Geodesically-convex optimization for averaging partially observed covariance matrices

Florian Yger, Sylvain Chevallier, Quentin Barthélemy, Suvrit Sra

Proceedings of Machine Learning Research 129:417–432, 2020 ACML 2020

Geodesically-convex optimization for averaging partially observed covariance matrices

Florian Yger* FLORIAN.YGER@DAUPHINE.FR
LAMSADE, CNRS, Université Paris-Dauphine, Université PSL
Paris, France

Sylvain Chevallier* SYLVAIN.CHEVALLIER@UVSQ.FR
Université Paris-Saclay, UVSQ, LISV
Vélizy-Villacoublay, France

Quentin Barthélemy Q.BARTHELEMY@FOXSTREAM.FR
Foxstream
Vaulx-en-Velin, France

Suvrit Sra SUVRIT@MIT.EDU
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology
Cambridge, MA, USA

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Symmetric positive definite (SPD) matrices permeates numerous scientific disciplines, including machine learning, optimization, and signal processing. Equipped with a Riemannian geometry, the space of SPD matrices benefits from compelling properties and its derived Riemannian mean is now the gold standard in some applications, *e.g.* brain-computer interfaces (BCI). This paper addresses the problem of averaging covariance matrices with missing variables. This situation often occurs with inexpensive or unreliable sensors, or when artifact-suppression techniques remove corrupted sensors leading to rank deficient matrices, hindering the use of the Riemannian geometry in covariance-based approaches. An alternate but questionable method consists in removing the matrices with missing variables, thus reducing the training set size. We address those limitations and propose a new formulation grounded in geodesic convexity. Our approach is evaluated on generated datasets with a controlled number of missing variables and a known baseline, demonstrating the robustness of the proposed estimator. The practical interest of this approach is assessed on real BCI datasets. Our results show that the proposed average is more robust and better suited for classification than classical data imputation methods.

Keywords: SPD matrices, average, missing data, data imputation.

1. Introduction

With from the pioneering work of Candès and Recht (2009), a new light was shed on the matrix completion problem and the imputation of missing data. It demonstrated that data imputation could be formulated as an optimization problem benefiting from the underlying structure of the data. Such method has been successfully applied to recommendation

* Equal contributions

systems (Rennie and Srebro, 2005), multitask learning (Obozinski et al., 2010) or manifold learning (Weinberger and Saul, 2006). The bound of the number of entries for an optimal reconstruction of a low-rank matrix are provided in Candès and Recht (2009) and improved in Recht (2011), with only minimal assumptions on the coherence of the matrix to recover.

Several problems connected to low-rank matrix completion have been considered, including recovering spectrally sparse objects from missing time-domain samples (Chen and Chi, 2013), or in graph data (Narang et al., 2013). Positive definite and positive semi-definite matrices are indeed good candidates for matrix completion approaches as those matrices lives in a low-dimensional space, and have been the subject of early works in matrix completion (Johnson, 1990; Johnson and Tarazaga, 1995). Symmetric positive-definite (SPD) matrices are commonly found in covariance-based approaches, such as financial time series processing (Bingham, 2014), radar detection (Arnaudon et al., 2013), computer vision (Harandi et al., 2017) or medical imaging (Pennec et al., 2006).

In this contribution, we focus on yet another application of positive-definite matrices: the brain-computer interfaces (BCI). Since the seminal work of Barachant et al. (2012) that introduced a new approach for BCI, the covariance matrices are directly handled on the manifold of positive-definite matrices. This implied to reformulate common machine learning algorithms and processing pipelines for this specific geometry. These approaches became the new gold standard for BCI (Yger et al., 2017; Congedo et al., 2017) with high accuracy on various kinds of brain signals, such as motor imagery (Jayaram and Barachant, 2018), evoked potentials (Korczowski et al., 2015) or steady-state visually-evoked potential (Chevallier et al., 2020), and won several Kaggle competitions.

In all these applications, from BCI to financial time series, machine learning models rely on average centers of SPD matrices. The estimation of average (or center of gravity) is the key to obtain reliable results for the target application, and must be robust to missing variables. However, in BCI for example, unreliable sensors or artifact-suppression techniques removing corrupted sensors leading to rank deficient matrices, hindering the use of the usual Riemannian averaging. To our knowledge, there is no systematic review of the processing of SPD matrices with missing data in the literature. This is the main objective of this paper, by confronting the existing approaches and by introducing a new framework.

Our contributions are the following:

i) a sound framework for SPD matrices averaging with missing variables, and a theoretical analysis of the geodesic convexity of this optimization,

ii) an evaluation of the convergence of the proposed approach on a synthetic dataset,

iii) a comparison with Euclidean mean and common imputation methods in machine learning,

iv) a validation on a real dataset for BCI application.

2. Review of related works

There are several types of missing data that could be encompassed in the literature. In an attempt to make a summary of the studies on the subject, we could describe the different cases that are illustrated on Fig. 1. Our approach comes from practical concerns occurring

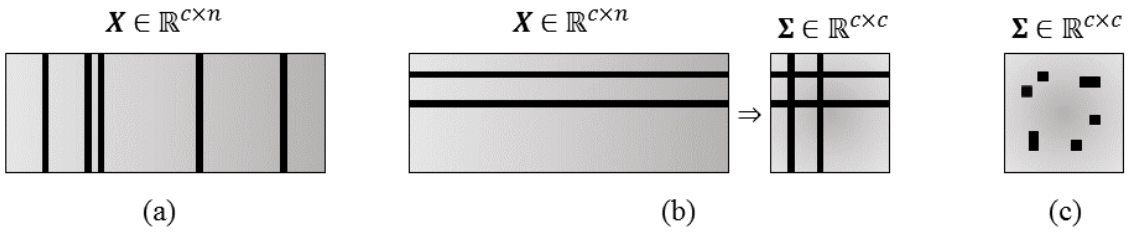

Figure 1: Illustration of the different types of missing data: (a) missing samples / observations in matrix X, (b) missing variables in matrices X and Σ, (c) missing elements in the matrix Σ.

when considering multivariate data. We call $X \in \mathbb{R}^{c \times n}$ a matrix composed of c input variables (sensors, or components, shown in rows) and of n samples (observations, visible in columns). We call $\Sigma \in \mathbb{R}^{c \times c}$ its covariance matrix, which is SPD[1]. As discussed hereafter, the data can be corrupted and missing data can occur with particular patterns.

