Group Learning by Joint Alignment in the Riemannian Tangent Space

MARCO CONGEDO[1], ALEXANDRE BLEUZÉ[1], JÉRÉMIE MATTOUT[2]

[1] GIPSA-lab, Université Grenoble Alpes, CNRS, Grenoble-INP
11 rue de la Piscine, Grenoble Campus BP46, F-38402, France
[2] Lyon Neuroscience Research Centre, INSERM, CNRS, Université Claude Bernard Lyon 1, Université de Lyon
CHS Le Vinatier, 95 Bd Pinel, 69635 Cedex Bron, France

[1]Marco.Congedo@gipsa-lab.fr, Alexandre.Bleuze@gipsa-lab.fr,
[2]Jeremie.Mattout@inserm.fr

Résumé – Dans le domaine des interfaces cerveau-ordinateur (ICO), les modèles d'apprentissage sont typiquement entrainés sur chaque sujet et chaque session séparément, comme les données ne sont pas alignées entre sessions et entre sujets. Ici nous proposons une méthode pour l'apprentissage de groupe, c'est-à-dire pour l'apprentissage simultané à l'aide de plusieurs sujets et/ou sessions, après les avoir alignés de façon conjointe. Notre méthode s'inspire de la littérature sur la séparation aveugle de source. Comme démonstration, nous entrainons un modèle unique d'apprentissage sur une base de données de 22 sujets et nous appliquons ce modèle de groupe pour prédire les données de test de façon analogue pour tous les sujets. Nous observons une augmentation moyenne conséquente de 6.8 points de précision comparé à un paramétrage entrainement-test individuel classique. Notre méthode est générique et peut être employée dans n'importe quelle application. Elle peut aussi être utilisée pour entrainer des modèles d'apprentissage qui requièrent un très grand nombre de données, tel que les réseaux de neurones profonds.

Abstract – In the brain-computer interface (BCI) field the machine learning models are usually trained for each subject and each session separately, since data are misaligned between subjects and between sessions. In this article we propose a method for group learning, that is, for learning from many different subjects and/or sessions after jointly aligning them. Our method is inspired from the literature on joint blind source separation. As a demonstration, we fit a unique machine learning model on a 22-subject BCI database and we apply such a group model to predict test data on all subjects alike. We observe a highly significant average 6.8-point accuracy increase as compared to the classical individual train-test setting. Our method is general and may be applied in any applications. For instance, it may be used to fit machine learning models requiring very large amount of data, such as deep neuronal networks.

1 Introduction

A Brain-Computer Interface (BCI) is a computerized system for on-line prediction of cognitive states and intentions of the user [1]. In this article we focus on BCIs based on electroencephalography (EEG), a modality that is completely non-invasive, safe and silent, but also affordable to the large public due to recent advances in micro-technology. Typical BCIs operate in two phases : in the *training phase* the classifier is calibrated in a supervised fashion, that is, with examples of EEG data corresponding to labeled classes ; the actual use of a BCI is named the *test phase* and must be unsupervised, that is, the BCI must classify EEG data to infer the classes they belong to. In general, a training phase precedes every actual use of the BCI because the machine learning (ML) model cannot be generalized to subsequent sessions of the same user and, even less so, to other users. However, the calibration phase is time consuming and tiring, hence highly impractical both for healthy and clinical users [2]. For this reason, recent research has focused on transfer learning (also named *domain adaptation*) methods in order to transfer the ML model from one session/subject to another. In this field the domain we ought to use for learning is referred to as the *source* and the domain we want to apply the learning to as the *target*. Recently, there have been significant advances in transfer learning methods thanks to the inception in the BCI field of Riemannian geometry [3]. In this framework EEG segments are encoded in the form of symmetric positive-definite (SPD) matrices and manipulated as points in their natural Riemannian manifold [4]. The first proposition aimed at parallel-transporting all points for both the source and target domain so as to *recenter* them around the identity [5]. This amounts to whitening the observations, a well-known pre-processing step in the signal processing community. Similarly, in [6] the authors recentered to the midpoint between the centers of mass of the source and target data.

The above recentering procedures acts as *translations*. Inspired by Procrustes analysis, the authors in [7] proposed to add two more matching steps following recentering: a *stretching* operation to match the dispersion of the points of the source and target sets and a *rotation*, reminiscent of the well-known whitening + rotation procedure used in signal processing for blind source separation.

