features) and the right-hand side is the total Riemannian distance between $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ (the Riemannian feature). We see that the CSP approximates the distances considered by the MDM algorithm, with the approximation getting closer as $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ approach diagonal form. The approximation is in general good due to the noise suppression operated by the CSP ($P \leq N$), but the method is not robust for trials contaminated by noise making $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ far from diagonal form.

The robustness of the Riemannian MDM method with respect to noise comes handy in real-world BCI; with N small enough (<32) the difference in accuracy obtained by the CSP and Riemannian MDM methods is negligible [58, 102], whereas for N large the CSP proves superior because more and more irrelevant components are ignored by the CSP ($P \ll N$), but not in the computation of the Riemannian distance between $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$. As we have seen in section 3.1, one may apply a CSP transformation and then apply the MDM in the reduced (denoised) space, i.e., on $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$, exploiting the advantages of both approaches, as it has

been done in [103]. This way higher accuracy may be obtained, however we need again to estimate a spatial filter and unless we adapt it on line such a filter is specific to the available training data, losing generalization power. Also, the Riemannian distance and mean are robust with respect to the noise if we do not filter the data whenever N is small. In conclusion, in real-world applications, where a small number of electrodes is applied, the MDM method offers a very competitive option, in that high accuracy is obtained without estimating spatial filters.

5. Advanced Riemannian Classifiers

The MDM algorithm is the simplest Riemannian approach. In this section we briefly report on more sophisticated Riemannian classifiers, which are capable to clearly outperform the CSP and other state-of-the-art methods. Those make use of the concept of *tangent space*. Tangent space mapping is a local projection that maps the elements of the manifold into an Euclidean space, keeping their distance relationship intact. This operation can be visualized as a *local* unfolding of the manifold curved structure into a plane (Fig. 5). Once projected into the tangent space, data points (i.e., positive matrices) can be vectorized so as to form a standard feature vector and thus fed into any standard classifier (e,g., linear discriminant analysis, logistic regression, support-vector machine, etc.) This operation allows the opportunity to produce complex decision functions, depending on the chosen classifier in the tangent space, combining the advantages of the chosen classifier with the advantages of using a Riemannian metric, as we have discussed. While the overall performance of Riemannian methods based on tangent space mapping is superior to the MDM method and clearly outperform the state of the art [102], they are less suitable for online operation because of the increased algorithmic complexity and possible need of intense learning inherited by the classifier. Studies that have been employing the MDM and the tangent space approach for BCI data will be considered in next section.

6. A review of studies applying Riemannian Geometry to EEG

The use of Riemannian geometry in the EEG arena has been introduced independently in two groups, for classifying sleep stages [104-107] and for BCI based on motor imagery [102, 108-110] and P300 [58, 111]. The interest in these works has led to a rapid follow up by the EEG community, especially in the BCI field: the connections of the CSP algorithm and the tools of information geometry have been investigated considering several divergence functions in alternative to the Riemannian distance [112-115]. Reference [98, 99] successfully tested the Riemannian geometric mean for estimating the average class covariance matrices for CSP spatial filters. They concluded that this choice should be preferred for N small, which is in line with our theoretical analysis. Reference [11] proposed a simple data augmentation approach for improving the performance of the Riemannian MDM algorithm. A hierarchical MDM classifier for multi-class problem has been tested in [117]. An adaptive (online) MDM algorithm has been implemented and tested in [58, 98, 102] for motor imagery-based and ERP-based BCIs and in [116] for SSVEP-based BCIs. Cross-subject learning with the MDM Riemannian classifier has been treated in a supervised fashion in [58, 118] and in an unsupervised fashion in [119, 120]. The use of power means as a generalization of the geometric mean has been proposed in [88]. The MDM algorithm for multi-user P300-based BCIs has been tested in [46, 47].

Advanced Riemannian classifiers based on the tangent space on the Riemannian manifold of positive matrices is also receiving increasing attention: in [102] it was shown that motor imagery classification can be improved significantly over the MDM by mapping the covariance matrices in the tangent space and applying a feature selection + LDA in the tangent space. In [110] a support vector machine embedded with a Riemannian kernel was used. These two methods outperform the state of the art, but they require tuning parameters to be learned. In [98] an extension of the Fisher discriminant analysis to the manifold of positive matrices was proposed. Matrices were mapped in the tangent space where a supervised projection of the points (regularized LDA with automatic regularization) was applied in order to increase the class separation and then remapped into the Riemannian manifold where the MDM was applied. This letter method does not require tuning parameters, but it still is more involving as compared to the simple MDM and inferior to complex decision functions taken in the tangent space. In [121] simultaneous EEG and EMG recordings were analyzed in the tangent space by means of canonical partial least squares to reconstruct spatial patterns in a

pinch-and-hold task. A multi-class linear discriminant analysis in the tangent space was proposed in [122]. In [123] it has been showed that the geometric mean of data is not the optimal basepoint for defining the tangent space and that a trimmed geometric mean can improve the classification performance in motor imagery-based BCI. Taken together these studies have shown that using more sophisticated classification methods in the Riemannian framework one may outperform the state of the art, but only at the expenses of the ergonomic requirements of the BCI as considered in the introduction.

Metric/kernel learning, a hot topic in the machine leaning community, has been investigated in [124-125]. In parallel, geometry-aware dimensionality reduction (see section 3.1) inspired by the Riemannian framework is currently intensely investigated. Works related to BCI data include [126-131]. A relevant work, which can be readily borrowed from the computer vision community, is [132, 133].

Outside the BCI field, but still using EEG data, Riemannian classifiers have been tried in the diagnosis of clinical populations such as bipolar disorders and schizophrenia [134], in the detection of EEG artifacts [41], epileptic seizures [135], mental fatigue [136], respiratory states with use for a brain-ventilator interface [137] and sleep apnea events [138]. The geometric mean of event-related potentials (ERPs) have been studied in [139] as an example of a general method for visualizing and computing means of data in their native space, which is otherwise impossible once covariance matrices have been used to summarize the data point. In that work we have illustrated the robustness of the geometric mean with respect to outliers by computing the geometric mean of event-related potentials.

It should be noted that articles using Riemannian geometry are currently appearing in several and diverse applied fields that are often completely disconnected. Nonetheless, developed methods are very similar, thus more or less interesting advances, possibly relevant in the BCI field, may be found in other scientific domains. Also, the Riemannian geometry framework is not limited to theoretical studies and applications to classification problems. An expanding line of research on Riemannian optimization is today at the forefront in the signal processing community and is gaining attention in biomedical engineering. The mathematical connection between the concept of Riemannian mean and the blind source separation/independent component analysis problem in EEG data has been explored in [91].

Finally, it is worth mentioning that Riemannian methods (MDM and/or tangent space projection) submitted by author A.B. in five BCI-related *predictive modeling* competitions have scored first over 352 competitors on the average (table 1). These results provide overwhelming evidence of the great potential of this approach and have been instrumental for the current attention Riemannian methods are receiving in the BCI community.