The first case one occurs when some multivariate samples have not been acquired by the system, resulting in missing columns in the matrix X. Missing values in this case is often interpolated or imputed from average or median values. In the second case, some sensors have a trouble during acquisition. Consequently, some variables/rows are missing in X, and these rows are thus missing in Σ, but also their corresponding columns due to the symmetry.

In Fukuda et al. (2001), the authors rely on matrix completion techniques to propose a solver in semi-definite programming that is able to cope with partial information in the constraint space of the primal-dual interior-point methods. Tsuda et al. (2003) propose to complete missing variables for SVM kernel estimation in genetic applications, relying on an auxiliary kernel matrix that integrates another source of information.

Some works consider the case where some elements of the matrix are missing (Bishop and Byron, 2014; Choi et al., 2019), but without a particular link between these elements, excepted symmetry. The work of Johnson and Tarazaga (1995) is on matrix completion for graph data with positive semi-definite constraint, for missing entries outside the diagonal. Another approach by Laurent (2001), using the same setup, try to estimate missing information on the diagonal.

Several problems can be stated when processing incomplete matrices, *i.e.* matrices with missing data. The first one concerns the estimation of covariance from an incomplete matrix X, with the objective to build a robust estimator (Little, 1998; Schneider, 2001; Lounici, 2014). The second category of problems is the covariance completion. From an incomplete covariance matrix Σ, the goal is to find the best way to complete its missing values (Johnson and Tarazaga, 1995; Bishop and Byron, 2014).

A two-step approach is presented in (Rodrigues et al., 2019): firstly, covariance matrices are imputed on the manifold (performing on the matrix a whitening, a scaling, completing it

1. In theory, covariance matrices belong to the set of symmetric semi-definite matrices but making the realistic assumption that X is corrupted by a Gaussian noise leads to $\Sigma \in \mathcal{P}_c$.

	Missing samples in X	Missing variables in X and Σ	Missing elements in Σ
Covariance estimation	Little (1998) Schneider (2001) Lounici (2014)	Vinci et al. (2019)	$\times$
Covariance completion	$\times$	Rodrigues et al. (2019)	Johnson and Tarazaga (1995) Fukuda et al. (2001) Laurent (2001) Tsuda et al. (2003) Bishop and Byron (2014)
Covariance averaging	$\times$	Our work	Choi et al. (2019) for $m = 2$

Table 1: Summary of the state-of-the-art classified along two criteria: the type of missing data (horizontally) and the type of problems (vertically). Symbols $\times$ means that this problem cannot be considered.

by adding 1s on the diagonal where variables are missing, and then de-whitening), secondly a classical covariance averaging is applied on completed matrices. From several incomplete covariance matrices, the goal is to estimate the best average covariance matrices. In the work of Choi et al. (2019), the mean is limited to 2 covariance matrices, having identical missing elements. To sum up this state-of-the-art, all these works are classified in Table 1 along two criteria: the type of missing data (horizontal) and the type of problems (vertical).

The contribution of this paper belongs to covariance averaging, providing a way to average several SPD matrices with missing variables. It is important to note that missing variables can change for each matrix to average. The resulting matrix is defined on the union (and not the intersection) of variables of input matrices.

3. Estimating Riemannian mean with missing variables

After a brief recall of the Riemannian geometry to process SPD matrices, this section presents the problem of estimating an average matrix from a set of SPD matrices with missing variables, introducing masks to tackle missing variables. The cost function is shown to be geodesically convex under some restricted cases.

3.1. Preliminaries

We consider the set $\mathcal{P}_c$ of symmetric positive definite (SPD) of size $c \times c$. Formally, this set is defined as:

$$\mathcal{P}_c = \left\{ \Sigma \in \mathbb{R}^{c \times c} \mid \Sigma = \Sigma^\top,\ \forall x \in \mathbb{R}^c_\star \quad x^\top \Sigma x > 0 \right\} , \tag{1}$$

and it is endowed with the Löwner partial order, defined between $\Sigma_1, \Sigma_2 \in \mathcal{P}_c$ as:

$$\Sigma_1 \succcurlyeq \Sigma_2 \quad \Leftrightarrow \quad \Sigma_1 - \Sigma_2 \in \mathcal{P}_c \ .$$

The most straightforward way to handle $\mathcal{P}_c$ is to consider it as a subspace of $\mathcal{S}_c$, the space of symmetric matrices. As such, it is a Euclidean space and the tools of conic geometry and semi-definite programming could apply. However, this Euclidean geometry comes at the cost of dramatic drawbacks, like the swelling effect[2] (Fletcher et al., 2004; Horev et al., 2015). Non-Euclidean alternatives have been successfully applied in some applications (Yger et al., 2017; Congedo et al., 2017; Harandi et al., 2017).

The most common estimator of the covariance matrix Σ is the sample covariance matrix, defined as:

$$\Sigma = \frac{1}{n-1} X X^\top \ . \tag{2}$$

More robust estimators have been proposed (Daniels and Kass, 2001) and have been evaluated in the context of EEG-based BCI (Chevallier et al., 2018). The most common are the shrinkage estimator that combines the sample covariance matrix with a target covariance matrix, that could be similar to the identity matrix, to ensure that the estimated matrices avoid any ill-conditioning problem.

Due to the curvature of the space of SPD matrices, an adequate metric could rely on $\Sigma_1 \sharp_t \Sigma_2$, the geodesic between $\Sigma_1, \Sigma_2 \in \mathcal{P}_c$ and defined as:

$$\Sigma_1 \sharp_t \Sigma_2 = \Sigma_1^{\frac{1}{2}} \left(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}} \right)^t \Sigma_1^{\frac{1}{2}} = \Sigma_1^{\frac{1}{2}} \operatorname{Exp}\left(t \operatorname{Log}\left(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}} \right) \right) \Sigma_1^{\frac{1}{2}} \ , \ t \in [0,1] \ . \tag{3}$$

For clarity reason, when t is omitted $\Sigma_1 \sharp \Sigma_2 = \Sigma_1 \sharp_{0.5} \Sigma_2$, the geometric center of Σ_1 and Σ_2. We will consider hereafter the affine-invariant Riemannian (AIR) distance, that derived from the geodesic and that is widely used for its geometric properties.

