ON THE NEED FOR METRICS IN DICTIONARY LEARNING ASSESSMENT

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

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

ABSTRACT

Dictionary-based approaches are the focus of a growing attention in the signal processing community, often achieving state of the art results in several application fields. Albeit their success, the criteria introduced so far for the assessment of their performances suffer from several shortcomings. The scope of this paper is to conduct a thorough analysis of these criteria and to highlight the need for principled criteria, enjoying the properties of metrics. Henceforth we introduce new criteria based on transportation like metrics and discuss their behaviors w.r.t the literature.

***Index Terms*—** Dictionary learning, dictionary recovering, metric, transportation distance, detection rate.

1. INTRODUCTION

Dictionary learning approaches have been the subject of an increasing attention in the last decade, combining latest advances in sparse approximations [1] and overcomplete representations learning [2]. Important results have been obtained on the applicative side [3, 4] e.g. denoising, inpainting, compression or classification. A key point in all these applicative fields is the definition of a suitable assessment criterion. Common practices rely on measures, often task specific [4] suffering from several drawbacks, among which non-convexity, data/task-dependence, parameterization, etc. can be cited. These drawbacks can be overcome when the criteria are derived from a distance enjoying metric properties. Surprisingly enough, few works concerned a principled analysis of the existing criteria and the introduction of metrics aimed at dictionary learning assessment, apart from the noticeable work of [5].

The contributions of this paper are to evaluate existing criteria for dictionary learning assessment and to define a suited metric for dictionaries. The existing criteria and the proposed distance are evaluated on real and synthetic datasets. The results are threefold: (i) existing criteria fail to detect and capture changes during the first iterations of the optimization algorithm whereas the proposed distance does, (ii) the distances measured with the proposed metric have a smaller variance than those obtained by the existing criteria and (iii) the existing criteria completely fail at low SNR while the proposed distance does not.

The lack of true metrics in dictionary learning is often circumvented by assessing dictionaries through specific task performances, such as in [3, 4], using a black box approach. An alternative approach is to rely on the reconstruction error values [4].

A recurrent methodology, introduced in [6, 3], is to generate a dataset from an initial dictionary and to assess the ability of a dictionary learning algorithm to recover this initial dictionary from the generated dataset. The initial and the learned dictionaries are then compared through a threshold-based matching relying on the correlation between dictionary atoms. Several implementations follow from the definition of this criterion, referred to as *detection rate*, and lead to very different results. Furthermore, the detection rate should be carefully parameterized, as the results heavily depend on the chosen threshold. Using this criterion, small variations could result in important changes and large variations could go undetected.

A criterion, introduced in [5], is specifically designed to compare dictionaries. This is the first attempt to define a metric for overcomplete dictionaries with the aim of assessment. In fact, it defines a distance under specific conditions and it turns out to be a ℓ_∞ metric. This metric suffers however from known limitations, such as sensitivity to outliers, which have been intensively investigated in the literature [7].

In this paper, the dictionary learning algorithms and the existing criteria are discussed in Section 2. The existing metric for dictionaries is recalled in Section 3 along with the proposition of a new one and their comparison. The existing criteria and the proposed distance are evaluated through different experiments in Section 4. Section 5 concludes this paper and points out some research directions.

2. DICTIONARY LEARNING AND ASSESSMENT

2.1. Dictionary learning problem

The dictionary learning problem aims at capturing most of the energy of a set of training signals $Y = [y_1, \dots, y_q]$ with $y_j \in \mathbb{R}^n$ and representing it through a collection $\Phi = [\phi_1, \dots, \phi_m]$ in $\mathbb{R}^{n \times m}$ thanks to a set of sparse coefficients $X = [x_1, \dots, x_q]$ in $\mathbb{R}^{m \times q}$. This collection Φ, which

S. Chevallier is supported by the Continental Industrial Chair “High Tech Low Cost”.

is redundant ($m \gg n$), is called overcomplete dictionary. The admissible set of dictionaries is convex and is defined as $\mathcal{D}_\Phi = \{\Phi \in \mathbb{R}^{n\times m} : \|\phi_i\|_2 \leqslant 1,\, i = 1, \dots, m\}$. Classically, the dictionary learning problem can be formulated as:

$$\min_{\Phi\in\mathcal{D}_\Phi, X\in\mathbb{R}^{m\times q}} \|Y - \Phi X\|_F^2 \quad \text{s.t. } \|x_j\|_p \leqslant K,\, j = 1, \dots, q\,, \tag{1}$$

with $\|\cdot\|_F$ the Frobenius norm and $p = 0$ or 1. This nonconvex problem is tackled by dictionary learning algorithms (DLAs), in which energy patterns of the dataset are iteratively selected by a sparse approximation step, and then updated by a dictionary update step (see for instance [3, 4, 8, 9]).

