SUBSPACE METRICS FOR MULTIVARIATE DICTIONARIES AND APPLICATION TO EEG

Sylvain Chevallier[⋆] *Quentin Barthélemy*[†] *Jamal Atif*[‡]

[⋆] LISV, University of Versailles [†] Mensia Technologies [‡] LRI, University Paris-Sud

ABSTRACT

Overcomplete representations and dictionary learning algorithms are attracting a growing interest in the machine learning community. This paper addresses the emerging problem of comparing multivariate overcomplete dictionaries. Despite a recurrent need to rely on a distance for learning or assessing multivariate overcomplete dictionaries, no metrics in their underlying spaces have yet been proposed. Henceforth we propose to study overcomplete representations from the perspective of matrix manifolds. We consider distances between multivariate dictionaries as distances between their spans which reveal to be elements of a Grassmannian manifold. We introduce set-metrics defined on Grassmannian spaces and study their properties both theoretically and numerically. Thanks to the introduced metrics, experimental convergences of dictionary learning algorithms are assessed on synthetic datasets. Set-metrics are embedded in a clustering algorithm for a qualitative analysis of real EEG signals for Brain-Computer Interfaces (BCI). The obtained clusters of subjects are associated with subject performances. This is a major methodological advance to understand the BCI-inefficiency phenomenon and to predict the ability of a user to interact with a BCI.

Index Terms— Dictionary Learning, Metrics, Frames, Grassmannian Manifolds, Multivariate Dataset

1. INTRODUCTION

Dictionary learning approaches and sparse approximations attracted a lot of attention in several application fields, achieving often state of the art results. Within this context, the contribution of this paper is of double nature. The first one is to be found on the theoretical machine learning side. We introduce metrics in the space of multivariate dictionaries; a topic that has not been tackled yet despite its importance. The second contribution is applicative. Indeed, the introduced metrics are embedded in a clustering algorithm to tackle the very challenging problem of Brain-Compute Interfaces.

Despite the profusion of research papers dealing with overcomplete representations, aside from some noticeable exceptions [1, 2], few results have been reported on how the constructed representations should be compared. Thus, to qualitatively assess a specific dictionary learning algorithm, one has to indirectly evaluate it through a benchmark based on a task performance [3, 4, 5]. Meanwhile, one can find in the literature some hints for dictionaries comparison with the aim of learning assessment [6, 7, 2, 8], but they fall short to define a true metric. However, a related topic has been studied in non-harmonic analysis. In [1], the question of comparing frames is addressed by considering a mapping from the frame space to a continuous functional space. The constructed functions allow then for the definition of an equivalence class, a partial order and a distance. Nonetheless this distance is not invariant to linear transforms, a desirable property in several application fields. Furthermore, its extension to the multivariate case is not straightforward since one should define a new mapping from the space of multivariate frames to an unknown continuous functional space.

In this paper, we introduce metrics exhibiting strong properties, among them the invariance to linear transforms. We propose to study overcomplete representations from the perspective of matrix manifolds. The proposed metrics are built in two stages: first we consider distance between two multivariate atoms as a distance between their spans, which reveal to be elements of a Grassmannian manifold. Then, the collection of spans are then compared by considering transportation distances, e.g. Wasserstein distance.

From the BCI standpoint, one challenging problem is the inter-individual variability. A large proportion of BCI users ($\approx$ 15 to 30%) leads to very poor results [9, 10] even with state of the art algorithms. This phenomenon is known as "BCI-inefficiency" or "BCI-illiteracy". A major difficulty of EEG signal processing is that many artifacts and noise sources (electrical, muscular, etc) corrupt the signal, lowering the SNR of brain signals, while several brain sources are continuously active and mixed in the recording electrodes. The linear transform invariance property of the proposed metric is highly desirable: the metric is insensitive to changes in the electrode positions. We aim at tackling this problem by embedding the proposed metrics in a hierarchical clustering algorithm. The obtained clusters of subjects are then associated with the user's ability to interact with the BCI.