Definition 1 *Affine-invariant Riemannian distance (Förstner and Moonen, 1999). For $\Sigma_1, \Sigma_2 \in \mathcal{P}_c$, two SPD matrices, the Riemannian distance is defined as*

$$\delta_R(\Sigma_1, \Sigma_2) = \left\| \operatorname{Log}(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}) \right\|_F = \left(\sum_{k=1}^{c} \log^2 \lambda_k \right)^{\frac{1}{2}} \ , \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm, and λ_k, $k = 1, \dots, c$, are the eigenvalues of $\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}$.

Among the numerous properties of the affine-invariant Riemannian distance (Bhatia, 2009), the congruence-invariance[3] states that, for any invertible matrix $W \in \mathbb{R}^{c \times c}$:

$$\delta_R(\Sigma_1, \Sigma_2) = \delta_R(W^\top \Sigma_1 W, W^\top \Sigma_2 W) \ . \tag{5}$$

We now want to estimate the mean of a set of m SPD matrices $\{\Sigma_i\}_{i=1}^m$. It may actually be helpful to estimate a weighted mean, where the weights $\{w_i\}_{i=1}^m$ are non-negative and sum up to a constant (in practice, 1 or m). These weights can be fixed to encode prior knowledge, or tuned by cross-validation to optimize some given criterion.

2. In the Euclidean geometry of definite positive matrices, the determinant of $\frac{\Sigma_1 + \Sigma_2}{2}$ can be greater than the determinants of Σ_1 or Σ_2. Hence, the Euclidean interpolation can bring some spurious information, contrary to the Riemannian geometry.
3. Using this transformation as a whitening, this invariance could be used in a transfer learning setup as in Yger and Sugiyama (2015).

Definition 2 *Riemannian mean (Moakher, 2005; Fletcher et al., 2004). The Riemannian mean* $\bar{\Sigma} \in \mathcal{P}_c$ *of a set of* m *SPD matrices* $\{\Sigma_i\}_{i=1}^m$*, associated to non-negative weights* $\{w_i\}_{i=1}^m$*, is defined as*

$$\bar{\Sigma} = \arg\min_{\Sigma \in \mathcal{P}_c} f_R(\Sigma) = \arg\min_{\Sigma \in \mathcal{P}_c} \frac{1}{2} \sum_{i=1}^{m} w_i \, \delta_R^2(\Sigma_i, \Sigma) \ . \tag{6}$$

This mean, also known as geometric mean, has no closed form and therefore has to be computed iteratively, for example through a gradient descent algorithm. The gradient of the cost function f_R is defined as (Fletcher et al., 2004):

$$\nabla f_R(\Sigma) = -\sum_{i=1}^{m} w_i \, \Sigma^{\frac{1}{2}} \operatorname{Log}(\Sigma^{-\frac{1}{2}} \Sigma_i \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}} = -\sum_{i=1}^{m} w_i \, \operatorname{Log}_{\Sigma}(\Sigma_i) \ , \tag{7}$$

where $\operatorname{Log}_{\Sigma}(\Sigma_i)$ is the logarithmic map, projecting the matrix Σ_i from manifold to tangent space at point Σ. The geometric mean is the unique matrix $\bar{\Sigma}$ of the manifold $\mathcal{P}_c$ nullifying the sum of the m tangent vectors.

As defined in Zadeh et al. (2016) and explained in Boumal (2020, Chap 11), the geometric mean is a geodesically convex function. Even if it is not convex in the Euclidean sense, geodesic convexity ensures that functions are convex along the manifold and enjoying similar properties to convex functions.

Definition 3 *Geodesically convex function (Sra and Hosseini, 2015). Let* $\mathcal{P}$ *be a Riemannian manifold and* $\mathcal{Q} \subset \mathcal{P}$ *a geodesically convex set. A function* $f : \mathcal{Q} \to \mathbb{R}$ *is said geodesically convex, if for all points* Σ_1 *and* Σ_2 *of this set, it satisfies*

$$f(\Sigma_1 \sharp_t \Sigma_2) \ \leq \ t f(\Sigma_1) + (1-t) f(\Sigma_2) \ , \ t \in [0,1] \ . \tag{8}$$

3.2. Masked Riemannian mean

A mask $M \in \mathbb{R}^{c \times (c-p)}$ is defined as the identity matrix, but with the columns of the p missing variables removed, with $0 \leq p < c$. In our problem, the mask is not constrained to be identical for all input matrices, *ie.* each input matrix can be associated to a particular mask. If an input $X \in \mathbb{R}^{c \times n}$ is incomplete, with p missing variables, the available (non-missing) variables can be extracted thanks to the product by a mask $M \in \mathbb{R}^{c \times (c-p)}$. The resulting complete submatrix is $\check{X} = M^\top X \in \mathbb{R}^{(c-p) \times n}$. Thus, its covariance matrix $\check{\Sigma} \in \mathcal{P}_{c-p}$ estimated with Eq. (2) is:

$$\check{\Sigma} = \frac{1}{n-1} \check{X} \check{X}^\top = \frac{1}{n-1} M^\top X X^\top M = \frac{1}{n-1} M^\top \Sigma M \ , \tag{9}$$

where $\Sigma = X X^\top$ is an incomplete covariance matrix. Incomplete matrices X and Σ are illustrated in Fig. 1(b). Consequently, a mask is a full column-rank matrix such that the submatrix $M^\top \Sigma M$ is SPD, of size $(c-p) \times (c-p)$[4].

4. Hence, mask M is a semi-orthogonal matrix, *i.e.* $M^\top M = \mathbb{I}_{(c-p)}$, and our setup could be generalized to missing subspaces in a *compressive sensing* manner using orthogonal matrices.