2.2. Criteria for dictionary learning assessment

As the objective function of Eq. (1) is to minimize a reconstruction error, a common assessment on real datasets is to rely directly on the error value [4]. However, this criterion could only offer a comparison of signals approximations with original signals, not a comparison of dictionaries. It is also possible to evaluate the learned dictionary using a black box approach for specific applications (denoising, inpainting, demosaicing, etc.) but this approach suffers from the same drawbacks as the objective function-based comparison.

As said before, the most widely adopted methodology so far has been proposed in [6, 3] and applied over simulated data: it is called *dictionary recovering*. An initial dictionary $\Phi \in \mathbb{R}^{n\times m}$ is randomly generated from a uniform distribution. A dataset of signals Y is created from this reference dictionary, by mixing K random atoms with random coefficients. A dictionary $\hat{\Phi} \in \mathbb{R}^{n\times m'}$ is then learned on this dataset. The dictionary recovering goal is to uncover most of the atoms of Φ from the training set Y. This task is meant to qualitatively evaluate dictionary learning methods. This evaluation relies on a similarity criterion c computed between Φ and $\hat{\Phi}$.

The Frobenius norm $\| \Phi - \hat{\Phi} \|_F^2$ is also not appropriate since the learned atoms are not necessarily in the same order as the original ones. Furthermore, some original atoms can be learned several times while others can be missing.

Another way to compare dictionaries in dictionary recovering task is to rely on a detection rate based on the correlation between the original and the learned atoms, $\langle \phi_i, \hat{\phi}_j \rangle$. An original atom ϕ_i is considered as recovered if there is at most one learned atom $\hat{\phi}_j$ which provides a scalar product-based score s greater than a given threshold t, commonly fixed at 0.99 [3]. This general recovering definition can lead to different implementations which are not necessarily equivalent. Two approaches for computing the detection rate can be found in the literature: in the *pairwise detection rate* [3], each considered atom could be matched with one and only one target atom, while in *precision/recall detection rate* [10], a given atom could be matched with several target atoms.

Concerning the pairwise detection rate, to not be sensitive to atoms permutation, the best practice is to rely on the cross-Gramian matrix $G = \Phi^T \hat{\Phi}$ containing all correlations between atoms $\langle \phi_i, \hat{\phi}_j \rangle$. Summed up in Algorithm 1, the pairwise detection rate c_g is iteratively computed using:

$$s_\mathrm{g}(i,j) = \max_{j=1\dots m'} \max_{i=1\dots m} \left| \langle \phi_i, \hat{\phi}_j \rangle \right| . \tag{2}$$

However, it is important to note that the detection rate c_g can be different w.r.t to the matching choice, as illustrated in Fig. 1 which compares a greedy matching with a different one.

A second approach [10] offers the possibility for an atom to be matched several times in the score computation:

$$s_\mathrm{r}(i) = \max_{j=1\dots m'} \left| \langle \phi_i, \hat{\phi}_j \rangle \right| . \tag{3}$$

The drawback of this score is that it is not symmetric with respect to original and learned dictionaries. Two scores, precision and recall, have been proposed in [10] to overcome this limitation. For each original atom ϕ_i, the recall score s_r defined in Eq. (3) gives the maximal scalar product with the learned dictionary, and for each learned atom $\hat{\phi}_j$, the precision score s_p gives the maximal scalar product with the original dictionary:

$$s_\mathrm{p}(j) = \max_{i=1\dots m} \left| \langle \phi_i, \hat{\phi}_j \rangle \right| . \tag{4}$$