The paper is organized as follows. Section 2 introduces some formal definitions. In Section 3 we define set-metrics on dictionaries. Section 4 provides experimental validations. On synthetic datasets, the convergence of dictionary learning algorithm (DLA) is shown with set-metrics. On BCI Competition datasets, set-metrics are embedded in a clustering algorithm to characterize BCI-inefficient subjects. Section 5 concludes this paper and points out some future research directions.

2. PRELIMINARIES

We consider an n-dimensional real vector space $\mathcal{V}$ (Euclidean n-space). The vectors of $\mathcal{V}$ will be denoted by u, w, the matrices by U, W, and the transpose as $(\cdot)^T$. The inner product induces the ℓ_2 norm $\|u\|_2^2 = \langle u, u\rangle$. The pseudo-norm $\|u\|_0$ is defined as the number of nonzero elements in the vector u. The Frobenius norm is defined as $\|U\|_F^2 = \mathrm{trace}(U^T U)$ and its associated inner product as $\langle U, W\rangle_F = \mathrm{trace}(W^T U)$. Elements U, W have respective spans: $\mathrm{span}(U) = \mathcal{U}$ and $\mathrm{span}(W) = \mathcal{W}$, with $\dim(\mathcal{U}) = \dim(\mathcal{W}) = \varrho$. The indexed families of matrices will be denoted as $\mathbf{U} = \{U_i\}_{i\in I}$ and $\mathbf{W} = \{W_j\}_{j\in J}$ respectively. Indexed families of subspaces will be denoted by $\mathbb{U} = \{\mathcal{U}_i\}_{i\in I}$ and $\mathbb{W} = \{\mathcal{W}_j\}_{j\in J}$ respectively.

S. Chevallier is supported by the Cerebraptic project of EADS Foundation.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.1. Dictionary learning problem

In its classical formulation, the dictionary learning problem aims at capturing most of the energy of a set of training signals $Y = [y_1, \ldots, y_q]$ with $y_j \in \mathbb{R}^n$ and representing it through a collection $U = [u_1, \ldots, u_m]$ in $\mathbb{R}^{n\times m}$ thanks to a set of sparse coefficients $A = [a_1, \ldots, a_q]$ in $\mathbb{R}^{m\times q}$. This collection U, which is redundant ($m \gg n$), is called overcomplete dictionary. The admissible set of dictionaries is convex and is defined as $\mathcal{D}_U = \{U \in \mathbb{R}^{n\times m} : \|u_i\|_2 \leqslant 1,\ i = 1, \ldots, m\}$. Formally, the dictionary learning problem writes as:

$$\min_{U\in\mathcal{D}_U, A\in\mathbb{R}^{m\times q}} \|Y - UA\|_F^2 \quad \text{s.t. } \|a_j\|_0 \leqslant K,\ j = 1, \ldots, q\,. \quad (1)$$

This problem is tackled by dictionary learning algorithms (DLAs), in which energy representative patterns of the dataset are iteratively selected by a sparse approximation step, and then updated by a dictionary update step (see for instance [6, 3, 11, 12]).

2.2. Multivariate dictionary learning problem

Several DLA approaches have been proposed to handle multivariate signals $\mathbf{Y} = [Y_1, \ldots, Y_q]$, with $Y_i \in \mathbb{R}^{n\times\rho}$, the additional dimension ρ being either supported by the coefficients (multichannel approach, [13, 14, 15]) or by the dictionary atoms (multivariate approach, [16]).

The multivariate reformulation of the dictionary learning problem, called M-DLA hereafter, allows to preserve the input space geometry by considering a multivariate dictionary $\mathbf{U}$ as a collection of m multivariate atoms U_i. The considered convex set of dictionaries is defined as $\mathcal{D}_M = \left\{\{U_i\}_{i=1}^m \in \mathbb{R}^{n\times\rho} : \|U_i\|_F \leqslant 1,\ i = 1, \ldots, m\right\}$. The dictionary learning problem thus writes:

$$\min_{\mathbf{U}\in\mathcal{D}_M} \sum_{j=1}^{q} \min_{a_j\in\mathbb{R}^m} \left\| Y_j - \sum_{i=1}^{m} a_{i,j}\, U_i \right\|_F^2 \quad \text{s.t. } \|a_j\|_0 \leqslant K, j = 1, \ldots, q. \quad (2)$$

where $a_{i,j} \in \mathbb{R}$, $a_j \in \mathbb{R}^m$ are the coding coefficients. The sparse decomposition is achieved using a greedy approach [17, 16]. Remark that after vectorization, this model is computationally equivalent to the univariate model of Equation (1), under the hypothesis that the ρ components are independent. The multivariate approach allows the dictionary learning algorithm to take into account existing interactions between components, such as rotations or unconstrained linear transformations (case studied in this paper).