Now, we consider the set $\{\Sigma_i\}_{i=1}^m$ of incomplete SPD matrices, *ie.* each matrix Σ_i has p_i missing variables, with $0 \leq p_i < c$. Consequently, their associated masks $\{M_i\}_{i=1}^m$ are potentially all different. The masked Riemannian mean $\bar{\Sigma} \in \mathcal{P}_c$ is defined as:

$$\bar{\Sigma} = \arg\min_{\Sigma \in \mathcal{P}_c} f_{R_M}(\Sigma) = \arg\min_{\Sigma \in \mathcal{P}_c} \frac{1}{2} \sum_{i=1}^{m} w_i \, \delta_R^2(M_i^\top \Sigma_i M_i, M_i^\top \Sigma M_i) \ . \tag{10}$$

Note that, each matrix $M_i^\top \Sigma_i M_i$ being SPD, the distance δ_R can be applied and the cost function is well defined. As explained in Absil et al. (2009) and Boumal (2020, Chap 4), the Riemannian gradient can be deduced from the Euclidean gradient and then plugged into a Riemannian gradient descent. Combining the chain rule of derivatives with the gradient formula given in Eq. (7), the Euclidean gradient of the cost function $f_{R_M}(\Sigma)$ can be written as:

$$\nabla f_{R_M}(\Sigma) = -\sum_{i=1}^{m} w_i \, M_i \operatorname{Log}_{M_i^\top \Sigma M_i} \left(M_i^\top \Sigma_i M_i \right) M_i^\top \ . \tag{11}$$

The cost being computed on subparts of the matrix Σ, the gradient is null for the elements not involved on the compressed matrix $M_i^\top \Sigma M_i$. Hence, the Euclidean gradient of δ_R^2, although computed on a subpart of Σ can be broadcast to the whole space using $M_i(\cdot)M_i^\top$ operator. Remark that Eq. (11) falls back to Eq. (7) when there is no missing variables, *i.e.* when all masks $M_i = \mathbb{I}_c$.

The illustration of the averaging of two masked matrices is displayed in Fig. 2.

3.3. Link with the NaN-mean

If we replace the Riemannian distance δ_R by the Euclidean distance $\delta_E(\Sigma_1, \Sigma_2) = \|\Sigma_1 - \Sigma_2\|_F$ into Eq. (10), the gradient of the masked Euclidean mean would have been:

$$\nabla f_{E_M}(\Sigma) = -\sum_{i=1}^{m} w_i \, M_i \left(M_i^\top \Sigma_i M_i - M_i^\top \Sigma M_i \right) M_i^\top = -\sum_{i=1}^{m} w_i \, M_i M_i^\top \left(\Sigma_i - \Sigma \right) M_i M_i^\top \ , \tag{12}$$

where the operator $M_i M_i^\top (\cdot) M_i M_i^\top$ extracts only the non-missing values of Σ_i.

Defining $\mathcal{I}_{u,v}$ as the set containing the indices of matrices without missing values for entry (u, v), the gradient can be expressed element-wise:

$$\nabla f_{E_M}(\Sigma)(u, v) = -\sum_{i \in \mathcal{I}_{u,v}} w_i \ (\Sigma_i(u, v) - \Sigma(u, v)) \ . \tag{13}$$

This gradient is nullified by the matrix:

$$\bar{\Sigma}(u, v) = \frac{\sum_{i \in \mathcal{I}_{u,v}} w_i \, \Sigma_i(u, v)}{\sum_{i \in \mathcal{I}_{u,v}} w_i} \ . \tag{14}$$

When all weights are equal, this masked Euclidean mean is commonly called *NaN-mean* and is written as:

$$\bar{\Sigma}(u, v) = \frac{1}{|\mathcal{I}_{u,v}|} \sum_{i \in \mathcal{I}_{u,v}} \Sigma_i(u, v) \ , \tag{15}$$

where $|\mathcal{I}_{u,v}|$ denotes the cardinality of $\mathcal{I}_{u,v}$.

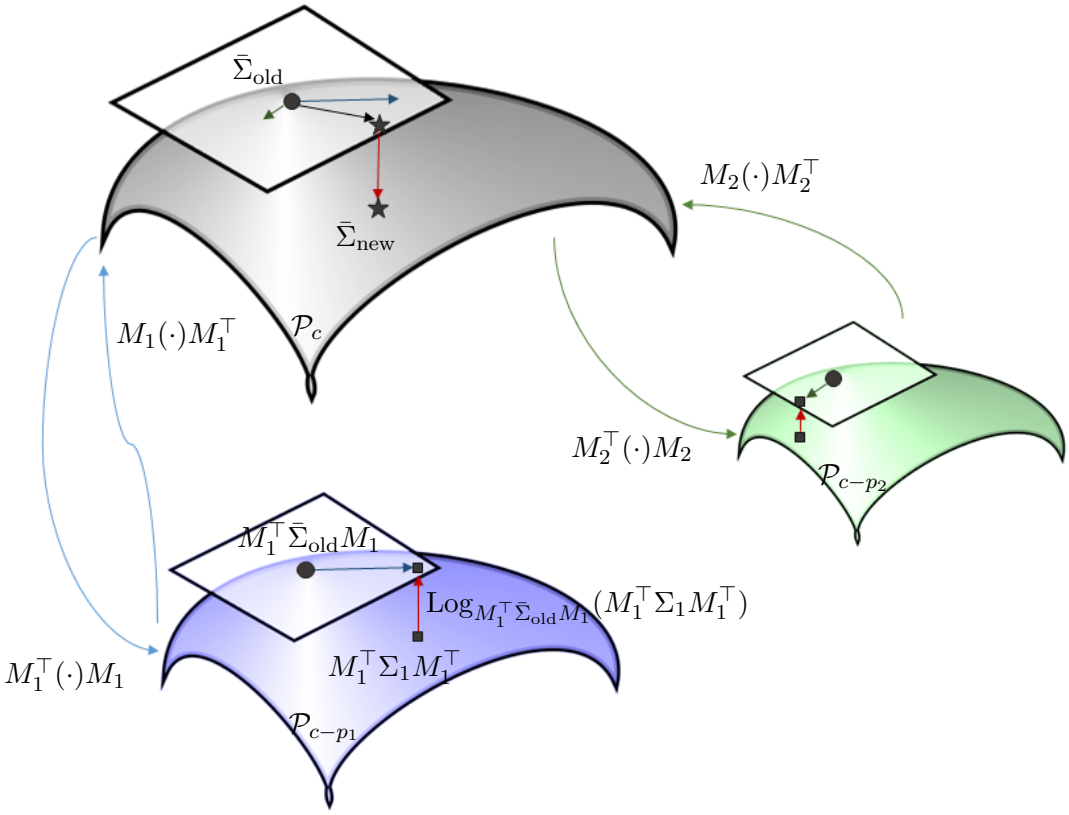

Figure 2: Illustration of the masked Riemannian mean $\bar{\Sigma} \in \mathcal{P}_c$ of two matrices, Σ_1 (*resp.* Σ_2) containing p_1 (*resp.* p_2) missing variables encoded by mask M_1 (*resp.* M_2).