To obtain the final criterion, associated detection rates c_r and c_p can be computed based on the thresholding of s_r and s_p.

```
Data: Cross-Gramian matrix $G = \Phi^T \hat{\Phi}$
repeat $m$ times
    find line $i$ and column $j$ of $s_\mathrm{g}(i,j)$ from $|G|$
    if $s_\mathrm{g}(i,j) \geqslant t$ then $c_\mathrm{g} \leftarrow c_\mathrm{g} + 1$
    set to 0 the line $i$ and the column $j$ of $G$
end
Result: Detection rate $c_\mathrm{g} \leftarrow \frac{c_\mathrm{g}}{m} \times 100$
```

Algorithm 1: Greedy pairwise detection rate, where $|G|$ is the $m \times m'$ matrix of the absolute values of G elements. Indices i and j refer respectively to matrix line and column.

2.3. Limitations of existing criteria

The criteria introduced in the previous section suffer from several drawbacks. The objective function is non convex: two different dictionaries can obtain similar scores. The complexity of this criterion directly depends on the number of training signals q, which requires to solve the minimization problem of Eq. (1). As this is the most computationally demanding part of the dictionary learning algorithm and as the complexity is a function of the dataset size q, this criterion becomes not efficient. Since the objective function values strongly depend on the dataset, one cannot rely on such criterion to define a generic stopping criterion.

	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
ϕ_1	0.998	0.996	–
ϕ_2	0.997	–	–
ϕ_3	–	–	0.995

Fig. 1: Matrix $|G|$ of scalar products of a toy example, with - for any value lower than 0.99. Greedy method matches atoms ϕ_1 and $\hat{\phi}_1$, and then ϕ_3 and $\hat{\phi}_3$, giving a detection rate of 66 %. The matching ϕ_1 and $\hat{\phi}_2$, ϕ_2 and $\hat{\phi}_1$, and ϕ_3 and $\hat{\phi}_3$ provides a rate of 100 %.

The main issue with the detection rate criterion, independently of the chosen implementation, is the threshold effect. Due to the binary nature of the thresholding, an atom just under the threshold will not be counted. The detection rate values also strongly depend on the chosen threshold which should be carefully parameterized for each problem/dataset.

The study of these existing criteria suggests that they are mostly a default choice because of the lack of criteria with strong properties, like smoothness, differentiability, sign and permutation invariance or convexity. This calls for the definition of real metric for overcomplete representations.

3. METRICS FOR DICTIONARIES

Formally, a metric is a function $d : U \times U \to \mathbb{R}^+ = [0, \infty)$ defined for an arbitrary non empty set U and verifies the following three axioms: $d(x, y) = 0$ *iff* $x = y, \forall x, y \in X$ (*A1*: Separability), $d(x, y) = d(y, x)$, $\forall x, y \in X$ (*A2*: Symmetry) and $d(x, y) \leqslant d(x, z) + d(z, y)$, for all $x, y, z \in X$ (*A3*: Triangle inequality).

3.1. Existing metric

To the best of our knowledge, the unique work proposing a metric between dictionaries is the one by Skretting *et al.* [5]. Combining the recall and precision scores, s_r and s_p, they introduced the following distance:

$$d_\beta = \frac{1}{m + m'} \Big(\sum_{j=1}^{m'} \mathrm{acos}\ (s_\mathrm{p}(j)) + \sum_{i=1}^{m} \mathrm{acos}\ (s_\mathrm{r}(i)) \Big).$$

This distance could take into account dictionaries of different size, such as $m \neq m'$, but then the triangular inequality does not hold and d_β fails to be a true metric in that case.

3.2. New metric

The main contribution of this paper is to introduce a new metric enjoying nice desired properties. We proceed as follows: (i) first, a suitable metric is defined between atoms, then (ii) a set-metric (e.g. Hausdorff or Wasserstein metrics) is defined based on this atom to atom "ground" metric.

For dictionaries, the following Euclidean-based distance could act as a ground metric:

$$d_\mathrm{E}(\phi_i, \hat{\phi}_j)^2 = 2 \times \big(1 - |\langle \phi_i, \hat{\phi}_j \rangle|\big),$$

assuming that $\|\phi_i\|_2 = \|\hat{\phi}_j\|_2 = 1$. The distance d_E is related to the scalar product based detection rate, including the sign invariance: $\hat{\phi}_j$ is considered recovered if it is close to $-\phi_i$ or ϕ_i. The absolute value allows to be independent from the sign of the scalar product, since there is no positivity constraints on atoms in the classical dictionary problem.

