A Data Imputation Method for Matrices in the Symmetric Positive Definite Manifold

Pedro Luiz Coelho Rodrigues, Marco Congedo, Christian Jutten

A Data *Imputation* Method for Matrices in the Symmetric Positive Definite Manifold

Pedro L. C. RODRIGUES[1], Marco CONGEDO[1], Christian JUTTEN[1]

GIPSA-lab, CNRS, University Grenoble Alpes, Grenoble Institute of Technology
11 Rue des Mathématiques, 38400 Saint-Martin d'Hères, France
pedro.rodrigues@gipsa-lab.fr

Résumé – L'*imputation* de données manquantes (aussi connue sous le nom de complétion de données) est une technique standard d'analyse statistique de données utilisée pour remplacer des valeurs manquantes sur une base de données. Elle est utile, par exemple, lorsqu'un ou plusieurs capteurs d'un enregistrement exprimental présentent des problèmes et doivent être rejetés. Dans ce travail, nous présentons une méthode pour compléter les valeurs d'une matrice symétrique définie positive lorsqu'elle est utilisée comme descripteur statistique d'une série temporelle multiple. Nous illustrons notre contribution dans des tâches de classification pour des Interfaces Cerveau-Machine et comparons sa performance à celle d'une méthode de référence.

Abstract – Missing-data *Imputation* (sometimes called data completion) is a standard technique from statistical data analysis used for replacing missing values in databases. It is useful, for instance, when one (or several) of the sensors on an experimental recording presents a problem and has to be discarded. In this work, we present a method for completing missing values of symmetric positive definite matrices when they are used for describing the statistics of a multivariate time series recorded on multiple sensors. We illustrate our proposal on classification tasks in the field of Brain-Computer Interfaces and compare its performance to a reference method.

1 Introduction

Missing-data *Imputation* (sometimes called data completion) is a standard technique from statistical data analysis used for replacing missing values in databases [7]. It is useful when working with spreadsheets, where the values of some attributes on certain rows may be missing due to technical problems, and also when monitoring physical experiments, where one (or more) sensors may present recording problems.

One might argue that it would be easier to simply discard data points presenting problems in one (or more) of their dimensions instead of filling missing values via an imputation method. However, there are many situations where gathering data points may be expensive and discarding even just a few of them is not acceptable. An example of such an application are Brain-Computer Interfaces (BCI), where each data point of a dataset is the recording of electroencephalographic (EEG) activity of a subject that may last for a few seconds. In this case, discarding a trial because of only one or a few malfunctioning electrodes is not desirable.

In this work, we propose a method for imputation when the data points are n-dimensional symmetric positive definite (SPD) matrices. In this case, the goal is to be able to augment an m-dimensional SPD data point ($m < n$) into an n-dimensional data point respecting some geometric constraints. An important application is in the field of BCI, where the state-of-the-art methods for classification employ spatial covariance matrices as statistical features, which are SPD matrices [5]. An EEG epoch recorded using n electrodes yields an n-dimensional SPD matrix, but if p electrodes present problems and have to be discarded, the data point becomes a m-dimensional SPD matrix, with $m = n - p$. Our imputation method transforms the m-dimensional matrix into an n-dimensional matrix respecting geometric and statistical constraints.

The rest of the present paper is organized as it follows : we start with a brief introduction to concepts of Riemannian geometry for SPD matrices. Then, we present our data imputation method as well as its interpretations in the context of multivariate time series. Finally, we describe our numerical illustrations on BCI datasets and discuss the results.

2 The SPD manifold

We denote the set of n-dimensional SPD matrices by $\mathcal{P}(n)$ and endow it with the Affine Invariant Riemannian Metric [2]. This choice turns $\mathcal{P}(n)$ into a symmetric Riemannian manifold with non-negative curvature whose geodesic distance between $X_i, X_j \in \mathcal{P}(n)$ is [2]

$$\delta_R^2(X_i, X_j) = \left\| \log\left(X_i^{-1/2} X_j X_i^{-1/2} \right) \right\|_F^2 . \qquad (1)$$

The geometric mean (or center of mass) $M_{\mathcal{X}}$ of a set of points

$$\mathcal{X} = \{X_1, \ldots, X_K\} \subset \mathcal{P}(n)$$

is defined as [2]

$$M_{\mathcal{X}} = \underset{X \in \mathcal{P}(n)}{\operatorname{argmin}} \sum_{k=1}^{K} \delta_R^2(X, X_k) \,, \tag{2}$$

and we denote by $d_{\mathcal{X}}$ the dispersion of the set of points, that is, the minimum value attained by the cost function in (2).