Table 1: *Results of five international predictive modeling data science competitions on BCI data where author* A.B. has scored first in all occasions. *Results in bold concern a submission that has used Riemannian geometry. Scores : Percent Accuracy for DecMeg 2014 and Decoding Brain Signal (chance level=50%) and area under the ROC for the other three (chance level=0.5). For Grasp and Lift EEG Challenge the reported area is the average across the 6 classes (see Appendix II for references to code submitted in these competitions).*
Legend: * *https://www.kaggle.com*; ° *https://gallery.cortanaintelligence.com*;
+ *https://sites.google.com/site/hubertcecotti/home/biomag2016*

Name of Challenge	***Associated Event or Organizer***	***Closing Date***	***Data Type (# of Classes)***	***Challenge***	***Participants (Teams)***	***Three Best Scores***
DecMeg 2014*	BIOMAG 2014 Conference	27/07/2014	MEG visual ERF (2 classes)	Cross-subject transfer learning	301(267)	**75.5** 72.6 71.3
BCI Challenge*	IEEE NER 2015 Conference	24/02/2015	EEG Error Potentials (2 classes)	Cross-subject transfer learning. Unbalanced data	311(260)	**0.872** 0.856 0.818
Grasp & Lift EEG Challenge*	WAY European Project	31/08/2015	Hand Movement Sequences (6 classes)	Multiclass asynchronous prediction of time structured events	452(379)	**0.9810** 0.9802 0.9799
Decoding Brain Signal °	Microsoft	01/07/2016	ECoG Face vs. House (2 classes)	Classification of ERP and induced activity	688 (688)	**93.75** 92.5 **88.33**
Biomag2016 competition #3 +	BIOMAG 2016 Conference	25/09/2016	MEG visual ERF (2 classes)	Classification of facial expression	7(5)	**0.956** **0.866** 0.771

7. Conclusions and Discussion

Riemannian geometry is a relatively new classification framework operating a paradigmatic shift in the field of brain-computer interface. In this article we have provided a primer on the simpler Riemannian classification method, the minimum distance to mean (MDM), providing rationale on its efficacy without requiring any specific knowledge on differential geometry.

Instead, we have highlighted the simplicity of its application in practice and we have relied mainly on intuitive (geometrical) explanations. The Riemannian MDM approach is based entirely on two simple concepts: the *distance* between two data points and a *mean* of a number of them. These two concepts are fundamental in many branches of mathematics. They can be easily understood and easily communicated to peers, yet, they allow a precise and robust classification method comparing well with the performance of more complex state-of-the-art methods. To make a metaphor, it appears that we have started a long time ago measuring distances with a biased ruler (Euclidean distance for variances and covariance matrices). Since it did not work well, we have developed complex instruments in order to replace the malfunctioning ruler (spatial filters and other matrix decomposition methods). Finally, we have found a valid ruler (the Riemannian metric for SPD matrices) and we can now come back to the simple concept of measuring distances between observables for classifying BCI data. Riemannian geometry provides the natural framework to treat symmetric positive-definite matrices and many kind of structured covariance matrices are of this type. It does not matter how the covariance matrices are defined, the MDM Riemannian classifier remains the same for all the three BCI modalities, namely, motor/mental imagery, event-related potentials and steady state evoked potentials. Defining appropriate covariance matrices embedding relevant information depending on the data in the BCI field is the task of brain scientists, linking electrophysiological knowledge to mathematical formalism. Formal definitions for the three main BCI modalities are given in Appendix I. These definitions can be improved and further research on this topic is expected.

A distinctive approach of the Riemannian MDM algorithm is that at no point there is a parameter to be tuned; it is all deterministic and completely parameter-free. This is in contrast with more sophisticated machine learning approach such as support vector machine, where one or more parameters must be optimized, typically, by cross-validation [55]. For this reason we claim that the MDM approach can be used purposefully in all BCI modalities; In fact, taken together its simplicity, its ability to learn rapidly (with little training data), its good cross-subject and across-session generalization, it suits well real-world BCIs according to requirements a)-h) listed in the introduction.

Most research on BCI signal processing still today proceeds by improving spatial filters [75, 140, 141], thus we may say that since the inceptions of the CSP more than 15 years ago, the first major paradigmatic shift that has been operated in the BCI arena has been the

introduction of Riemannian geometry. We have observed that refinements of spatial filters bring upon only moderate improvement for classification purposes and that such improvement does not translate easily in a significant increase of reliability and robustness. In fact, a spatial filter is by its very nature highly specific to the data on which it has been estimated, that is to say, it is intrinsically subject- and session-specific, which jeopardizes its robustness and constitutes a major limit to achieve effective transfer learning. Nonetheless, the CSP and related methods are efficient option in a classical test-training mode of operation, as more than 15 years of practice has demonstrated.

Aknowledgements

This research has been partially founded by European project ERC-2012-AdG-320684-CHESS.

Appendix I

For all BCI modalities we work in practice with *trials*, which are time-windows of EEG data. We indicate them by $\boldsymbol{X}z$, with $z \in \{1,..,Z\}$ the index of Z classes. Trials $\boldsymbol{X}z$ are $N \cdot T$ EEG data matrices, where N is the number of electrodes and T the number of samples. The MDM works in the same way for all BCI modalities, however, for each of them the way covariance matrices are defined differs. In fact, the performance of the MDM depends on our ability to capture all relevant information related to the task in the form of a positive matrix. In this appendix we detail appropriate covariance matrix definitions for the three main BCI modalities. Far from being exhaustive, these definitions can be further enriched and improved. In the sequel, the definitions are the same for the training trials and for the test trials; once obtained positive matrices from the available trials, one can compute the geometric mean of each class on training data and measure distances of unlabeled trials to these means.

Classification of Motor/Mental Imagery

The sample covariance matrix of trial $\boldsymbol{X}z$ is

$$\boldsymbol{C}_z = \frac{1}{T-1}\left(\boldsymbol{X}_z \boldsymbol{X}_z^T\right). \tag{21}$$

The diagonal elements hold the variance of the signal at each electrode and the off-diagonal elements hold the covariance between all electrode pairs. The sample covariance matrix contains all *spatial information*, particularly, its second order statistics. If the data are centered (zero mean) and they follows a multivariate Gaussian distribution, those statistics describe the spatial process exhaustively. Sample covariance matrices suffice for classifying motor/mental imagery (MI) trials because MI trials for different classes do indeed produce a different scalp spatial pattern [142]. Then for MI we just use definition (21). The only other pre-processing step requested is filtering the data in the frequency band pass regions involved in the task, for example, for motor imagery, 8-30 Hz, in order to reduce noise. If several frequency band-pass regions are of interest, one can estimate the geometric mean for each region separately and then sum the (squared) distance obtained in all regions, a procedure we will encounter for the case of steady state evoked potentials.