2.3. Grassmannian manifolds and their metrics

The Grassmannian $\mathbf{Gr}(\varrho, n)$ is the set of all ϱ-dimensional linear subspaces of $\mathcal{V}$. The notion of principal angles is central when characterizing the distance between subspaces and hence for metrics on the Grassmannians. Principal angles $0 \leqslant \theta_1 \leqslant \ldots \leqslant \theta_\varrho \leqslant \frac{\pi}{2}$ between two subspaces $\mathcal{U}$ and $\mathcal{W}$ rely on the singular value decomposition of the bases $A = \{a_i\}_{i\in I}$ spanning $\mathcal{U}$ and $B = \{b_i\}_{i\in J}$ spanning $\mathcal{W}$, with I and J two indexing sets:

$$A^T B = \underline{Y}\,\Sigma\underline{Z}^T = \underline{Y}\,(\cos\boldsymbol{\theta})\underline{Z}^T\,, \quad (3)$$

where $\underline{Y} = [y_1, \ldots, y_\varrho]$ and $\underline{Z} = [z_1, \ldots, z_\varrho]$ are orthonormal bases. We denote by $\boldsymbol{\theta}$ the ϱ-vector formed by the principal angles θ_k, $k = 1, \ldots, \varrho$. Here $\cos\boldsymbol{\theta}$ is the diagonal matrix formed by $\cos\theta_1, \ldots, \cos\theta_\varrho$ along the diagonal. It is also known as principal correlations or canonical correlations [18].

Let $\mathcal{U}, \mathcal{W}$ be two elements of a Grassmannian manifold $\mathbf{Gr}(\varrho, n)$ and let $\{\theta_1, \ldots, \theta_\varrho\}$ be their associated principal angles. The chordal distance is probably the most known Grassmannian metric [19, 20]. A detailed and complete study of other Grassmannian metrics could be found in [21] or in [22]. The chordal distance is defined as:

$$d_c(\mathcal{U}, \mathcal{W}) = \|\sin\boldsymbol{\theta}\|_2 = \left(\sum_{k=1}^{\varrho} \sin^2\theta_k\right)^{\frac{1}{2}} = \left(\varrho - \|\underline{U}^T\underline{W}\|_F^2\right)^{\frac{1}{2}}, \quad (4)$$

where $\underline{U}$ and $\underline{W}$ are the orthonormal bases for $\mathcal{U}$ and $\mathcal{W}$, that is $\underline{U}^T\underline{U} = I_\varrho$ and $\mathrm{span}\,(\underline{U}) = \mathcal{U}$.

3. METRICS FOR MULTIVARIATE DICTIONARIES

In this section, we will exploit the chordal metric described in the previous paragraph to act on subsets of a Grassmannian manifold as a ground distance. This ground distance has a key role for the definition of a metric between sets of points in a Grassmannian space.

The Grassmannian manifold $\mathbf{Gr}(\varrho, n)$ together with a distance d, such as the chordal distance, defines a metric space, or pseudo-metric space depending on the properties of the underlying distance. We will denote it in the sequel as $(\mathbb{G}, d)$, and when there is no confusion as $\mathbb{G}$. A result from [23] states that $\mathbf{Gr}(\varrho, n)$ is a Hausdorff, compact, connected smooth manifold of dimension $\varrho(n - \varrho)$. This result is of prime importance since one can define a Borel measure, denoted π, on Grassmannian spaces and consequently a transportation metric.