The methods described so far carry out the transfer learning on the manifold. Usually the data are then projected onto the *tangent space* for classification purposes, since the tangent space is Euclidean and powerful ML algorithms can be applied therein. In [8] a recentering on the manifold is followed by the projection

onto the tangent space, where the tangent vectors are submitted to a principal component analysis (PCA), independently for the source and target data sets. The two PCAs being unrelated, this method proves insufficient for aligning the source and target tangent vectors and suffers from an unsolvable sign ambiguity. In [9] the present authors proposed to align the tangent vectors using *Procrustes Analysis* in the tangent space. Under certain conditions this is equivalent to aligning the source and target tangent vectors by *Maximum Covariance Analysis* (MCA), which does align the data and does not suffer from sign ambiguities.

All the above studies have been concerned with transfer learning from one domain to another. In this article we take a more general route. Given M > 2 domains, we wondered how they can be employed jointly for learning purposes. Rather than aligning a target data set to a source data set, or vice versa, we require to use *many* source domains and obtain *group learning*. As a matter of fact, we have now access to a large quantity of BCI data performed under similar, sometimes even identical, conditions. For example, for a given BCI system, all previous users of the system constitute a (possibly large) database. Not using such resource is, evidently, a waste in several respects. Motivated by this, we prepare now to present our method.

2 Method

We present and test the method for the case of BCIs based on event-related potentials (ERPs). Only the pre-processing and encoding parts of the entire pipeline are to be adapted in case of other paradigms. We will describe the pre-processing steps only briefly as they are state-of-the-art procedures in the field of ERP-based BCIs.

2.1 ERP-based BCIs

In ERP-based BCI the user is confronted with a continuous stream of discrete sensory stimuli. Among them, the one the user wants to select (e.g., the flashing of a specific letter in a BCI speller) is *salient* and all the others (e.g., the flashing of all other letters) are *non-salient*. All stimuli evoke stereotypical electric potential in the brain, lasting up to 1s [10]. The EEG stream is hence segmented in 1s-segments, named *trials*, starting at the exact moment the stimulations are delivered. The goal of such a BCI is to understand what stimulus is salient for the user at a given time, given several stimulations. This is possible because the ERPs are different for salient and non-salient stimuli. A 2-class (salient vs. non-salient) ML problem is posed. In the BCI data we analyze here the ratio between salient and non-salient stimuli is 1:6, thus the classes are unbalanced.

2.2 Pre-processing

The EEG data is band-pass (1-16 Hz) filtered applying a second-order forward-backward Butterworth digital filter featuring linear phase response and segmented to extract the trials, as explained above. The trials with excessive artefacts are excluded from analysis by means of a data-driven amplitude-thresholding procedure.

2.3 Encoding

Let $m \in \{1,\ldots,M\}$ be the index of M subjects. Let $X_{ml} \in \mathbb{R}^{N \cdot T}$ be the matrix holding the EEG data of the l^{th} trial, where N is the number of electrodes and T the number of samples comprising one second. In this work we assume N and T be the same for all subjects, although this is not a requirement. First, the stereotypical ERP response for the salient stimuli is estimated using the weighted least-square estimation detailed in [10]. Only the first D=4 principal components of the stereotypical response for each subject, denoted $P_m \in \mathbb{R}^{D \cdot T}$, are retained. Then, the trials are augmented [11], such as

$$Y_{ml} = \begin{bmatrix} X_{ml} \\ P_m \end{bmatrix}, \quad (1)$$

their covariance matrix is estimated using the linear Ledoit and Wolf shrinkage estimator [12] and normalized so as to have unit determinant. Let us denote those normalized covariance matrix estimators $C_{ml} \in \mathbb{R}^{N+D \cdot N+D}$. The center of mass G_m of all matrices C_{ml} (i.e., of both salient and non-salient trials) is found for each subject as the weighted *geometric mean* according to the affine-invariant (Fisher-Rao) metric on the Riemannian manifold of SPD matrices [4]. Being the classes unbalanced, the weights are chosen so as to give equal overall weight to the salient and non-salient trials.