There are many ways to describe the geometry of a set of points. Some consider all the pairwise distances between them, others are based on an empirical estimation of their statistical distribution [8]. In this work, we describe the datasets via their center of mass and their dispersion. We use the following notation to refer to the geometry of a dataset $\mathcal{X}$ to which a data point X_i belongs to :

$$X_i \sim (M_{\mathcal{X}}, d_{\mathcal{X}}) \,.$$

3 The imputation method

3.1 Problem statement

Consider a dataset $\mathcal{X} \subset \mathcal{P}(n)$ and a data point $Y \in \mathcal{P}(m)$ ($m < n$). We assume that Y is the only element to which we have access from a dataset $\mathcal{Y}$ whose geometry can be summarized by $(M_{\mathcal{Y}}, d_{\mathcal{Y}})$. Our imputation method is a transformation from $\mathcal{P}(m)$ to $\mathcal{P}(n)$ such that, when it is applied to Y, the new data point $\widetilde{Y}$ has the same geometric characteristics as the points coming from $\mathcal{X}$. We require then :

$$Y \sim (M_{\mathcal{Y}}, d_{\mathcal{Y}}) \xrightarrow{\text{imputation}} \widetilde{Y} \sim (M_{\mathcal{X}}, d_{\mathcal{X}}) \,. \tag{3}$$

3.2 The transformations

The imputation method consists of 4 transformations :

Step 1 **(re-center to Identity)** : transform Y so its description is centered around the m-dimensional Identity matrix :

$$Y \sim (M_{\mathcal{Y}}, d_{\mathcal{Y}}) \xrightarrow{\text{re-center}} Y_{\text{rct}} \sim (I_m, d_{\mathcal{Y}}) \,, \tag{4}$$

where

$$Y_{\text{rct}} = M_{\mathcal{Y}}^{-1/2} \, Y \, M_{\mathcal{Y}}^{-1/2} \,. \tag{5}$$

Step 2 **(scaling)** : adapt the dispersion around the mean for Y_{rct} so that it is equal to $d_{\mathcal{X}}$:

$$Y_{\text{rct}} \sim (I_m, d_{\mathcal{Y}}) \xrightarrow{\text{scaling}} Y_{\text{str}} \sim (I_m, d_{\mathcal{X}}) \,, \tag{6}$$

where

$$Y_{\text{str}} = (Y_{\text{rct}})^s \quad \text{and} \quad s = \frac{d_{\mathcal{X}}}{d_{\mathcal{Y}}} \,. \tag{7}$$

Step 3 **(expand)** : expand Y_{str} to an n-dimensional matrix :

$$Y_{\text{str}} \sim (I_m, d_{\mathcal{Y}}) \xrightarrow{\text{expand}} \widetilde{Y}_{\text{str}} \sim (I_n, d_{\mathcal{X}}) \,, \tag{8}$$

where

$$\widetilde{Y}_{\text{str}} = \begin{bmatrix} Y_{\text{str}} & 0_{m \times (n-m)} \\ 0_{(n-m) \times m} & I_{(n-m)} \end{bmatrix} \tag{9}$$

and $0_{p \times q}$ is a $p \times q$ matrix filled with zeros.