3.1. Wasserstein distance

Let us denote by $\mathcal{C}(\mathbb{G})$ any collection on $\mathbb{G}$ and by $\mathbb{G}_W = \{(\mathcal{U}, \pi_\mathcal{U}) : \mathcal{U} \in \mathbb{G}\}$ where $\pi_\mathcal{U}$ is a Borel measure. A measure π on the product space $\mathbb{U}\times\mathbb{W}$, with $\mathbb{U}, \mathbb{W} \in \mathbb{G}_W$, is a coupling of $\pi_\mathbb{U}$ and $\pi_\mathbb{W}$ if:

$$\pi(\bar{\mathbb{U}}\times\mathbb{W}) = \pi_\mathbb{U}(\bar{\mathbb{U}}), \quad \pi(\mathbb{U}\times\bar{\mathbb{W}}) = \pi_\mathbb{W}(\bar{\mathbb{W}}) \quad (5)$$

for all Borel sets $\bar{\mathbb{U}} \subset \mathbb{U}$, $\bar{\mathbb{W}} \subset \mathbb{W}$. We denote by $\mathcal{M}(\pi_\mathbb{U}, \pi_\mathbb{W})$ the set of all couplings of $\pi_\mathbb{U}$ and $\pi_\mathbb{W}$.

We could then define the Wasserstein distance, given $p \geqslant 1$, $\mathbb{U}$, $\mathbb{W} \in \mathbb{G}_W$ and a coupling π, as:

$$d_W^p(\mathbb{U}, \mathbb{W}) = \inf_{\pi\in\mathcal{M}(\pi_\mathbb{U}, \pi_\mathbb{W})} \left(\int_{\mathbb{U}\times\mathbb{W}} d(\mathcal{U}, \mathcal{W})^p d\pi(\mathcal{U}, \mathcal{W})\right)^{\frac{1}{p}} \quad (6)$$

This distance is also called the Wasserstein-Kantorovich-Rubinstein [24, Chap. 6].

3.2. Set-Metrics for dictionaries

Metric in Equation (6) is defined on Grassmannian spaces. Then $(\mathbb{G}, d)$ is a separable metric space allowing to compute a distance between the collection of subspaces spanned by a dictionary and the collection of subspaces spanned by another. The distance between two dictionaries in the dictionary space, that is $\mathcal{V}$, is defined as follows.

Let $\mathbf{U} = \{U_i\}_{i\in I}$ and $\mathbf{W} = \{W_j\}_{j\in J}$ be two dictionaries of the $n \times \rho$ vector space $\mathcal{V}$. We define a distance between these two dictionaries as:

$$d_D(\mathbf{U}, \mathbf{W}) = d_{W,p}(\mathbb{U}, \mathbb{W})\,, \quad (7)$$

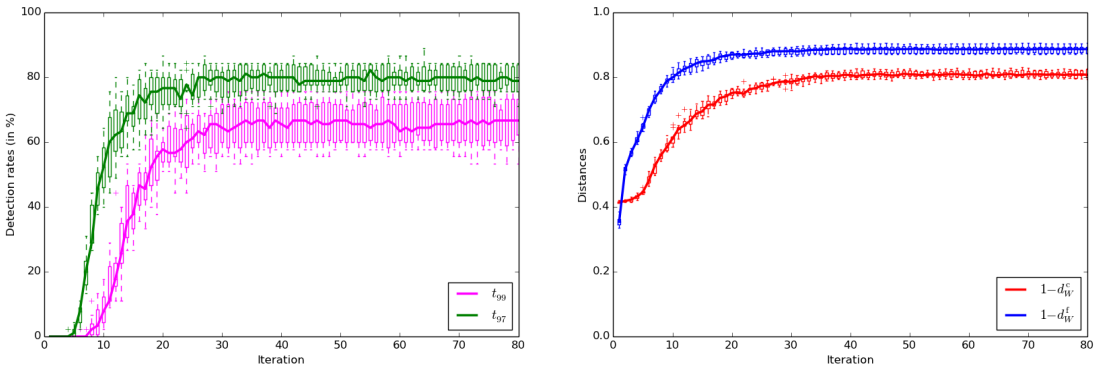

Fig. 1. Left: Detection rates with a threshold of 0.99 and 0.97 as a function of the learning iteration. Right: Wasserstein set-metrics for M-DLA using chordal ($1 - d_{\mathrm{W}}^{\mathrm{c}}$) and Frobenius-based distance ($1 - d_{\mathrm{W}}^{\mathrm{f}}$).

where $\mathbb{U} = \{\mathcal{U}_i : \mathcal{U}_i = \mathrm{span}(U_i), i \in I\}$ and $\mathbb{W} = \{\mathcal{W}_j : \mathrm{span}(W_j), j \in J\}$. In the following, we will denote $d_{W,p}$ as d_W to simplify the notation.