2.3.1 Tangent space projection

The aforementioned recentering to the identity matrix (whitening) simplifies the projection onto the tangent space of the trials. Both operations are carried out as

$$S_{ml} = \log\left(G_m^{-1/2} C_{ml} G_m^{-1/2} \right), \quad (2)$$

where log(.) is the matrix logarithm of the argument. The upper (or lower) triangle of symmetric matrices S_{ml} are then vectorized giving weight 1 to the diagonal elements and $\sqrt{2}$ to the off-diagonal elements, yielding the vectorized tangent vector v_{ml}; the weights ensure that the 2-norm of v_{ml} is the same as the Frobenius norm of S_{ml} [3]. Furthermore, we remove from v_{ml} the corresponding elements of S_{ml} that depends only on P_m in (1) ; those elements are in fact identical for all trials, thus do not hold any discriminant information. The final vectors v_{ml} are therefore of dimension $E=(N^2+N+2DN)/2$. In this study (N=16, D=4), E=200.

2.4 Group Alignment

Let $k \in \{1,\ldots,K\}$ be the index of K classes. In this study K=2, but in the sequel it may be any natural number. Let

T_{mk} be the matrix formed by stacking horizontally a number of bootstrapped average estimations of vectors v_{ml} (section 2.3.1), for each subject and each class separately. In this work, E of such bootstraps are extracted, that is, as many as the dimension of the vectors, and each one is obtained averaging 10 vectors randomly drawn with replacement. The Euclidean metric is used for averaging. This yields a matrix of E column tangent vectors $T_{mk}\ \mathbb{R}^{\mathrm{E}\cdot\mathrm{E}}$. Now, let us define all cross-products

$$R_{ijk} = T_{ik} T_{jk}^{T} \text{ , for all } i \neq j \in \{1,\ldots,\mathrm{M}\} \text{ and } k \in \{1,\ldots,\mathrm{K}\}. \quad (3)$$

In order to align the tangent vectors of all subjects we require to find M matrices B_m such that the *cross-products* in (3) transformed such as

$$B_i^T R_{ijk} B_j \quad (4)$$

are as diagonal as possible, according to some criterion. This is a generalization of the case K=1 and M=2, yielding a single cross-product $R_{12,}$ the left and right singular values of which provide the sought solutions B_1 and B_2. Such solution, known as *maximum covariance analysis*, diagonalizes R_{12} exactly, with B_1, B_2 being orthogonal matrices. For the general case (M>2, K>1), which is of concern here, the diagonalizations of the cross-products R_{ijk} is only approximate. Furthermore, we do not need to constraint the solution matrices to the orthogonal group.

The same problem has been encountered by the signal processing community in completely different contexts, known as *independent vector analysis* and *joint blind source separation* [13]-[16]. To solve it, we adopt the widespread off-diagonal least-squares criterion and a known *approximate joint diagonalization* (AJD) gradient descent scheme. Possible sign and permutation ambiguities, which are inherent to non-crossed AJD problems, can be solved in our case finding signed permutation matrices for $B_1,\ldots,B_{\mathrm{M}}$ so as to make the diagonal elements of cross-products (4) positive and sort in descending order their sum across *m* and *k*. We report here the *Group Alignment Algorithm* in pseudo-code ; for the cost function, the derivation of the gradient and the optimization scheme, the reader is referred to [15], [16].

Once optimized matrices $B_1,\ldots,B_M$, we jointly align all tangent vectors for all subjects, regardless their class, by means of the following non-orthogonal projection :

$$v_{ml} \leftarrow B_m \left(B_m^T B_m \right)^{-1} B_m^T v_{ml} \,. \quad (5)$$

Group Alignment Algorithm

Input

subspace dimension P<<E,

matrices $R_{ijk} = T_{ik} T_{jk}^T \ \forall\ i, j \in 1{:}\mathrm{M}$ and $k \in 1{:}\mathrm{K}$ (3)