3.4. Geodesic convexity

In practice, we noticed that a simple Riemannian gradient descent always converged to the same minimum (independently of its initialization) on our experiments and we tested it numerically in Pymanopt (Townsend et al., 2016). This lead us to formulate a conjecture on the geodesic convexity of the problem. Unfortunately, we could generate some counter-examples to this conjecture in simplistic cases, but we also found some restricted cases where the geodesic convexity holds.

Applying the congruence-invariance of Eq. (5) with $W = (M_i^\top\Sigma_iM_i)^{-\frac{1}{2}} \in \mathbb{R}^{(c-p_i)\times(c-p_i)}$, we can express the cost function f_{R_M}, defined in Eq. (10), as follows:

$$\begin{aligned} f_{R_M}(\Sigma) &= \frac{1}{2}\sum_{i=1}^{m} w_i\,\delta_R^2\left(\mathbb{I}_{(c-p_i)}, ((M_i^\top\Sigma_iM_i)^{-\frac{1}{2}})^\top M_i^\top\Sigma M_i(M_i^\top\Sigma_iM_i)^{-\frac{1}{2}}\right) \\ &= \frac{1}{2}\sum_{i=1}^{m} w_i\,\delta_R^2\left(\mathbb{I}_{(c-p_i)}, \tilde{M}_i^\top\Sigma\tilde{M}_i\right)\ , \end{aligned} \tag{16}$$

where $\tilde{M}_i = M_i(M_i^\top\Sigma_iM_i)^{-\frac{1}{2}} \in \mathbb{R}^{c\times(c-p_i)}$ is a full column-rank matrix.

In the case where we have $\tilde{M}^\top \tilde{M} \succcurlyeq \mathbb{I}_{(c-p)}$, then, for any $\Sigma_1, \Sigma_2 \succcurlyeq \mathbb{I}_c$:

$$\mathbb{I}_{(c-p)} \preccurlyeq \tilde{M}^\top \tilde{M} \preccurlyeq \tilde{M}^\top (\Sigma_1 \sharp \Sigma_2) \tilde{M} \preccurlyeq (\tilde{M}^\top \Sigma_1 \tilde{M}) \sharp (\tilde{M}^\top \Sigma_2 \tilde{M}) \ , \tag{17}$$

and because $\|\mathrm{Log}(\cdot)\|_F^2 = \delta_R^2(\mathbb{I}, \cdot)$ is monotonically increasing above $\mathbb{I}$, we have:

$$\delta_R^2 \left(\mathbb{I}_{(c-p)}, \tilde{M}^\top (\Sigma_1 \sharp \Sigma_2) \tilde{M} \right) \leq \delta_R^2 \left(\mathbb{I}_{(c-p)}, (\tilde{M}^\top \Sigma_1 \tilde{M}) \sharp (\tilde{M}^\top \Sigma_2 \tilde{M}) \right) \ . \tag{18}$$

From this, g-convexity follows upon using g-convexity of δ_R (Sra and Hosseini, 2015) in the usual way for the very last occurrence of δ_R. This proof is detailed in the supplementary material of this article.

In the case where we have $\tilde{M}_i^\top \tilde{M}_i \succcurlyeq \mathbb{I}_{(c-p_i)}, \forall i$, then, each element of the sum in $f_{R_M}(\Sigma)$ is g-convex on Σ on the set of SPD matrices such that $\Sigma_i \succcurlyeq \mathbb{I}_c$, where matrices $\Sigma_i \in \mathcal{P}_c$ ideally exist but are only partially observed.

4. Experimental analysis

The first experiment evaluates the proposed approach on an artificial dataset and compares it with an Euclidean approach. The second experiment demonstrates the advantage of averaging incomplete matrices over the classical methods of data imputation, that are the matrix deletion and channel (or variable) deletion. The third experiment, provided in the supplementary material of this article, evaluates the convergence rate of the new algorithm. The last experiment is conducted on a real dataset, acquired for a motor imagery experiment in BCI. For all experiments, we used uniform weights w_i for the matrices (hence leading to un-weighted averages). Experiments are made using pyRiemann 0.2.6 and Pymanopt 0.2.5 (Townsend et al., 2016).

4.1. Evaluation of convergence on synthetic dataset

For the first experiment, the dataset is defined such that the training samples are generated according to a distribution parametrized by dispersion σ and reference G. To this end, the reference matrix is generated as $G = U^\top D U$, where the diagonal matrix $D \in \mathbb{R}^{c \times c}$ has strictly positive values drawn from a triangular distribution and where the orthogonal matrix U is obtained as the eigenvectors of a random matrix $A \in \mathbb{R}^{c \times 2c}$. The sample matrices of the dataset $\{\Sigma_i\}_{i=1}^m$ are built as $\Sigma_i = U^\top (D + \epsilon_i) U$, where ϵ_i is an additive white noise drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. This perturbation is introduced to control the dispersion of matrices on the manifold during generation. Indeed, if $\lim_{m \to \infty} \bar{G} = G$, where $\bar{G}$ is the geometric mean of $\{\Sigma_i\}_{i=1}^m$, in practice it is equal to $\bar{G} = U^\top (D + \bar{\epsilon}) U$, with $\bar{\epsilon} = \frac{1}{m} \sum_i \epsilon_i$.

Then, the masks $\{M_i\}_{i=1}^m$ are generated with p missing variables. In these experiments, $m = 1000$, $c = 10$, and $p = 1, 2, 3$. The masked Riemannian mean $\bar{R}_M$ of incomplete matrices is computed by solving the optimization problem Eq. (10). The masked Euclidean mean $\bar{E}_M$ is computed using Eq. (14). Visualisations of this dataset are displayed in the supplementary material of this article.