From this definition, we can state the following: Let $\mathcal{C}(\mathcal{V})$ be any collection on $\mathcal{V}$. Then the following holds:

- d_D is pseudo-metric and hence $(\mathcal{C}(\mathcal{V}), d_D)$ is pseudo-metric space,
- d_D is invariant by linear combinations.

The proof is a direct consequence of Equation (7) since d_D is defined as a distance between subspaces.

The dictionary distance d_D is defined as the distance between their subspaces, that is d_D is acting in the dictionary space $\mathcal{V}$ whereas the distance d_W is acting in the Grassmannian space $\mathbb{G}$. As an element $\mathcal{U}$ of $\mathbb{G}$ is a subspace, there exists an infinite number of elements U in $\mathcal{V}$ spanning $\mathcal{U}$. Thus two distinct dictionaries $\mathbf{U}_1 \neq \mathbf{U}_2$, that is two collections $\{U_i^1\}_{i\in I}$ and $\{U_i^2\}_{i\in I}$ of elements in $\mathcal{V}$, could span the same collection of subspaces $\mathbb{U} = \{\mathcal{U}_i\}_{i\in I}$ in $\mathbb{G}$. In other words, a distance $d_D(\mathbf{U}_1, \mathbf{U}_2) = d_W(\mathbb{U}, \mathbb{U}) = 0$ could exist for two separate dictionaries $\mathbf{U}_1 \neq \mathbf{U}_2$. As the separability axiom does not hold, d_D is a pseudo-metric and the separability axiom is relaxed to the identity axiom: $d(x, x) = 0, \forall x \in X$.

4. EXPERIMENTS

This section demonstrates that set-metrics are able to capture the convergence of DLA on synthetic dataset. On real EEG-based dataset, from BCI Competition IV [25], set-metrics are embedded in a hierarchical clustering algorithm to investigate the BCI-inefficiency phenomenon.

4.1. Convergence evaluation in a dictionary recovering task

This section is devoted to demonstrate why relying on metrics allows to improve the assessment of dictionary learning algorithms. More precisely, a set of experiments is conducted to reproduce state-of-the-art results on synthetic datasets and to show how the different proposed metrics behave compared to the commonly used indicators.

Dictionaries criteria and metrics: A first measure to compare dictionaries is the known as *detection rate*. Given two dictionaries $\mathbf{U}$ and $\hat{\mathbf{U}}$, corresponding respectively to a collection of m atoms U_i and m atoms $\hat{U}_i$, a common methodology [26] is to match an atom $\hat{U}_i$ w.r.t. its corresponding atom U_i if their correlation value ν_i is above a chosen threshold s. The correlation is expressed as $\nu_i = |\langle U_i, \hat{U}_i \rangle| \geqslant s$. The detection rate is defined as the percentage of the U_i atoms in $\mathbf{U}$ matched with atoms of $\hat{\mathbf{U}}$.

The set-metric defined in Equation (7) provides a principled way of comparing dictionaries. In the following experiments, the Wasserstein is parametrized with $p = 1$ and the measures are uniform on the whole support, see Equation (6), it is also known as Earth Mover's distance or Mallows distance and many efficient implementations are available [27, 28]. The Wasserstein distance is applied with two different ground distances. The first one relies on the chordal distance, described in Equation (4), and is denoted $d_{\mathrm{W}}^{\mathrm{c}}$. The second one, denoted $d_{\mathrm{W}}^{\mathrm{f}}$, relies on a Frobenius distance and is defined as $(d^{\mathrm{f}}(U_i, \hat{U}_j))^2 = \| \hat{U}_j - U_i \|_F^2 = 2(1 - \langle U_i, \hat{U}_j \rangle)$, assuming that $\|U_i\|_F = \|\hat{U}_j\|_F = 1$. The distance d^{f} is related to the detection rate ν_i, but without the sign invariance: $\hat{U}_j$ is not considered recovered if it is close to $-U_i$. This distance is not invariant to linear transforms.