Classification of Event-Related Potentials (ERPs)

For ERP-based BCI the sample covariance matrix (21) is not efficient since ERPs feature amplitude much smaller as compared to the background EEG, thus the spatial structure contained in the covariance matrix of a single trial does not hold sufficient information for classification. As a matter of fact the sample covariance matrix (21) does not contain *any temporal information at all*, which is easily realized if we consider that shuffling at random the samples of trial $\boldsymbol{X}_z$ the sample covariance matrix (21) is unchanged. However ERPs have a specific time signature; it is this signature that differentiates an ERP from another or an ERP from the absence of the ERP, so this is the information we need to extract and embed in a "covariance matrix". In order to do so let us consider again trials $\boldsymbol{X}_z$, for $z \in \{1,\ldots,Z\}$ classes (Single-trial ERPs). In this case each class corresponds to a different ERP and a no-ERP class is usually added. For example, in P300-based BCI, one class is the *target* class, containing a P300, and the other is the *non-target* class (Z=2). Let us now construct the *super-trial*

$$\boldsymbol{X}_z^{ERP} = \begin{pmatrix} \bar{\boldsymbol{X}}_{(1)} \\ \cdots \\ \bar{\boldsymbol{X}}_{(Z)} \\ \boldsymbol{X}_z \end{pmatrix} \in \Re^{N(Z+1)xT}, \tag{22}$$

where $\bar{\boldsymbol{X}}_{(1)}^T, \cdots, \bar{\boldsymbol{X}}_{(Z)}^T$ are the ensemble average ERPs obtained on the training data, on previous sessions of the user or even on a database of other users (transfer learning using a grand average). Note that we have introduced index (*z*) in parenthesis to highlight the difference with the z^{th} training class of the trial $\boldsymbol{X}_z$. We call the averages in Eq. (22) the *temporal prototypes*. An appropriate method for estimating them when ERPs overlap, which is usually the case in BCI systems, is described in [143]. Because of the invariance by congruence, theses temporal prototypes can be estimated in the sensors space or in the source space without any loss of performance. For instance, a principal component analysis can be applied to a grand average, retaining only a few components, with ensuing computational advantage. We specify a prototype for each class. Now, for a training trial $\boldsymbol{X}_z$ the covariance matrix of the super-trial has the following block structure:

$$\boldsymbol{C}_z = \frac{1}{(T-1)}\left(\boldsymbol{X}_z^{ERP}\left(\boldsymbol{X}_z^{ERP}\right)^T\right) = \frac{1}{(T-1)}\begin{pmatrix} \bar{\boldsymbol{X}}.\bar{\boldsymbol{X}}.^T & \left(\boldsymbol{X}_z\bar{\boldsymbol{X}}.^T\right)^T \\ \boldsymbol{X}_z\bar{\boldsymbol{X}}.^T & \boldsymbol{X}_z\boldsymbol{X}_z^T \end{pmatrix} \in \Re^{N(Z+1)x\,N(Z+1)}, \tag{23}$$

$$\text{where } \bar{\boldsymbol{X}}.\bar{\boldsymbol{X}}.^T = \begin{pmatrix} \bar{\boldsymbol{X}}_{(1)}\bar{\boldsymbol{X}}_{(1)}^T & \cdots & \bar{\boldsymbol{X}}_{(1)}\bar{\boldsymbol{X}}_{(Z)}^T \\ \vdots & \ddots & \cdots \\ \bar{\boldsymbol{X}}_{(Z)}\bar{\boldsymbol{X}}_{(1)}^T & \cdots & \bar{\boldsymbol{X}}_{(Z)}\bar{\boldsymbol{X}}_{(Z)}^T \end{pmatrix} \in \Re^{NZxNZ} \tag{24}$$

$$\text{and } \boldsymbol{X}_z\bar{\boldsymbol{X}}.^T = \left(\boldsymbol{X}_z\bar{\boldsymbol{X}}_{(1)}^T, \cdots, \boldsymbol{X}_z\bar{\boldsymbol{X}}_{(Z)}^T\right) \in \Re^{NxNZ}. \tag{25}$$

Let us take a close look to the structure of this covariance matrix:

The *N*x*N* diagonal blocks of $\bar{\boldsymbol{X}}.\bar{\boldsymbol{X}}.^T$ in (23) - see (24) for the relevant expansions - hold the covariance matrices of the *Z* temporal prototypes and its *N*x*N* off-diagonal blocks their cross covariance. All these blocks are not useful for classification since they, being based on fixed prototypes, do not change from trial to trial.

The *N*x*N* block $\boldsymbol{X}_z\boldsymbol{X}_z^T$ in (23) holds the sample covariance matrix (21) of the trial $\boldsymbol{X}_z$, which contains the *spatial information* of the trial and will be little useful for classification, as we have said.

The *N*x*N* blocks of $\boldsymbol{X}_z\bar{\boldsymbol{X}}.^T$ in (23) - see (25) for the relevant expansion - contains the cross-covariances between the trial and the *Z* prototypes, that is, these blocks contain the *temporal covariances*. Notice that shuffling the samples of the trial now *does disrupt* these covariances. These blocks contain the relevant information for classification as the cross-covariance will be large only in the blocks where the class of the trial coincides with the class of the prototype. The only other pre-processing required is to filter the data in the frequency band pass region containing the ERPs, typically 1-16 Hz, so as to reduce noise.

It is worth mentioning that often we deal only with the presence and absence of an ERP, as it is the case of P300-bases BCIs, where there are only two classes, a *target* (P300 present) and *non-target* (P300 non-present) class. In this case one can equivalently use a simplified version of super-trial (22) given by

$$\boldsymbol{X}_z^{P300} = \begin{pmatrix} \bar{\boldsymbol{X}}_{(+)} \\ \boldsymbol{X}_z \end{pmatrix} \in \Re^{2N\text{x}T}, \tag{26}$$

where $\bar{\boldsymbol{X}}_{(+)}^T$ is the temporal prototype of the P300 (target class).

Finally, notice that in typical P300-based BCIs such as spellers we classify after several ERPs have been collected, that is, after several repetitions of exhaustive flashing of all elements. The classification then is based either on the cumulating sum of (squared) distances across repetitions or on a single distance obtained on the average trial computed across repetitions, the two approaches being equivalent.