To evaluate of convergence of the masked Riemannian mean $\bar{R}_M$, we compute its distance to the groundtruth mean $\bar{G}$. This distance is compared to the distance between the masked

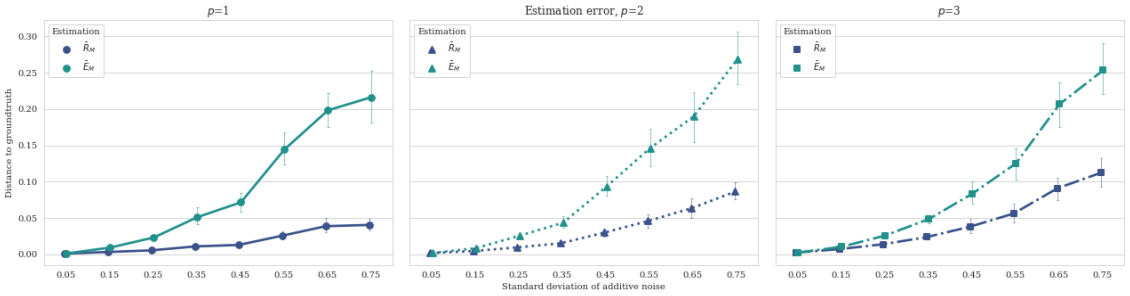

Figure 3: Convergence error for the masked Riemannian mean $\bar{R}_M$ (in dark blue) and the masked Euclidean mean $\bar{E}_M$ (light green). Distances to groundtruth $\bar{G}$ are evaluated with 10 repetitions, for $p = \{1, 2, 3\}$ missing variables and for $\sigma \in [0.05, 0.75]$.

Euclidean mean $\bar{E}_M$ and the groundtruth. To have a good overview, the comparison is repeated 10 times, computing the mean and standard deviations of distances. Moreover, increasing the standard deviation σ of additive noises ϵ_j allows to control the dispersion of matrices. Different number of missing variables p are also tested in this experiment.

Figure 3 compares the distances $\delta_R^2(\bar{R}_M, \bar{G})$ and $\delta_R^2(\bar{E}_M, \bar{G})$ when varying parameters σ and p. We observe that all distances increase with dispersion of matrices. Moreover, the masked Riemannian mean is always closer to the groundtruth than the Euclidean one. This result is extremely consistent across all tested parameters, showing that the masked Riemannian mean captures more information allowing to converge to the groundtruth faster than masked Euclidean one. This difference tends to decrease when p increases. We can conclude that the masked mean is more robust to missing data than its Euclidean counterpart, and that the convergence is effective even in case where only a fraction of the information is available.

4.2. Comparison with other strategies on synthetic dataset

When some variables are missing or have been removed after a *trimming* step (outlier rejection of very noisy variables), several methods of machine learning exist for data imputation. Notably, the *matrix deletion* consists of removing a matrix if any single variable is missing (Allison, 2001). Another common approach is to remove each variable that is missing in some of the matrices, called *variable deletion*.

Indeed, when the number of matrices with missing variables is small, matrix deletion is a good choice, that could even act as some sort of regularization. When the proportion of matrices with missing variables increases, the variable deletion may be a good strategy as it ensures all matrices are considered (even if it is with less information in each). The objective of this experiment is to compare the Riemannian masked mean $\bar{R}_M$ to the Riemannian mean applied after a matrix deletion strategy, denoted $\bar{R}_{md}$, and after a variable deletion strategy, denoted $\bar{R}_{vd}$.

We use the same data generation as previous experiment, using $m = 1000$, $c = 10$ and $\sigma = 0.5$. However, in this experiment only a given proportion of matrices (from 10

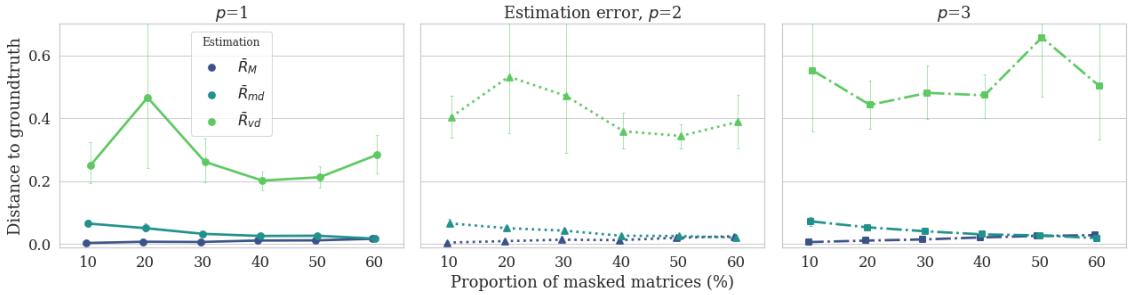

Figure 4: Riemannian mean estimation with different strategies: masked Riemannian mean $\bar{R}_M$, Riemannian mean computed after matrix deletion $\bar{R}_{md}$, and Riemannian mean after variable deletion $\bar{R}_{vd}$. Distances to groundtruth $\bar{G}$ are evaluated with 10 repetitions, for $p = \{1, 2, 3\}$ missing variables and for different proportions of matrices with missing variables (10 to 60 %).

to 60 %) have missing variables and are associated with a mask. Distances between the groundtruth and the different means are computed, and the comparison is repeated 10 times. As previously, different values of missing variables are tested.

Figure 4 compares the distances $\delta_R^2(\bar{R}_M, \bar{G})$, $\delta_R^2(\bar{R}_{md}, \bar{G})$ and $\delta_R^2(\bar{R}_{vd}, \bar{G})$ when varying the number of missing variables p and the proportion of masked matrices. This experiment shows that the Riemannian masked mean is close to the groundtruth and better than the other strategies, matrix deletion and variable deletion. This indicates that masked mean captures more information than other strategies, and that this information helps to converge to the groundtruth. As observed before, this difference tends to decrease when p increases.

We can conclude that masked mean is a good alternative to the deletion strategies, allowing to keep more statistical power in analysis (Olinsky et al., 2003) and avoiding to introduce a bias like the matrix deletion when missing variables are not random.

4.3. Validation on BCI dataset

This BCI experiment is conducted on a motor imagery dataset, acquired by Yi et al. (2014), where 10 subjects are performing a left hand/right hand movement and that should be detected using only their brain waves recorded with EEG. There are $m = 160$ trials, 80 for left hand movement and 80 for right hand, recorded for each subject on $c = 60$ channels/sensors.