Experimental protocol: A dictionary $\mathbf{U}$ of $m = 135$ normalized multivariate atoms with $\rho = 10$ is created from white uniform noise. The atom length is $n = 20$ samples. A training dataset $\mathbf{Y}_{\mathrm{train}}$ is generated by combining atoms of $\mathbf{U}$. $\mathbf{Y}_{\mathrm{train}}$ contains $q = 2000$ training signals of length n. Each training signal is generated as the sum of three atoms, the coefficients and the atom indices being randomly drawn. A dictionary $\hat{\mathbf{U}}$ with at least m atoms is built from $\mathbf{Y}_{\mathrm{train}}$ using M-DLA [16] described in (2). The quality of the DLA is assessed by measuring the proportion of atoms in $\mathbf{U}$ recovered in $\hat{\mathbf{U}}$. Here, $\hat{\mathbf{U}}$ is initialized with random signals from $\mathbf{Y}_{\mathrm{train}}$ and 80 iterations of DLA are performed (sparse approximation and dictionary update).

Results: The results are presented in Fig. 1: the set-metrics and detection rates are computed at each iteration. It appears clearly that the dictionary has almost converged after only 15 to 20 iterations. It is thus interesting to investigate how the detection rates and the set-metrics capture the evolution of the dictionary $\hat{\mathbf{U}}$ during these first iterations. The detection rates completely fail to detect any modification ongoing in $\hat{\mathbf{U}}$ before 5 or 8 iterations. Then, they abruptly increase between iterations 5 and 20. The detection rates are converging toward 67% for $s = 0.99$ and 72% for $s = 0.97$ but display important variations. These variations are a direct consequence of the thresholding occurring in the detection rates.

All the set-metrics start with positive values because they detect

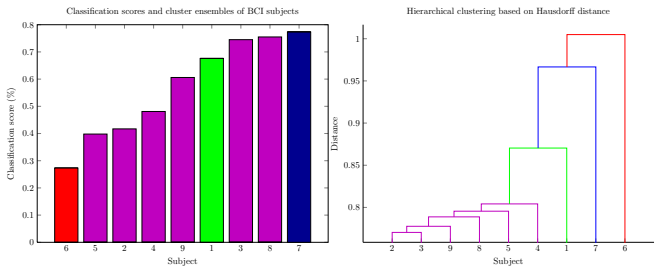

Fig. 2. Left: Performance of the subjects from the BCI Competition IV-2a with state of the art algorithm from [29], color indicates cluster ensembles obtained with consensus clustering on learned dictionaries. Right: Hierarchical clustering on the same dataset.

that $\hat{\mathbf{U}}$ is initialized with training signals. They provide a smoother and more accurate evaluation of the convergence. The $d_{\mathrm{W}}^{\mathrm{f}}$ metric provides the smoother results of all distances and the $d_{\mathrm{W}}^{\mathrm{c}}$ is sensible enough to capture modification of $\hat{\mathbf{U}}$ after the 20^{th} iteration.

Contrary to detection rates based indicators that provide oscillating values, one can note the regularity of the values provided by the set-metrics. This demonstrates their efficiency in the context of dictionary learning.

4.2. Clustering on EEG-based BCI

Experimental protocol: The aim here is to rely on set-metrics embedded in a clustering algorithm to investigate user-specific characteristics in BCI tasks. The individual variability of BCI subjects is still largely unharnessed: from 15 to 30% of BCI users obtain very poor results [9, 10] with the state of the art algorithms. This phenomenon is referred to as "BCI-illiteracy" or "BCI-inefficiency". We claim here that the comparison of subjects through computing the distances between their associated learned dictionaries could help to characterize the BCI-inefficiency.