Step 4 **(re-center to $M_{\mathcal{X}}$)** : transform $\widetilde{Y}_{\text{str}}$ so its description is centered around $M_{\mathcal{X}}$:

$$\widetilde{Y}_{\text{str}} \sim (I_m, d_{\mathcal{X}}) \xrightarrow{\text{re-center}} \widetilde{Y} \sim (M_{\mathcal{X}}, d_{\mathcal{X}}) \,, \tag{10}$$

where

$$\widetilde{Y} = M_{\mathcal{X}}^{1/2} \, \widetilde{Y}_{\text{str}} \, M_{\mathcal{X}}^{1/2} \,. \tag{11}$$

This sequence of steps applied to Y gives as output a new data point $\widetilde{Y}$ that satisfies Equation 3. The geometric motivation behind these transformations, as well as their formal justifications, are presented in [10].

3.3 A time series interpretation

In our formalism up to here, we have only said that the data points in $\mathcal{X}$ are n-dimensional SPD matrices. From now on, we will assume that these data points are in fact spatial covariance matrices describing the statistics of multivariate time series.

We consider having a dataset $\mathcal{S}$ with K recordings over n electrodes during T time samples so that

$$\mathcal{S} = \{S_1, \ldots, S_K\} \subset \mathbb{R}^{n \times T} \,. \tag{12}$$

To each element $S_i \in \mathcal{S}$ we associate a spatial covariance matrix X_i and form a dataset $\mathcal{X} \subset \mathcal{P}(n)$. In this context, having a data point $Y \in \mathcal{P}(m)$ means having in fact a recording with problems over p electrodes so that only $m = n - p$ time series can be used. We denote this recording $S \in \mathbb{R}^{m \times T}$.

We can interpret the steps of our data imputation method as operations over the multivariate time series S as follows :

1. First, apply a whitening matrix to make the time series on each electrode approximately uncorrelated to each other (Step 1) :

$$S_{\text{whitened}} = M_{\mathcal{Y}}^{-1/2} S \Rightarrow Y_{\text{whitened}} \simeq I_m. \tag{13}$$

2. Then, add p new dimensions to the multivariate time series and fill them with uncorrelated white noise (Step 3) :

$$\widetilde{S}_{\text{whitened}} = \begin{bmatrix} S_{\text{whitened}} \\ \mathbf{s}_p \end{bmatrix} \Rightarrow \widetilde{Y}_{\text{whitened}} \simeq I_n \,, \tag{14}$$

where $\mathbf{s}_p$ is a realization over T time samples of a p-dimensional time series with zero mean and covariance matrix I_p. For instance, it could be a spatially uncorrelated zero-mean Gaussian white noise.

3. Dewhiten $\widetilde{S}_{\text{whitened}}$ by doing (Step 4) :

$$\widetilde{S} = M_{\mathcal{X}}^{1/2} \widetilde{S}_{\text{whitened}} \Rightarrow \widetilde{Y} \simeq M_{\mathcal{X}} \,. \tag{15}$$

Note that this step mixes the newly added $\mathbf{s}_p$ with the time series from other electrodes.

The imputation method can be seen as a general way to fill p dimensions of S with time series whose second order statistics have some desired structure. For particular applications, there are other methods to solve this problem. For instance, for magnetoencephalographic (MEG) and EEG signals the method of reference is the *spherical spline interpolation* [9], which fills the signals on problematic channels by taking linear combinations of the time series on electrodes which are spatially close to them.

3.4 Estimating $M_{\mathcal{Y}}$ and $d_{\mathcal{Y}}$

Our imputation method relies on the estimation of parameters $M_{\mathcal{Y}}$ and $d_{\mathcal{Y}}$, but we have access to just one data point Y from $\mathcal{Y}$. We cope with this limitation by doing the following : discard the same p problematic electrodes related to S from all n-dimensional data points in $\mathcal{S}$. Estimate their spatial covariance matrices and denote this new dataset $\mathcal{X}^{(-p)} \subset \mathcal{P}(m)$. Estimate the geometric mean and the dispersion for the data points in $\mathcal{X}^{(-p)}$ and use them as estimates for $M_{\mathcal{Y}}$ and $d_{\mathcal{Y}}$. This procedure relies on the assumption that datasets $\mathcal{X}^{(-p)}$ and $\mathcal{Y}$ are similar to each other, which is reasonable when considering that Y was obtained during the same experiment that generated the data points from $\mathcal{X}$.