Following the previous experiments, we generate random masks M_i to simulate missing channels, and apply them to covariance matrices Σ_i computed from EEG trials X_i. Then, we compare different means: a masked Riemannian mean $\bar{R}_M$ trained on all matrices of the training set, a Riemannian mean $\bar{R}_{md}$ trained only on the complete matrices (*i.e.* after applying a matrix deletion strategy on the raw training set), and a masked Euclidean mean $\bar{E}_M$. We rely on minimum-distance-to-mean (MDM) (Barachant et al., 2012) to embed these different means on a same classifier. This is a simple, robust, and online classifier that is part of the state-of-the-art for motor imagery BCI (Jayaram and Barachant, 2018). In this experiment, the gold standard is defined as the MDM trained on non-masked training

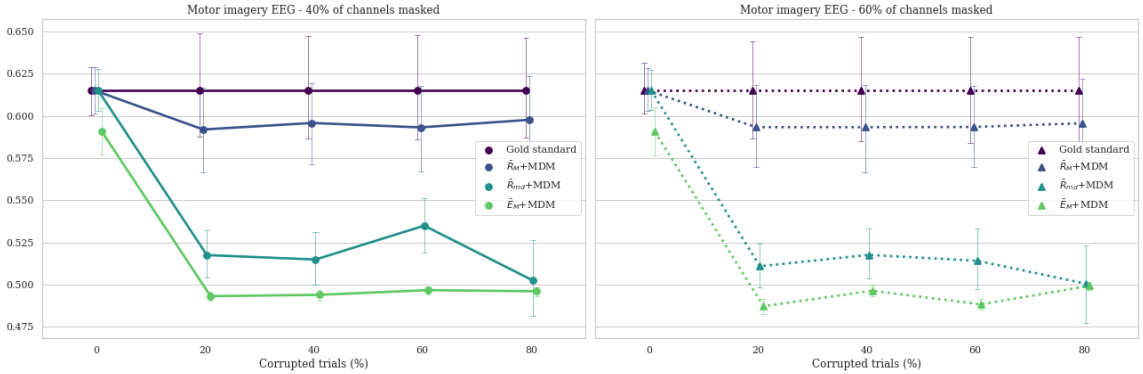

Figure 5: Comparison of MDM computed on non-masked matrices (gold standard, or $\bar{R}$+MDM), MDM based on masked Riemannian mean ($\bar{R}_M$+MDM), MDM based on Riemannian mean after matrix deletion ($\bar{R}_{md}$+MDM), and MDM based on masked Euclidean mean ($\bar{E}_M$+MDM). Left: Accuracy averaged for the 10 subjects, with 0 to 60% of matrices having 40% of masked channels (missing variables). Right: Same for 60% of masked channels.

set (with access to all information, with no masked or deleted variables). This gold standard thus indicates the results that could be obtained in optimal condition, without missing data.

We compare the classification accuracies obtained with a 5-fold cross-validation for each subject, while varying the proportion of masked matrices, from 0 to 60% of matrices with missing variables. The classifiers are evaluated in the two cases: when $p = 24$ out of the $c = 60$ EEG channels are masked (thus 40% of the channels) and when $p = 36$ channels are masked (60% of channels). These numerical experiments encompass the different situations where there is a moderate corruption in few trials to the case where the majority of the dataset is corrupted. The MDM classifier that relies on masked Riemannian mean $\bar{R}_M$ is called "$\bar{R}_M$+MDM". It is compared to MDM based on matrix deletion strategy, called "$\bar{R}_{md}$+MDM", and to MDM based on masked Euclidean mean, called "$\bar{E}_M$+MDM".

Accuracies are displayed on Fig. 5, with mean and standard deviation computed across subjects. The results clearly show that $\bar{R}_M$+MDM outperforms other classifiers. $\bar{E}_M$+MDM is not performing well, even while using a nan-mean is a common strategy during preprocessing. Euclidean MDM is known to perform poorly when compared to Riemannian MDM (Kalunga et al., 2015), as it is subject to the swelling effect. The matrix deletion strategy $\bar{R}_{md}$+MDM, as it uses only a subpart of the available information, achieves a better accuracy than $\bar{E}_M$+MDM, but is lower than $\bar{R}_M$+MDM. $\bar{R}_M$+MDM yields results that are almost at the level of the gold standard, and seems not sensitive to the proportion of matrices with missing channels.

5. Conclusion

This paper describes a method for the robust estimation of an average matrix from a set of SPD matrices with missing variables. This *masked mean* method could cope with any

number and distribution of missing variables on the considered set. The proposed method solves an optimization problem formulated as a matrix factorization, where the missing variables are discarded to estimate the solution on a collection of embedded Riemannian submanifold. While the convergence of this method is attested in practice, we have proved the geodesic convexity of the problem on a restricted class of SPD matrices. The experiments on synthetic and real datasets exhibit the practical interest of this method. The applications target covariance matrices, but without loss of generality this method could be applied on any set of SPD matrices.

The applicative contributions of the proposed method are multiple: (i) it is a true alternative to data imputation, avoiding to exclude data from analysis or to complete data with mean or median values (Experiments 4.1 and 4.2), (ii) it gives a more robust calibration of Riemannian classifiers, to obtain better classification scores (Experiment 4.3), and (iii) it provides a more robust covariance average for spatial filters estimation (Yger et al., 2015), allowing a supervised dimension reduction for high-dimensional signals enhancing the separation of classes (Barachant, 2014).

This method is flexible and paves the way to the extension of several other Riemannian approaches. For example, it could be used as a basis for extending SPD networks (Huang and Van Gool, 2017; Brooks et al., 2019) or dictionary of SPD matrices (Cherian and Sra, 2017) with missing data. Moreover, using a similar matrix decomposition, the proposed approach could go beyond missing variables as we are currently investigating the case of missing subspaces. A following work could address the data imputation problem with a *compressive sensing* flavor for SPD matrices.

Acknowledgments

FY acknowledges the support of the ANR as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). SC acknowledges that this work is not supported by any ANR or IDEX project but by the university recurrent funding.

References

P-A Absil, R Mahony, and R Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

P D Allison. *Missing data*, volume 136. Sage publications, 2001.

M Arnaudon, F Barbaresco, and L Yang. Riemannian medians and means with applications to radar signal processing. *IEEE J Sel Top Signal Process*, 7(4):595–604, 2013.