Dimensionality reduction by pre-whitening

find $W_1,\ldots,W_{\mathrm{M}}$ satisfying $W_m^T \sum_{k=1}^{\mathrm{K}} R_{mmk} W_m = I_{\mathrm{P}} \ \forall\ m \in 1{:}\mathrm{M}$

transform $R_{ijk} \leftarrow W_i^T R_{ijk} W_j \ \ \forall\ i \neq j \in 1{:}\mathrm{M}$ and $k \in 1{:}\mathrm{K}$

initialize $U_m = I_{\mathrm{P}} \ \forall\ m \in 1{:}\mathrm{M}$

---

repeat

$\forall\ m \in 1{:}\mathrm{M}$ begin

$Z_{mp} = \sum_{k=1}^{K} \sum_{j \neq m=1}^{M} \left(R_{mjk} u_{pj} u_{pj}^T R_{mjk}^T \right) \ \forall\ p \in 1{:}\mathrm{P}$

(u_{pj} is the p^{th} column vector of U_j and u_{pj}^T its transpose)

compute Cholesky decomposition $\sum_{p=1}^{\mathrm{P}} Z_{mp} = LL^T$

$\forall\ p \in 1{:}\mathrm{P}$ begin

solve $Lf = Z_{mp} u_{pm}$ for f and $L^T g = f$ for g

update $u_{pm} \leftarrow g \left(g^T Z_{mp} g \right)^{-1/2}$

end $\forall\ p$

end $\forall\ m$

until convergence of all matrices $U_1,\ldots,U_M$

normalize all columns of matrix $U_m \ \forall\ m \in 1{:}\mathrm{M}$

Output

$B_m = W_m U_m F_m \ \ \forall\ m \in 1{:}\mathrm{M}$,

with $F_1,\ldots,F_{\mathrm{M}}$ signed permutation matrices (see text)

3 Results

We have tested the proposed group learning approach on 22 subjects of the BCI `BI.EEG.2012-GIPSA` database [17]. EEG data were acquired at 16 scalp locations using silver/silver-chloride electrodes and sampled at 128 samples per second. For details on the experimental procedure the reader is referred to [17].

For each subject we have divided in half the available salient and non-salient tangent vectors v_{ml} (section 2.3.1), allowing two data *splits;* one split is used for training and the other for testing, then vice versa. This splitting procedure is repeated 10 times; the accuracy we report is the average of the two training-test procedures repeated 10 times (10 2-fold cross-validation procedure).

Since the classes are unbalanced, we have employed the balanced accuracy as performance index. As ML model we have employed the lasso logistic regression with a fully automated cross-validation procedure to find the best parameters for fitting the data. As parameter P in the group alignment algorithm we have set four; this allows two discriminant dimensions per class, which is a typical value used when spatial filtering is applied to ERP-based BCI data.

We have compared the balanced accuracy obtained with the group alignment method to the one obtained with a classical subject-wise train-test procedure. The training and test data are always *identical* in the comparisons; the only difference between the two is that in the case of group alignment only one ML model is trained using the training data of all subjects, whereas in

the subject-wise case M models are trained, each one using only the training data of the subject under test.

A laptop computer has been instructed to perform all computations using the *Julia* language [18]. The group alignment algorithm is available as function `majd` in the open-source package `Diagonalizations.jl` [19].

The results are shown in Figure 1. The group learning approach proves superior for all 22 subjects. The minimum(maximum) improvement in balanced accuracy is 0.028(0.108). The average(sd) is 0.068(0.023). The 6.8-point average accuracy increase is impressive, first because the subject-wise accuracy is already high for these data, therefore it is difficult to improve further, and second because we did not optimize the pre-processing pipeline.

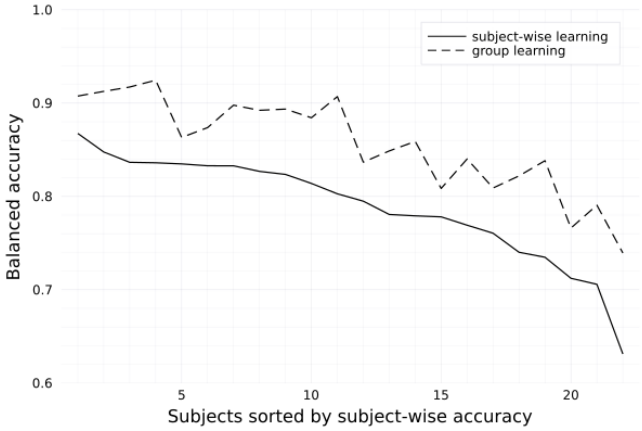

FIGURE 1 – Results. See text for details

4 Discussion and Conclusions

We have presented an original method to jointly align the data of M subjects in the Riemannian tangent space. The same can be done for M sessions. Since our method acts onto a Euclidean space, it can be used to align all kinds of feature vectors, therefore it is general and by no means restricted to the BCI field. Since we have not searched for optimal hyper-parameters in the pre-processing pipeline, there is likely room for improvement.