Classification of Steady-State Evoked Potentials

We make here the example of steady-state visually evoked potentials (SSVEP). The *Z* classes here represent *F* different flickering frequencies and a no-flickering class can be added as well, if sought. In this case the relevant information is the diversity of the frequencies

engendering oscillations in the visual cortex, while the spatial pattern may be the same for different frequencies. In order to exploit the frequency diversity we construct super trial

$$\boldsymbol{X}_z^{SSEP} = \begin{pmatrix} \boldsymbol{X}_{(1)} \\ \cdots \\ \boldsymbol{X}_{(F)} \end{pmatrix} \in \Re^{NFxT} \tag{27}$$

where $\boldsymbol{X}_{(f)}^T$ is the trial filtered in the band-pass region for flickering frequency $f \in \{1,...,F\}$. The covariance matrix of super-trial (27) has the following block structure:

$$\boldsymbol{C}_z = \frac{1}{(T-1)}\left[\boldsymbol{X}_z^{SSEP}\left(\boldsymbol{X}_z^{SSEP}\right)^T\right] = \frac{1}{(T-1)}\begin{pmatrix} \boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T & \cdots & \boldsymbol{X}_{(1)}\boldsymbol{X}_{(F)}^T \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_{(F)}\boldsymbol{X}_{(1)}^T & \cdots & \boldsymbol{X}_{(F)}\boldsymbol{X}_{(F)}^T \end{pmatrix} \in \Re^{NFxNF}. \tag{28}$$

The *N*x*N* diagonal blocks holds the covariance matrices of the *F* frequencies. When comparing an unlabeled trial with the mean of the different classes, only the mean with the block indexing the frequency corresponding to the frequency of the trial will have large values (see [58]). Thus the diagonal blocks will be useful for classification. On the other hand the off-diagonal blocks hold the cross-covariance between frequencies, thus are not meaningful. We can put them to zero since the resulting matrix

$$\boldsymbol{C}_z = \frac{1}{(T-1)}\begin{pmatrix} \boldsymbol{X}_{(1)}\boldsymbol{X}_{(1)}^T & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X}_{(F)}\boldsymbol{X}_{(F)}^T \end{pmatrix} \in \Re^{NFxNF} \tag{29}$$

is still symmetric positive definite. Note that these blocks can be estimated simply by the Fourier cospectra [58, 144] corresponding to the flickering frequencies. Therefore, the only pre-processing required is either to filter the data in the frequencies corresponding to the SSVEP flickering frequencies and then compute (29) or, equivalently, estimating the blocks of (29) directly by the Fourier cospectra at the *F* flickering frequencies. Note that if the phase of the SSVEP is known thanks to precise data tagging, as it is done in [32], or code modulation is used [33], one can exploit both the frequential and the temporal information, constructing a super trial mixing the strategy used here for ERP (22) and for SSVEP (27). Finally, note that when covariance matrices have a block-diagonal structure as in (29), it is

easy to show that the Riemannian (squared) distance between two of them equals the sum of the Riemannian (squared) distance obtained on each diagonal block separately. Similarly, the geometric mean of a set of block-diagonal matrices can be more easily formed by the geometric means of the blocks. This greatly speed up computation when the number of blocks (frequencies in this case) is large.

Appendix II: Open Source Code Resources for Riemannian Geometry

Implementations for all concepts and methods found in this document are freely available as part of two open-source toolboxes written by author A.B.: *Covariance Toolbox*[14] for the programming language Matlab and *pyRiemann*[15] for the programing language Python. They share a common core of methods allowing the replication of the authors' work. The Python toolbox pyRiemann offer more flexibility and support. Specifically, in addition to support parallelization for efficient computing, pyRiemann's API is compatible with the popular machine learning library Scikit-Learn[16], allowing to pipeline any of the Riemannian methods with the most advanced classification and regression algorithms available to date (such as random forest, eXtreme gradient boosting or deep learning). In addition to the Riemannian metric, these toolboxes support several other metrics adapted to manipulation of positive matrices. Also, code source to reproduce the results achieved in the five international predictive modeling challenges (table 1) are available on the github account of author A.B.[17].

[14] https://github.com/alexandrebarachant/covariancetoolbox

[15] https://github.com/alexandrebarachant/pyRiemann

[16] http://scikit-learn.org/stable/

[17] https://github.com/alexandrebarachant

References

[1] Allison BZ, Dunne S, Leeb R, Millán JdR, Nijolt A. Toward Practical Brain-Computer Interfaces (Eds.), London: Springer; 2012.

[2] Schomer DL, Lopes da Silva F. Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Sixth Edition, Philadelpjia (PA): Lippincott Williams & Wilkins; 2011.

[3] Kübler A, Kotchoubey B, Kaiser J, Wolpaw JR, Birbaumer N. Brain-computer communication: unlocking the locked in. Psycholog Bull. 2001; 127(3): 358-75.

[4] Tan DS, Nijholt A. Brain-Computer Interfaces (Eds.). London: Springer; 2012.

[5] Wolpaw J, Wolpaw EW. Brain-Computer Interfaces: Principles and Practice. Oxford: Oxford University Press; 2012

[6] Chaudhary U, Birbaumer N. Communication in locked-in state after brainstem stroke: a brain-computer-interface approach. Ann Transl Med. 2015; 3 (S1), S29.

[7] Curado MR, Cossio EG, Broetz D, Agostini M, Cho W, Brasil FL, et *al*. Residual Upper Arm Motor Function Primes Innervation of Paretic Forearm Muscles in Chronic Stroke after Brain-Machine Interface (BMI) Training. PLoS ONE. 2015; 10(10): e0140161.

[8] Li Y, Nam CS. Evaluation of collaborative brain-computer interface for people with motor disabilities IEEE Comput Intell Mag. 2016; 11: 56-66.

[9] McCane LM, Heckman SM, McFarland DJ, Townsend G, Mak JN, Sellers EW, et *al*. P300-based brain-computer interface (BCI) event-related potentials (ERPs): People with amyotrophic lateral sclerosis (ALS) vs. age-matched controls. Clin Neurophysiol. 2015; 126(11): 2124-31.

[10] Münβinger JI, Halder S, Kleih SC, Furdea A, Raco V, Hösle A, Kübler A. Brain painting: first evaluation of a new brain-computer interface application with ALS-patients and healthy volunteers. Front Neurosci. 2010; 4: 1-11.

[11] Salisbury D, Driver S, Parsons TD. Brain-computer interface targeting non-motor functions after spinal cord injury. Spinal Cord. 2015; 53: S25-S26.

[12] Sellers EW, Donchin E. A P300-based brain-computer: initial tests by ALS patients. Clin Neurophysiol. 2006; 117(3): 538-548.

[13] Riccio A, Mattia D, Simione L, Olivetti M, Cincotti F. Eye-gaze independent EEG-based brain-computer interfaces for communication. J Neural Eng. 2012; 9(4): 045001.

[14] Volosyak I, Cecotti H, Valbuena D, Gräser A. Evaluation of the Bremen SSVEP based BCI in real world conditions. Proc. of IEEE 11th Int. Conf. on Rehabilitation Robotics; 2009 Jun 23-26. Kyoto, Japan: 322-331.

[15] Käthner I, Kübler A, Halder S. Comparison of eye tracking, electrooculography and an auditory brain-computer interface for binary communication: a case study with a participant in the locked-in state. J Neuroeng Rehabil. 2015; 12: 76.

[16] Mayaud L, Cabanilles S, Van Langhenhove A, Congedo M, Barachant A, Pouplin S, *et al.* Brain-computer interface for the communication of acute patients: a feasibility study and a randomized controlled trial comparing performance with healthy participants and a traditional assistive device, BCI, 2016, 3(4): 197-215.