4 Numerical Illustrations

4.1 The dataset

We illustrate our method on a database containing electroencephalographic (EEG) recordings of a brain-computer interface (BCI) experiment [3]. The database contains recordings on 23 electrodes of 52 subjects executing a left-hand/right-hand motor imagery paradigm. The signals are band-pass filtered between 8 Hz and 35 Hz (sampling frequency is 512 Hz) and epoched into one hundred 3-second trials : 50 trials on the left-hand class (labeled 0) and 50 trials on the right-hand class (labeled 1). Such pre-processing yields for each subject a set of data points $\mathcal{S}$ as defined in (12) with $m = 23$, $T = 1536$, and $K = 100$, as well as a set of labels associated to it

$$\mathcal{L} = \{\ell_1, \ldots, \ell_{100}\} \text{ with } \ell_i \in \{0, 1\} \,. \tag{16}$$

For each element $S_i \in \mathcal{S}$ we estimate a spatial covariance matrix X_i using Ledoit-Wolf shrinkage [6], which ensures its good numerical conditioning. The set of spatial covariances forms the dataset $\mathcal{X} \subset \mathcal{P}(23)$.

When an epoch has a problem on p electrodes, it generates a data point $S \in \mathbb{R}^{m \times 1536}$ where $m = 23 - p$. Without loss of generality, we will consider that the dimensions related to these discarded electrodes correspond to the p last dimensions of the data points in $\mathcal{X}$. The spatial covariance matrix Y estimated from this epoch is an element of $\mathcal{P}(m)$.

4.2 The classification procedure

In this paper, every classification task is done using the Minimum Distance to Mean classifier (MDM) [1], which is a generalization of the nearest-centroid classifier to the space of SPD matrices. It works by first estimating the geometric mean of the elements of each class in the training dataset (the class means). Then, it assigns to each unlabeled data point the class of the nearest class mean according to the geodesic distance defined in (1). The classification score is simply the average accuracy of the classifier.

We demonstrate the usefulness of our imputation method by comparing the classification score obtained via MDM on three different methods. Firstly, we get the accuracy on a dataset where there is no problem on any electrode (we call this the **full** method). Then, we consider the case when a few data points have problems on a set of p arbitrarily chosen electrodes. For this, we select a few epochs from the dataset used in the **full** method and emulate the problematic electrodes by discarding them from the recordings. We apply our imputation method to augment the dimension of these data points and compute the accuracy of MDM for classifying them (**imputation** method). As a comparison, we consider the case when the matrices are augmented via spherical spline interpolation (**interpolation** method).

The scores are obtained via cross-validation. The data points in the *training* dataset are always n-dimensional and used for estimating the centroids for the MDM classifier. The unlabeled data points from the *testing* dataset are either n-dimensional (**full** method) or m-dimensional (**imputation** and **interpolation** methods).

5 Results and discussion

In the results described below, we have used knowledge of the neurophysiology of BCI experiments in the motor imagery paradigm to consider settings with different combinations of EEG electrodes as problematic. We chose channels located in the motor cortex, which are known to carry important information for classifying the trials (C3 and C4), as well as electrodes which are not relevant for this kind of paradigm (Fz and Pz) [4]. See Figure 1 for a representation of the spatial disposition of the 23 electrodes used for the recordings in the database.

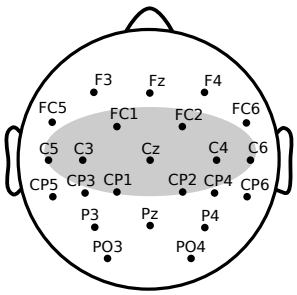

FIGURE 1 – Diagram with the electrodes configuration. The gray area indicates where the sensory motor cortex is approximately located, which is the region expected to be the most relevant for classification tasks in the motor imagery paradigm.

Table 1 displays the classification scores for the **imputation** and **interpolation** methods when different electrodes are considered as problematic. The score obtained with the **full** method was 0.663 and serves as a reference for our comparisons.