Multivariate dictionaries are learned with M-DLA [16] using the datasets from BCI Competition IV set 2a [25]. This is a motor imagery experiment, where 9 subjects have been instructed to imagine four tasks (imagination of left hand, right hand, tongue or feet movements) during two sessions. Each session consists of 288 trials (72 trials for each task) and a trial is 3 seconds recording of $\rho = 22$ electrodes sampled at 250 Hz. For a given subject, a dictionary is learned for each task using the first session dataset, with a sparsity parameter $K = 1$. Thus for all nine subjects, $9 \times 4 = 36$ dictionaries are learned.

Relation between subject clusters and BCI-inefficiency: A distance matrix G^t between subjects is computed for each task t using the Wasserstein set-metric based on the chordal distance $d_{\mathrm{W}}^{\mathrm{c}}$. For a given task, the distances $d_{\mathrm{W}}^{\mathrm{c}}(\mathbf{U}_i, \mathbf{U}_j)$ between dictionaries of subjects i and j are converted to Gaussian similarities: $s_{ij}^t = \exp\left(-(d_{\mathrm{W}}^{\mathrm{c}}(\mathbf{U}_i, \mathbf{U}_j))^2/2\right)$. Subject's clusters are gathered using affinity propagation [30] on G^t. The preference value of subjects for affinity propagation is set to the median value of G^t.

The subject's clusters obtained for each task are combined using cluster ensembles techniques [31]. Hence, a partition of the subjects in C clusters is obtained by maximizing the shared mutual information of all tasks. The results show that the subjects are aggregated in $C = 3$ or 4 stable cluster ensembles. This clustering is stable since increasing the C value results in empty clusters. These clusters are represented with different colors on the left-hand side of Fig. 2. The histograms in Fig. 2 report the subject performances based on Filter Bank CSP [29], the state-of-the-art algorithm. The subjects are sorted according to their performance, the best performing subject is on the right-hand side. Using $C = 3$ cluster ensembles, the subject with the highest performance is always alone in a cluster. The same holds for the subject with the worst performance.

To highlight the relation between clusters of subjects, a hierarchical clustering is shown on the right-hand side of Fig. 2. This hierarchy is built using the distance matrix G^t of only one task, the one computed from feet imaginary motion[1]. The obtained dendrogram is shown on the right-hand side of Fig. 2. One can note that subjects 2, 3, 9, 8, 5 and 4 belong to the same cluster and that each one of the subjects 1, 7 and 6 constitute three separate clusters.

On this BCI dataset, most of the subjects share a common profile except for the two extreme cases of the most BCI-inefficient and BCI-efficient subjects. Set-metrics based on multivariate dictionaries offer new opportunities to qualitatively assess datasets used in competitions and challenges. We hope that our approach could help the community to propose more consistent and more complete benchmarks or evaluations.

5. CONCLUSION

This contribution relies on advances from algebraic geometry and matrix manifolds to define suited metrics for multivariate dictionaries. It is the first attempt with respect to this emerging field. The distance between dictionaries is computed as the distance between their subspaces, yielding pseudo-metrics which are invariant to linear transformations, a very desirable property when dealing with multivariate dictionaries.

The interest of the described metrics has been shown through its direct application on two examples: a synthetic dataset and real dataset of EEG-based BCI. The proposed metrics allow to estimate empirically the convergence of a dictionary learning algorithm with a precision outperforming the classical measurements based on detection rates. On the BCI dataset, we have shown how these metrics applied on multivariate dictionaries learned from EEG data can help assessing the "BCI-inefficiency" of subjects. The chordal distance endows the set metric with an invariance to linear transforms, a desirable property for analyzing EEG brain signals as the set metric is thus not affected by variations in electrods positions. Thus, the proposed hierarchical clustering approach allows to gain new neurophysiological insight on the user's ability to interact with BCI systems.

Future work will be devoted to the extensive analysis of BCI datasets, using various clustering approaches based on the introduced set-metrics. From a theoretical perspective, an ongoing work concentrates on the embedding of these set-metrics in dictionary learning algorithms.

6. REFERENCES

[1] R. Balan, "Equivalence Relations and Distances between Hilbert Frames," *Proceedings of the American Mathematical Society*, vol. 127, no. 8, pp. 2353–2366, 1999.

[2] K. Skretting and K. Engan, "Learned dictionaries for sparse image representation: properties and results," in *SPIE Conference*, San Diego, USA, 2011, vol. 8138.