[17] Lee J-H, Lim J-H, Han C-H, Kim Y-W Im C-H. Global EEG synchronization as an indicator of emotional arousal and its application for tracking emotional changes during video watching, Proc. Of the 6th Int. BCI Conf.; 2014 Sept 16-19. Graz, Austria.

[18] Lau TM, Gwin JT, McDowell KG, Ferris DP. Weighted phase lag index stability as an artifact resistant measure to detect cognitive EEG activity during locomotion. J Neuroeng Rehabil. 2012; 9: 47.

[19] Allison BZ, Pineda JZ. ERPs Evoked by Different Matrix Sizes: Implications for a Brain Computer Interface (BCI) System. IEEE Trans Neural Syst Rehabil Eng. 2003; 11 (2): 110-113.

[20] Guan C, Thulasidas M, Wu J. P300 Speller for Brain-Computer Interface. Proc. of the IEEE Int. Workshop on Biomedical Circuits & Systems. 2004 Dec 1-3: S3/5/INV- S3/13-16.

[21] Li Y, Bahn S, Nam CS, Lee J. Effects of luminosity contrast and stimulus duration on user performance and preference in a P300-based brain-computer interface (BCI). Int J Hum Comput Interact. 2014; 30: 151-163.

[22] Li Y, Nam CS, Shadden BB, Johnson SL. A P300-Based Brain-Computer Interface (BCI): Effects of Interface Type and Screen Size. Int J Hum Comput Interact. 2010; 27: 52-68.

[23] Mainsah BO, Collins LM, Colwell K, Throckmorton CS. Improving Dynamic Data Collection in P300 Spellers With a Language Model. Proc. Fifth Int. BCI Meeting: 2013 June 3-7, Pacific grove, California. 107.

[24] Kaufmann T, Völker S, Gunesch L, Kübler A. Spelling is just a click away – a user-centered brain-computer interface including auto-calibration and predictive text entry. Front Neurosci. 2012; 6(72): 1-10
[25] Kindermans P-J, Schrauwen B. Dynamic Stopping in a Calibration-less P300 Speller. Proc. Fifth Int. BCI Meeting; 2013 June 3-7, Pacific Grove, California. 075.
[26] Pinegger A, Decker L, Halder S, Faller J, Käthner I, Wriessnegger SC et *al*. Automatic pause detection during P300 web browsing, Proc. 6th Int. BCI Conf.; 2014 Sept 16-19, Graz, Austria.
[27] Congedo M, Goyat M, Tarrin N, Varnet L, Rivet B, Ionescu G, et *al*. "Brain Invaders": a prototype of an open-source P300-based video game working with the OpenViBE platform. Proc of the 5th Int BCI Conference; 2011 Sept 22-24, Graz, Austria. 280-283.
[28] Jin J, Allison BZ, Sellers EW, Brunner C, Horki P, Wang X, Neuper C. Optimized stimulus presentation patterns for an event-related potential EEG based brain-computer interface. Med Biol Eng Comput. 2011; 49(2): 181-91.
[29] Townsend G, LaPallo BK, Boulay CB, Krusienski DJ, Frye GE, Hauser CK, et *al*. A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns. Clin Neurophysiol. 2010; 121(7): 1109-20.
[30] Verhoeven T., Buteneers P, Wiersema JR, Dambre J, Kindermans PJ. Towards a symbiotic brain-computer interface: exploring the application-decoder interaction, J Neural Eng. 2015; 12: 066027.
[31] Schreuder M, Höhne J, Blankertz B, Haufe S, Dickhaus T, Tangermann M. Optimizing event-related potential based brain-computer interfaces: a systematic evaluation of dynamic stopping methods. J Neural Eng. 2013; 10(3): 036025.
[32] Jia C, Gao X, Hong B, Gao S. Frequency and phase mixed coding in SSVEP-based brain--computer interface. IEEE Trans Biomed Eng, 2011; 58(1): 200-206.
[33] Bin G, Gao X, Wang Y, Li Y, Hong B, Gao S. A high-speed BCI based on code modulation VEP. J neural eng. 2011; 8(2): 025015.
[34] Tong J, Zhu D. Multi-phase cycle coding for SSVEP based brain-computer interfaces, Biomed Eng OnLine. 2015; 14: 5.
[35] Kleih SC, Kübler A. Empathy, motivation, and P300 BCI performance. Front Hum Neurosci. 2013; 7 (642); 00642.
[36] Kasahara K, DaSalla CS, Honda M, Hanakawa T. Neuroanatomical correlates of brain-computer interface performance. NeuroImage, 2015; 110: 95-100.
[37] Mayaud L, Congedo M, Van Laghenhove A, Orlikowski D, Figère M, Azabou E, *et al.* A comparison of recording modalities of P300 event-related potentials (ERP) for brain-computer interface (BCI) paradigm. Neurophysiol Clin, 2013; 43(4): 217-227.
[38] Mayaud L, Filipe S, Petegnief L, Rochecouste O, Congedo M. Robust Brain-computer interface for virtual Keyboard (RoBIK): project results, IRBM, 2013; 34 (2): 131-138.
[39] Nijboer F, Laar B, Gerritsen S, Nijholt A, Poel M. Usability of three electroencephalogram headsets for brain-computer interfaces: a within subject comparison. Interact comput. 2015; 27(5): 500-511.
[40] Lin CT, Liao LD, Liu YH, Wang IJ, Lin BS, Chang JY. Novel dry polymer foam electrodes for long-term EEG measurement. IEEE Trans Biomed Eng. 2011; 58(5): 1200-7.
[41] Barachant A, Andreev A, Congedo M. The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. TOBI Workshop lV; 2013 Jan 23-25. Sion, Switzerland. 19-20.
[42] Sagha H, Perdikis S, Millán JdR, Chavarriaga R. Quantifying Electrode Reliability During Brain–Computer Interface Operation. IEEE Trans Biomed Eng. 2015; 62(3): 858-864.
[43] Powers JC, Bieliaieva K, Wu S, Nam CS. The Human Factors and Ergonomics of P300-Based Brain-Computer Interfaces. Brain Sci. 2015; 5(3): 318-56.
[44] Yuan P, Gao X, Allison B, Wang Y, Bin G, Gao S. A study of the existing problems of estimating the information transfer rate in online brain–computer interfaces, J neural eng. 2013; 10(2); 026014.
[45] Bonnet L, Lotte F, Lécuyer A. Two Brains, One Game: Design and Evaluation of a Multi-User BCI Video Game Based on Motor Imagery, IEEE Trans Comput Intell AI, 2013; 5 (2): 185-198.
[46] Korczowski L, Congedo M, Jutten C. Single-Trial Classification of Multi-User P300-Based Brain-Computer Interface Using Riemannian Geometry. IEEE 39th Int Conference of IEEE Engineering in Medicine and Biology Society; 2015 Aug 25-29, Milano, Italy.
[47] Korczowski L, Barachant A, Andreev A, Jutten C, Congedo M. "Brain Invaders 2": an open source Plug & Play multiuser BCI videogame, Proc. of the 6th Int. Brain-Computer Interface Meeting; 2016 May 30 June 3, Asilomar (CA), USA..
[48] Schultze-Kraft R, Görgen K, Wenzel M, Haynes J-D, Blankertz B. Cooperating Brains: Joint Control of a Dual-BCI. Proc. Fifth Int. BCI Meeting; 2013 June 3-7, Pacific grove(CA), USA. 046.