The accuracy improvements we have observed is highly significant, however, before drawing firm conclusions the performance of the method should be confirmed on other databases and on different BCI paradigms, such as motor imagery and steady-state visually evoked potentials. If the performance is confirmed in future studies, this research may open the way to a renewed ML perspective in the BCI field : *leveraging on the massive amount of available data to build powerful machine learning models*. For instance, group learning may be used to fit ML models requiring very large amount of data, such as deep neuronal networks.

5 Acknowledgments

The research has been partly financed by the French Project ANR n°CE17-0023-01 - HIFI.

6 References

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller et *al*. Brain-computer interfaces for communication and control. *Neurophysiol Clin*, 113(6):767-91, 2002.

[2] L Mayaud, S Cabanilles, A Van Langhenhove, M Congedo at *al*. Brain-computer interface for the communication of acute patients… *Brain-Computer Interfaces*, 3(4) : 197-215, 2017.

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

[4] R. Bhatia. Positive definite matrices. Princeton University Press, 2009.

[5] P. Zanini, M. Congedo, C. Jutten, S. Said, et *al*. Transfer learning: a Riemannian geometry framework with applications to Brain-Computer Interfaces. *IEEE Trans Biomed Eng*, 65(5): 1107-1116.65, 2018.

[6] O. Yair, M. Ben-Chen, R. Talmon, Parallel Transport on the Cone Manifold of SPD Matrices for Domain Adaptation. *IEEE Trans Signal Process*, 67: 1797–1811, 2019.

[7] P. L. C. Rodrigues, C. Jutten, and M. Congedo. Riemannian procrustes analysis : Transfer learning for braincomputer interfaces. *IEEE Trans Biomed Eng*, 66(8) : 2390-2401, 2018.

[8] G. Maman, O. Yair, D. Eytan, R. Talmon. Domain adaptation using Riemannian geometry of SPD matrices. *Proc ICASSP Conf*, 2019.

[9] A. Bleuzé, J. Mattout, M. Congedo. Transfer learning for the Riemannian tangent space: Applications to Brain-Computer Interfaces. Proc ICEET Conf.

[10] M. Congedo M, L. Korczowski, A. Delorme, F. Lopes Da Silva. Spatio-Temporal Common Pattern; a Companion Method for ERP Analysis in the Time Domain. *J Neurosci Methods*, 267: 74-88, 2016.

[11] A. Barachant, M. Congedo, G. Van Veen, C. Jutten. Classification de potentiels évoqués P300 par géométrie riemannienne. *Proc GRETSI Conf*, 2013.

[12] O. Ledoit, M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal*, 88(2):365 – 411, 2004.

[13] J. Vía, M. Anderson, X.-L. Li, T. Adali. Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions, *Proc ICASSP Conf*, 2011

[14] M. Anderson, T. Adali, X.-L. Li. Joint Blind Source Separation With Multivariate Gaussian Model: Algorithms and Performance Analysis, *IEEE Trans Signal Process*, 60(4), 1672-1683, 2012.

[15] M. Congedo, R. Phlypo, J. Chatel-Goldman. Orthogonal and Non-Orthogonal Joint Blind Source Separation in the Least-Squares Sense, *Proc EUSIPCO Conf*, 2012.

[16] M. Congedo. EEG Source Analysis. Habilitation à diriger des recherches, Universite de Grenoble, 2013.

[17] G. van Veen, A. Barachant, A. Andreev, G. Cattan, et *al*. Building Brain Invaders: EEG data of an experimental validation. *Technical Report*, GIPSA-lab, Univesity Grenoble Alpes, CNRS, Grenoble-INP, 2019.

[18] J. Bezanson, A. Edelman, S. Karpinski, V.B. Shah. A fresh approach to numerical computing, *SIAM review*, 59(1) : 65-98, 2017.

[19] github.com/Marco-Congedo/Diagonalizations.jl