[1]Hierarchical clustering requires a distance matrix. Cluster ensembles using all tasks provide only a partition, thus the hierarchy is built using only one task.

[3] K. Engan, S.O. Aase, and J.H. Husøy, "Multi-frame compression: theory and design," *Signal Processing*, vol. 80, pp. 2121–2140, 2000.

[4] R. Grosse, R. Raina, H. Kwong, and A.Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, USA, 2007, pp. 149–158.

[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE ICCV*, Kyoto, Japan, 2009, pp. 2272–2279.

[6] M. Aharon, M. Elad, and A.M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, pp. 4311–4322, 2006.

[7] A. Aldroubi, "Portraits of Frames," *Proceedings of the American Mathematical Society*, vol. 123, no. 6, pp. 1661–1668, 1995.

[8] D. Vainsencher, S. Mannor, and A.M. Bruckstein, "The sample complexity of dictionary learning," *Journal of Machine Learning Research*, vol. 12, pp. 3259–3281, 2011.

[9] E.M. Hammer, S. Halder, B. Blankertz, C. Sannelli, T. Dickhaus, S. Kleih, K.-R. Müller, and A. Kübler, "Psychological predictors of SMR-BCI performance," *Biological Psychology*, vol. 89, no. 1, pp. 80–86, 2012.

[10] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain Topography*, vol. 23, no. 2, pp. 194–198., 2010.

[11] K. Engan, K. Skretting, and J.H. Husøy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Digital Signal Processing*, vol. 17, pp. 32–49, 2007.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[13] J.A. Tropp, "Algorithms for simultaneous sparse approximation; Part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, 2006.

[14] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multichannel sparse recovery using greedy algorithms," Tech. Rep. PI-1848, IRISA, 2007.

[15] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *Signal Processing*, vol. 91, pp. 1505–1526, 2011.

[16] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J.I. Mars, "Multivariate temporal dictionary learning for EEG," *Journal of Neuroscience Methods*, vol. 215, pp. 19–28, 2013.

[17] Q. Barthélemy, A. Larue, A. Mayoue, D. Mercier, and J.I. Mars, "Shift & 2D rotation invariant sparse coding for multivariate signals," *IEEE Trans. Signal Processing*, vol. 60, pp. 1597–1611, 2012.

[18] G.H. Golub and H. Zha, "The canonical correlations of matrix pairs and their numerical computation," in *Linear Algebra for Signal Processing*, A. Bojanczyk and G. Cybenko, Eds., vol. 69 of *The IMA Volumes in Mathematics and its Applications*, pp. 27–49. Springer, 1995.

[19] G.H. Golub and C.F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

[20] J.H. Conway, R.H. Hardin, and N.J.A. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Experimental Mathematics*, vol. 5, no. 2, pp. 139–159, 1996.

[21] I.S. Dhillon, R.W. Heath Jr., T. Strohmer, and J.A. Tropp, "Constructing Packings in Grassmannian Manifolds via Alternating Projection," *Experimental Mathematics*, vol. 17, no. 1, pp. 9–35, 2008.

[22] A. Edelman, T.A. Arias, and S.T. Smith, "The Geometry of Algorithms with Orthogonality Constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.

[23] J.W. Milnor and J.D. Stasheff, *Characteristic Classes. (AM-76)*, vol. 76, Princeton University Press, 1974.

[24] C. Villani, *Optimal transport: old and new*, vol. 338 of *Grundlehren der mathematischen Wissenschaften*, Springer, 2009.

[25] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K.J. Miller, G. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI Competition IV," *Frontiers in Neuroscience*, vol. 6, no. 55, 2012.

[26] M. Aharon, *Overcomplete Dictionaries for Sparse Representation of Signals*, Ph.D. thesis, Technion - Israel Institute of Technology, 2006.

[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *International Conference on Computer Vision (ICCV)*. 1998, pp. 59–66, IEEE.

[28] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *International Conference on Computer Vision (ICCV)*. 2009, pp. 460–467, IEEE.

[29] K.K. Ang, Z.Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b.," *Frontiers in Neuroscience*, vol. 6, pp. 1–9, 2012.

[30] B.J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[31] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.