[49] Buccino AP, Keles HO, Omurtag A. Hybrid EEG-fNIRS Asynchronous Brain-Computer Interface for Multiple Motor Tasks. PLoS ONE. 2016; 11(1): e0146610.
[50] Lee JS, Park KS. A New Stimulation Method of Virtual Speller for Simultaneous P300 and SSVEP Responses. Proc. Fifth Int. BCI Meeting; 2013 June 3-7, Pacific Grove (CA), USA. 162
[51] Pfurtscheller G, Allison BZ, Brunner C, Bauernfeind G, Solis-Escalante T, Scherer R et *al*. The Hybrid BCI, Front Neurosci. 2010; 4: 42.00003.
[52] Brunner C, Birbaumer N, Blankertz B, Guger C, Kübler A, Mattia A, et *al*. BNCI Horizon 2020: towards a roadmap for the BCI community. Brain-Comput Interfaces. 2015; 2(1): 1-10.
[53] Huggins J E, Guger C, Allison B, Anderson C W, Batista A, Brouwer A-M. Fifth Int. Brain-Computer Interface Meeting: Defining the Future; 2013 June 3-7. 1(1): 27-49.
[54] Obeid I, Picone J. Bringing Big Data to neural Interfaces. Proc. Fifth Int. BCI Meeting, 2013 June 3-7, Pacific Grove (CA), USA. 180
[55] Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain-computer interfaces, J Neural Eng. 2007; 4(2):. R1–R13.
[56] Acqualagna L, Botrel L, Vidaurre C, Kübler A, Blankertz B. Large-Scale Assessment of a Fully Automatic Co-Adaptive Motor Imagery-Based Brain Computer Interface. PLoS ONE. 2016; 11(2): e0148886.
[57] Colwell K, Throckmorton C, Collins L, Morton K. Transfer Learning for Accelerated P300 Speller Classifier Training, Proc. Fifth Int. BCI Meeting, 2013 June 3-7, Pacific Grove (CA), USA. 004.
[58] Congedo M. EEG Source Analysis. HDR thesis presented at the University of Grenoble Alpes; 2013.
[59] Herweg A, Kaufmann T, Kübler A. Using Generic Models to Improve Tactile ERO-BCI performance of Low Aptitude Users. Proc. Fifth Int. BCI Meeting, 2013 June 3-7, Pacific Grove (CA), USA. 097.
[60] Jayaram V, Alamgir M, Altun Y, Scholkopf B, Grosse-Wentrup M. Transfer Learning in Brain-Computer Interfaces, IEEE Comput Intell Mag, 2016; 11 (1): 20-31.
[61] Jin J, Sellers EW, Zhang Y, Daly I, Wang X, Cichocki A. Whether generic model works for rapid ERP-based BCI calibration. J Neurosci Meth, 2012; 212: 94-99.
[62] Kindermans P-J, Schreuder M, Schrauwen B, Müller K-R, Tangermann M. True Zero-Training Brain-Computer Interfacing – An Online Study. PLoS ONE. 2014; 9(7): e102504.
[63] Ray AM, Sitaram R, Rana M, Pasqualotto E, Buyukturkoglu K, Guan C, et *al*. A subject-independent pattern-based Brain-Computer Interface. Front Behav Neurosci. 2015; 9: 269.
[64] Kindermans P-J, Verstraeten D, Schrauwen B. A Bayesian Model for Exploiting Application Constraints to Enable Unsupervised Training of a P300-based BCI. PLoS ONE. 2012; 7(4): e33758.
[65] Panicker RC, Puthusserypady S , Sun Y. Adaptation in P300 Brain-Computer Interfaces: A Two-Classifier Co-Training Approach", IEEE Tran Biomed Eng. 2010; 57(12): 2927-35.
[66] Schettini F, Aloise F, Aricò P, Salinari S, Di Mattia D, Cincotti F. Self-Calibration in an Asynchronous P300-Based BCI. Proc. Fifth Int. BCI Meeting, 2013 June 3-7, Pacific Grove (CA), USA. 124.
[67] Mak JN, Arbel Y, Minett JW, McCane LM, Yuksel B, Ryan D, *et al.* Optimizing the P300-based brain-computer interface: current status, limitations and future directions. J Neural Eng, 2011; 8(2): 025003.
[68] Jrad N, Congedo M, Phlypo R, Rousseau S, Flamary R, Yger F, et *al*. sw-SVM : sensor weighting support vector machines for EEG-based Brain-Computer Interfaces. Journal Neural Eng. 2011: 8(5): 056004.
[69] Bhatia R. Positive Definite Matrices. Princeton (NJ): Princeton University press; 2007.
[70] Levi-Civita T. Lezioni di Calcolo Differenziale Assoluto [Lessons of Absolute Differential Calculus], Roma, Italy: Alberto Stock; 1925. Italian.
[71] Lotte F, Renard Y, Lécuyer A. Self-paced Brain-Computer Interaction with Virtual Worlds: a Quantitative and Qualitative Study 'Out of the Lab', 4th Int. Brain-Computer Interface Workshop and Training Course, 2008: 373-378.
[72] Solis-Escalante T, Müller-Putz G, Pfurtscheller G. Overt foot movement detection in one single Laplacian EEG derivation. J Neurosci Methods. 2008; 175(1): 148-53.
[73] Graf AB, Bousquet O, Rätsch G, Schölkopf B. Prototype classification: insights from machine learning. Neural Comput. 2009; 21(1) : 272-300.
[74] Legendre A-M. Nouvelles methods pour la determination des orbites des comètes; avec un supplement contenant divers perfectionnemens de ces methods et leure application aux deux comètes de 1805 [New methods for the determination of the orbits of comets ; with a supplement containing different improvements of these methods and their application to the two comets of 1805]. Paris, France : Courcier; 1806. French.
[75] Roijendijk L, Gielen S, Farquhar J. Classifying regularised sensor covariance matrices: an alternative to CSP. IEEE Trans Neural Syst Rehabil Eng. 2016; 24(8): 893-900.