We observe that the scores with the **imputation** method when only the Fz and/or the Pz electrodes are missing is not very different from that of the **full** method. In fact, a paired t-test comparing the scores for each subject of the database indicates no

TABLE 1 – Average accuracy scores for the **imputation** and **interpolation** methods over the 52 subjects in the database (standard deviation inside parenthesis). The missing electrodes column indicates which electrodes were discarded in each case. The average accuracy for the **full** method was 0.66.

missing electrodes	**imputation**	**interpolation**
$\{Fz\}$	0.66 (0.11)	0.64 (0.10)
$\{Pz\}$	0.66 (0.10)	0.63 (0.10)
$\{Fz, Pz\}$	0.66 (0.10)	0.61 (0.10)
$\{C3\}$	0.65 (0.10)	0.63 (0.10)
$\{C4\}$	0.65 (0.10)	0.61 (0.08)
$\{C3, C4\}$	0.64 (0.09)	0.61 (0.09)

evidence for rejecting the null hypothesis of them being equal. Such a result is not surprising, since the referred electrodes were not expected to carry relevant information to discriminate between the classes of the experiment. However, when the C3 and/or the C4 are missing, the important discriminative information provided by these channels can not be replaced by our imputation method, so the average classification score decreases.

We also note that the **imputation** method consistently yields better results, in average, as compared to the **interpolation** method. We performed paired t-tests to compare the results of the two methods and the null hypothesis of equal scores was always rejected with p-values smaller than 10^{-3} (corrected for the multiple comparisons problem via the Bonferroni correction). A possible explanation for this could be the diversity of information used by our imputation procedure as compared to the interpolation method, since it adds new dimensions to the problematic m-dimensional Y matrix using information from the rest of the dataset $\mathcal{X}$, whereas spherical spline interpolation uses only information from the time series S related to Y. Furthermore, because the p dimensions added to S are simply linear combination of its m time series, the rank of $\widetilde{S} \in \mathbb{R}^{n \times 1536}$ is just m. As a consequence, although the estimated $\widetilde{Y}$ has no zero eigenvalues (because of the Ledoit-Wolf shrinkage), some of its eigenvectors point to directions which are not descriptive and may prejudice the classification procedure.

It should be mentioned that the matrix augmentation scheme provided by our imputation method is purely based on the distribution of the spatial covariance matrices of each trial. This means that there is no physiological interpretation for the time series obtained on the p added dimensions. However, one could try to determine a physiologically plausible $\mathbf{s}_p$ in (14) with the statistical properties required by the imputation method.

6 Conclusions

We have presented a new method for augmenting the dimensions of an m-dimensional SPD matrix to make it compatible with datasets having n-dimensional data points. One important application of this method is in the case of EEG multivariate time series, where an electrode (or more than one) may present problems and have to be discarded. Our method provides a way for filling the values on this discarded electrode and allow it to be combined in a classification pipeline with other epochs. Comparisons with a reference method from the literature [9] show that our proposal is superior when the data imputation is used in classification pipelines.

Further work may include the use of physiological and spatial information from the location of the EEG electrodes to improve the imputation method and allow for the extraction of meaningful time series from the added dimensions.

Références

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

[2] Rajendra Bhatia. *Positive definite matrices*. Princeton university press, 2009.

[3] Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. EEG datasets for motor imagery brain–computer interface. *GigaScience*, 6(7) :1–8, may 2017.

[4] Marco Congedo. *EEG Source Analysis* . Habilitation à diriger des recherches, Université de Grenoble, October 2013.

[5] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces ; a primer and a review. *Brain-Computer Interfaces*, pages 1–20, 2017.

[6] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2) :365 – 411, 2004.

[7] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2002.

[8] Facundo Mémoli. A spectral notion of Gromov-Wasserstein distance and related methods. *Applied and Computational Harmonic Analysis*, 30(3) :363 – 401, 2011.

[9] F. Perrin, J. Pernier, O. Bertrand, and J.F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2) :184 – 187, 1989.

[10] P. L. C. Rodrigues, C. Jutten, and M. Congedo. Riemannian procrustes analysis : Transfer learning for brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, pages 1–1, 2018.