A Barachant. MEG decoding using riemannian geometry and unsupervised classification. Technical report, Grenoble University, 2014.

A Barachant, S Bonnet, M Congedo, and C Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Trans Biomed Eng*, 59(4):920–928, 2012.

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

NH Bingham. Modelling and prediction of financial time series. *Commun Stat Theory Methods*, 43(7):1351–1361, 2014.

W E Bishop and M Y Byron. Deterministic symmetric positive semidefinite matrix completion. In *Adv Neural Inf Process Syst*, pages 2762–2770, 2014.

N Boumal. An introduction to optimization on smooth manifolds, 2020. URL http://www.nicolasboumal.net/book.

D Brooks, O Schwander, F Barbaresco, J-Y Schneider, and M Cord. Riemannian batch normalization for SPD neural networks. In *Adv Neural Inf Process Syst*, pages 15463–15474, 2019.

E J Candès and B Recht. Exact matrix completion via convex optimization. *Found Comut Math*, 9(6):717, 2009.

Y Chen and Y Chi. Spectral compressed sensing via structured matrix completion. *arXiv preprint arXiv:1304.4610*, 2013.

A Cherian and S Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE Trans Neural Netw Learn Syst*, 28(12):2859–2871, 2017.

S Chevallier, EK Kalunga, Q Barthélemy, and F Yger. *Brain Computer Interfaces Handbook: Technological and Theoretical Advances*, chapter 19 - Riemannian classification for SSVEP based BCI : offline versus online implementations, pages 371–396. 2018.

S Chevallier, EK Kalunga, Q Barthélemy, and E Monacelli. Review of Riemannian distances and divergences, applied to SSVEP-based BCI. *Neuroinformatics*, pages 1–14, 2020.

H Choi, S Kim, and Y Shi. Geometric mean of partial positive definite matrices with missing entries. *Linear Multilinear A*, pages 1–26, 2019.

M Congedo, A Barachant, and R Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

M J Daniels and R E Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57 (4):1173–1184, 2001.

P T Fletcher, C Lu, S M Pizer, and S Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans Med Imaging*, 23(8):995–1005, 2004.

W Förstner and B Moonen. A metric for covariance matrices. Technical report, Bonn University, 1999.

M Fukuda, M Kojima, K Murota, and K Nakata. Exploiting sparsity in semidefinite programming via matrix completion I: General framework. *SIAM J Optim*, 11(3):647–674, 2001.

M Harandi, M Salzmann, and R Hartley. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE Trans Pattern Anal Mach Intell*, 40(1): 48–62, 2017.

I Horev, F Yger, and M Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *ACML*, pages 1–16, 2015.

Z Huang and L Van Gool. A Riemannian network for SPD matrix learning. In *AAAI CAI*, pages 2036–2042, 2017.

V Jayaram and A Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *J Neural Eng*, 15(6):066011, 2018.

C R Johnson. Matrix completion problems: a survey. In *Matrix Theory and Applications*, volume 40, pages 171–198, 1990.

C R Johnson and P Tarazaga. Connections between the real positive semidefinite and distance matrix completion problems. *Linear Algebra Appl*, 223:375–391, 1995.

E K Kalunga, S Chevallier, Q Barthélemy, K Djouani, Y Hamam, and E Monacelli. From Euclidean to Riemannian means: Information geometry for SSVEP classification. In *GSI*, pages 595–604. Springer, 2015.

L Korczowski, M Congedo, and C Jutten. Single-trial classification of multi-user p300-based brain-computer interface using Riemannian geometry. In *IEEE EMBC*, pages 1769–1772, 2015.

M Laurent. Polynomial instances of the positive semidefinite and Euclidean distance matrix completion problems. *SIAM J Matrix Anal Appl*, 22(3):874–894, 2001.

R J A Little. Robust estimation of the mean and covariance matrix from data with missing values. *J R Stat Soc Ser C Appl Stat*, 37:23–38, 1998.

K Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

M Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J Matrix Anal Appl*, 26(3):735–747, 2005.

S K Narang, A Gadde, and A Ortega. Signal processing techniques for interpolation in graph structured data. In *IEEE ICASSP*, pages 5445–5449, 2013.

G Obozinski, B Taskar, and M I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Stat Comput*, 20(2):231–252, 2010.

A Olinsky, S Chen, and L Harlow. The comparative efficacy of imputation methods for missing data in structural equation modeling. *Eur J Oper Res*, 151(1):53–79, 2003.

X Pennec, P Fillard, and N Ayache. A Riemannian framework for tensor computing. *Int J Comput Vis*, 66(1):41–66, 2006.

B Recht. A simpler approach to matrix completion. *J Mach Learn Res*, 12:3413–3430, 2011.

J DM Rennie and N Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, pages 713–719, 2005.

P Rodrigues, M Congedo, and C Jutten. A data imputation method for matrices in the symmetric positive definite manifold. In *GRETSI*, 2019.

T Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J Clim*, 14:853–871, 2001.

S Sra and R Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM J Optim*, 25(1):713–739, 2015.

J Townsend, N Koep, and S Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *J Mach Learn Res*, 17(1):4755–4759, 2016.

K Tsuda, S Akaho, and K Asai. The EM algorithm for kernel matrix completion with auxiliary data. *J Mach Learn Res*, 4:67–81, 2003.

G Vinci, G Dasarathy, and G I Allen. Graph quilting: graphical model selection from partially observed covariances. *arXiv preprint arXiv:1912.05573*, 2019.

K Q Weinberger and L K Saul. Unsupervised learning of image manifolds by semidefinite programming. *Int J Comput Vis*, 70(1):77–90, 2006.

F Yger and M Sugiyama. Supervised LogEuclidean metric learning for symmetric positive definite matrices. *arXiv preprint arXiv:1502.03505*, 2015.

F Yger, F Lotte, and M Sugiyama. Averaging covariance matrices for EEG signal classification based on the CSP: An empirical study. In *EUSIPCO*, pages 2721–2725, 2015.

F Yger, M Berar, and F Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Trans Neural Syst Rehabil Eng*, 25(10):1753–1762, 2017.

W Yi, S Qiu, K Wang, H Qi, L Zhang, P Zhou, F He, and D Ming. Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PloS One*, 9(12), 2014.

P Zadeh, R Hosseini, and S Sra. Geometric mean metric learning. In *ICML*, pages 2464–2471, 2016.