[76] Fisher RA (1924) On a Distribution Yielding the Error Functions of Several Well Known Statistics. Proc. of the Int. Congress of Mathematics. Toronto, Canada; 2: 805–813.
[77] Pfurtscheller G, Flotzinger D, Kalcher J. Brain-Computer Interface—a new communication device for handicapped persons. J of Microcomput Appl. 1993; 16 (3): 293-299.
[78] Bhatia R. The Riemannian Mean of Positive Matrices. Ch 2 in Nielsen F. and Bhatia R. (Eds.) Matrix Information Geometry. London: Springer; 2013.
[79] Moakher M. A differential geometric approach to the arithmetic and geometric means of operators in some symmetric spaces. SIAM J Matrix Anal Appl. 2005; 26(3): 735-747.
[80] Moakher M, Batchelor PG. Symmetric positive-definite matrices: From geometry to applications and visualization. In: Visualization and Processing of Tensor Fields (Weickert J, Hagen H Eds.). London: Springer, 2006: 285-298.
[81] Pennec X, Fillard P, Ayache N. A Riemannian Framework for Tensor Computing. Research Report #5255, INRIA, Sophie-Antipolis, France; 2004.
[82] Nakamura N. Geometric means of Positive Operators, KYUGPOOK Math J, 2009: 167-181.
[83] Sra S. A new Metric on the manifold of kernel matrices with application to matrix geometric means, NIPS Conference. 2012 Dec 3-8; Harrah's Lake Tahoe (CA), USA: 1-9.
[84] Arsigny V, Fillard P, Pennec X, Ayache N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J Matrix Anal Appl. 2007; 29(1): 328–347.
[85] Anderson TW. An Introduction to Multivariate Statistical Analysis, 2nd ed.. New York (NY), USA: John Wiley & Sons; 1984.
[86] Wilks SS. Multidimensional Statistical Scatter. In Collected Papers: Contributions to Mathematical Statistics (Anderson TW Eds). New York (NY), USA: John Wiley & Sons; 1967: 597–614.
[87] Manton JH. A globally convergent numerical algorithm for computing the centre of mass on compact Lie groups. ICARCV Conference proc, 2004 Dec 6-9: 2211-2216.
[88] Congedo M, Barachant A, Kharati Koopaei E. Fixed Point Algorithms for Estimating Power Means of Positive Definite Matrices, IEEE Trans Signal Process. 2017; 65(9): 2211–2220.
[89] Comon P, Jutten C. Handbook of Blind Source Separation: Independent Component Analysis and Applications, Oxford (UK): Academic Press; 2010.
[90] Congedo M, Gouy-Pailler C, Jutten C . On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. Clin Neurophysiol. 2008; 119: 2677-2686.
[91] Congedo M, Afsari B, Barachant A. Moakher M. Approximate Joint Diagonalization and Geometric Mean of Symmetric Positive Definite Matrices. PLoS ONE. 2005; 10(4): e0121423.
[92] Lotte F, Guan CT. Regularizing Common Spatial Patterns to Improve BCI Designs: Unified Theory and New Algorithms, IEEE Trans Biomed Eng. 2011; 58(2): 355-362.
[93] Fukunaga F, Koontz W. Applications of the Karhunen-Loève expansion to feature selection and ordering. IEEE Trans Comput. 1970; 19(5): 311-318.
[94] Koles ZJ. The Quantitative extraction and Topographic Mapping of the Abnormal Components in the Clinical EEG. Electroencephalogr Clin Neurophysiol. 1991; 79: 440-447.
[95] Koles ZJ, Soong A. EEG Source Localization: Implementing the Spatio-Temporal Decomposition Approach. Electroencephalogr Clin Neurophysiol. 1998; 107: 343-352.
[96] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal Spatial Filtering of single trial EEG during Imagined Hand Movement. IEEE Trans Rehabil Eng. 2000; 8(4): 441-446.
[97] Huo X. A statistical Analysis of Fukunaga-Koontz Transform, IEEE Signal process Lett. 2004; 11(2): 123-126.
[98] Barachant A, Bonnet S, Congedo M, Jutten C. Riemannian Geometry Applied to BCI Classification. Proc. of Latent Variable Analysis and Signal Separation Conference, 2010 Sept 27-30. St. Malo, France; 6365: 629-636
[99] Yger F., Lotte F., Sugiyama M. Averaging covariance matrices for EEG signal classification based on the CSP: An empirical study. Int. EUSIPCO Conf., 2015 Aug 31-Sept 4, Nice, France: 2721-2725.
[100] Fukunaga K. Statistical Pattern Recognition (2nd Eds). New York (NY), USA: Academic Press; 1990.
[101] Schott JR. Matrix Analysis for Statistics. New York (NY), USA: John Wiley & Sons; 1997.
[102] Barachant A, Bonnet S, Congedo M, Jutten C. Multi-Class Brain Computer Interface Classification by Riemannian Geometry, IEEE Trans Biomed Eng. 2012; 59(4): 920-928.
[103] Xie X, Yu ZL, Lu H, Gu Z, Li Y. Motor Imagery Classification based on Bilinear Sub-Manifold Learning of Symmetric Positive-Definite Matrices. IEEE Trans Neural Syst Rehabil Eng. 2016; Epub ahead of print.
[104] Li Y, Wong KM, de Bruin H, EEG Signal Classification Based on a Riemannian Distance Measure. IEEE Int. Conf. on Science and Technology for Humanity. 2009 Sep 26-27. Toronto, Canada: 268-273.

[105] Li Y, Wong KM, De Bruin H. EEG signals classification for sleep-state decision – A Riemannian geometry approach, IET Signal Process. 2012; 6(4): 288–299.
[106] Li Y, Wong KM, Signal classification by power spectral density: An approach via Riemannian geometry, Proc. IEEE Stat. Signal Processing, 2012 Aug 5-8. Ann Arbor (MI), USA
[107] Li Y., Wong K.M. de Bruin H. Electroencephalogram signals classification for sleep state decision - A Riemannian geometry approach, IET signal process. 2011; 6(4): 288-299.
[108] Barachant A, Bonnet S, Congedo M, Jutten C. Common Spatial Pattern revisited by Riemannian Geometry. IEEE Int. Workshop on Multimedia Signal Processing, 2010 Oct 4-6; St-Malo, France: 472-476.
[109] Barachant A, Bonnet S, Congedo M, Jutten C. A Brain-Switch using Riemannian Geometry. 5th Int. Brain-Computer Interface Conf, 2011 Sept 22-24; Graz, Austria: 64-67.
[110] Barachant A, Bonnet S, Congedo M, Jutten J. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. Neurocomput. 2013; 112: 172-178.
[111] Barachant A, Congedo M, Van Veen G, Jutten C. Classification de potentiels évoqués P300 par géométrie riemannienne, GRETSI Proc., 2013 Sept 3-6, St. Malo, France.
[112] Brandl S, Müller KR, Samek W. Robust common spatial patterns based on Bhattacharyya distance and Gamma divergence, Proc. of the Int. Winter Workshop on Brain-Computer Interface, 2015 Jan 12-14; Jeongsun-Kun, South Korea: 1-4.
[113] Samek W, Kawanabe M, Müller K-R. Divergence-based Framework for Common Spatial Patterns Algorithms, IEEE Rev Biomed Eng. 2014; 7: 50-72.
[114] Samek W, Müller KR. Information geometry meets BCI. Proc. of the International Winter Workshop on Brain-Computer Interface, 2014 Feb 17-19; Jeongsun-Kun, South Korea: 1-4.
[115] Samek W, Kawanabe M. Robust common spatial patterns by minimum divergence covariance estimator. Proc. of the Acoustics, Speech and Signal Processing Conference, 2014 May 4-9; Firenze, Italy.
[116] Kalunga E, Chevallier S, Barthélemy Q. Data augmentation in Riemannian space for Brain-Computer Interfaces, ICML Workshop on Statistics, Machine Learning and Neuroscience, 2015 July 6-11; Lille, France.
[117] Lindig-León C, Gayraud N, Bougrain L, Clerc M. Comparison of Hierarchical and Non-Hierarchical Classification for Motor Imagery Based BCI Systems. Proc. of the 6th Int. Brain-Computer Interface Meeting, 2016 May 30-June 3; Asilomar (CA), USA.
[118] Waytowich N, Lawhern V, Bohannon A, Lance B. Efficient Transfer Learning in Brain Computer Interfaces using Spectral Meta Learning. Proc. of the 6th Int. Brain-Computer Interface Meeting, 2016 May 30-June 3; Asilomar (CA), USA.
[119] Nasiri Ghosheh Bolagh S, Shamsollahi MB, Congedo M, Jutten C. Rank Of Subjects and Riemannian Geometry for Brain Signal Decoding Across Subjects, Proc. of the ESANN Conf., 2016 Apr 27-29, Bruges, Belgium.
[120] Waytowich N, Lawhern V, Bohannon A, Ball KR, Lance B. Spectral Transfer Learning Using Information Geometry for a User-Independent Brain-Computer Interface, Front Neurosci. 2016; 10: 430.
[121] Barachant A, Carmel JB, Friel KM, Gupta D. Extraction of motor patterns from joint EEG/EMG recording: A Riemannian Geometry approach. Proc. of the 6th Int. Brain-Computer Interface Meeting, 2016 May 30 June 3, Asilomar (CA), USA.
[122] Llera A, Gómez V, Kappen HJ. Adaptive classification on brain-computer interfaces using reinforcement signals. Neural Comput. 2012; 26: 1108-1127.
[123] Uehara T, Tanaka T, Fiori F, Robust averaging of covariance matrices by Riemannian geometry for motor-imagery brain–computer interfacing, Proc. fifth Int. Conf. on Cognitive Neurodynamics. 2015 Jun 3-7, Sanya, China: 347-353.
[124] Yger F. A review of kernels on covariance matrices for BCI applications, IEEE Int. Workshop on Machine Learning for Signal Processing, 2013 Sep 22-25; Southampton, UK.
[125] Yger F, Sugiyama M. Supervised LogEuclidean Metric Learning for Symmetric Positive Definite Matrices, arXiv:1502.03505; 2015.
[126] Davoudi A, Ghidary SS, Sadatnejad K. Dimensionality reduction based on Distance Preservation to Local Mean (DPLM) for SPD matrices and its application in BCI. arXiv:1608.00514; 2016.
[127] Horev I, Yger F, Sugiyama M. Geometry-Aware Principal Component Analysis for Symmetric Positive Definite Matrices. JMLR: Workshop and Conf. Proc. 2015; 45: 1–16.
[128] Horev I, Yger F, Sugiyama M. Geometry-aware Stationary Subspace Analysis. arXiv: 1605.07785v1; 2016.
[129] Krivov E, Belyaev M. Dimensionality reduction with isomap algorithm for EEG covariance matrices. 4th Int. Winter Conf. on Brain-Computer Interface; 2016 Feb 22-24, Yongpyong, South Korea.

[130] Sadatnejad K, Ghidary SS. Kernel learning over the manifold of symmetric positive definite matrices for dimensionality reduction in a BCI application, Neurocomput. 2016; 179: 152-160.
[131] Tanaka T, Uehara T, Tanaka Y. Dimensionality reduction of sample covariance matrices by graph Fourier transform for motor imagery brain-machine interface, IEEE Statistical Signal Processing Workshop, 2016 Jun 26-29, Palma de Mallorca, Spain: 1-5.
[132] Harandi MT, Salzmann M, Hartley R. From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices, European Conference on Computer Vision, 2014 Sep 6-12; Zurich, Switzerland: 17-32.
[133] Harandi MT, Salzmann M, Hartley R. Dimensionality Reduction on SPD Manifolds: The Emergence of Geometry-Aware Methods, arXiv:1605.06182; 2017.
[134] Alimardani F, Boostani R, B Blankertz. Presenting a Spatial-Geometric EEG Feature to Classify BMD and Schizophrenic Patients, Int J Adv Telecommun Electrotechnics Signals Syst. 2016; 5(2): 79-85.
[135] Yuan S, Zhou W, Wu Q, Zhang Y. Epileptic Seizure Detection with Log-Euclidean Gaussian Kernel-Based Sparse Representation. Int J Neural Syst. 2016; 26(3): 1650011.
[136] Roy RN, Charbonnier S, Bonnet S. Detection of mental fatigue using an active BCI inspired signal processing chain. Proc. of the 19th World Congress of the Int. Federation of Automatic Control, 2014 Aug 24-29, Cape Town, South Africa.
[137] Navarro-Sune X, Hudson AL, De Vico Fallani F, Martinerie J, Witon A, Pouget P, *et al*. Riemannian geometry applied to detection of respiratory states from EEG signals: the basis for a brain-ventilator interface, arXiv:1601.03022; 2016.
[138] Rutkowski TM. Data-Driven Multimodal Sleep Apnea Events Detection: Synchrosquezing Transform Processing and Riemannian Geometry Classification Approaches. J Med Syst, 2016; 40(7): 162.
[139] Congedo M, Barachant A. A Special Form of SPD Covariance Matrix for Interpretation and Visualization of Data Manipulated with Riemannian Geometry, Proc. MaxEnt Conference, 2014 Sep 21-26, Amboise, France: 495.
[140] Lu J, Xie K, McFarland DJ. Adaptive spatio-temporal filtering for movement related potentials in EEG-based brain-computer interfaces. IEEE Trans Neural Syst Rehabil Eng. 2014; 22(4): 847-57 .
[141] Treder M S, Porbadnigk A K, Shahbazi Avarvand F, Müller K R, Blankertz B. The LDA beamformer: Optimal estimation of ERP source time series using linear discriminant analysis. Neuroimage. 2016; 129: 279-291.
[142] Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles, Clin Neurophysiol. 1999; 110(11): 1842-57.
[143] Congedo M, Korczowski L, Delorme A, Lopes Da Silva F. Spatio-Temporal Common Pattern; a Companion Method for ERP Analysis in the Time Domain. J neurosci methods. 2016; 267: 74–88.
[144] Bloomfield P. Fourier Analysis of Time Series. New York: John Wiley & Sons; 2000.