Once the ground distance is defined, a first possibility to establish a metric between collections of atoms is to rely on the Hausdorff distance. For two non-empty finite[1] dictionaries $\Phi = \{\phi_i\}_{i=1}^m$ and $\Psi = \{\psi_j\}_{j=1}^{m'}$, we define:

$$d_\mathrm{H}(\Phi, \Psi) = \max \Big(\max_{\phi \in \Phi} \min_{\psi \in \Psi} d_\mathrm{E}(\phi, \psi), \max_{\psi \in \Psi} \min_{\phi \in \Phi} d_\mathrm{E}(\phi, \psi) \Big),$$

as a metric for dictionaries computed from the d_E Euclidean-related ground distance. The Hausdorff distance is a ℓ_∞ metric widely used in image processing: it is known to be sensitive to variations on the elements lying on the hull of the considered collections, such as outliers, and to be insensitive to variations of elements inside it.

A more appropriate metric for sparse representations is the Wasserstein distance, which could be defined as:

$$d_\mathrm{W}(\Phi, \Psi) = \min_T \Big(\sum_{i=1}^{m} \sum_{j=1}^{m'} T_{i,j}\, d_\mathrm{E}(\phi_i, \psi_j)^p \Big)^{1/p}\ ,\ p \geqslant 1,$$

where T is the $m \times m'$ transportation matrix. All the entries of T are non-negative: $T_{i,j} \geqslant 0$, for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant m'$. It also verifies the following properties: $\sum_j T_{i,j} = \frac{1}{m}$ for $1 \leqslant i \leqslant m$ and $\sum_i T_{i,j} = \frac{1}{m'}$ for $1 \leqslant j \leqslant m'$. With $p = 1$, this metric is closely related to the Earth Mover's Distance (EMD) or Mallows distance and many efficient implementations are available (e.g. [11]). We are applying this metric with d_E acting as ground distance and the measure is uniform on the support, that is T is $m \times m'$ matrix with all entries being equal to $\frac{1}{m \times m'}$.

The proposed criteria d_H and d_W are in fact pseudo-metrics, i.e. the separability axiom (A1) is relaxed to the identity axiom: $d(x, x) = 0, \forall x \in X$. This is a direct consequence of the choice of d_E as a ground distance: the sign invariance property of d_E does not allow to separate x from $-x$. This is not an issue as the sign invariance is a desired property for the pseudo-metric. One can note that d_β, for similar reasons, is also a pseudo-metric.

3.3. Comparison and discussion

The d_β distance is a pseudo-metric, under the hypothesis that all the dictionaries have the same number of atoms. Due to

[1] Without loss of generality, we consider only finite dictionaries here, but this metric extends to infinite case as well.

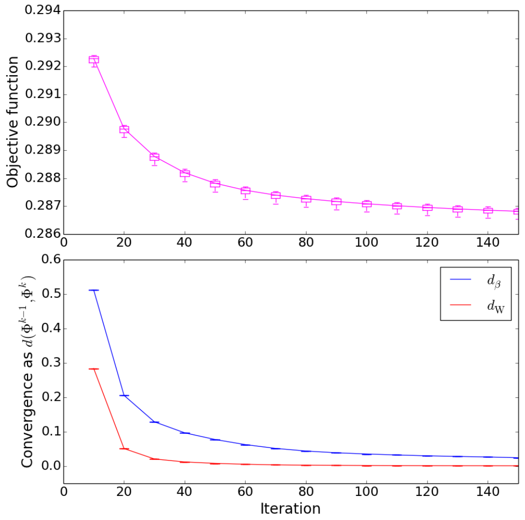

Fig. 2: Convergence analysis on image reconstruction task, objective function (top), d_β and d_{W} distances (bottom).

the $\max$ operator, this metric is sensitive to outliers and to changes occurring on the most “extreme” points of the considered collections. Consequently, it is also not suited to detect changes affecting points near to the “center” of the collections. The distance d_{W}, based on Wasserstein metric, is computed for all correspondences between dictionary elements (thanks to the transportation matrix T) and thereby not subject to those limitations.

A major advantage of the Wasserstein metric is that it could be easily modified to embed specific priors by enforcing a probability measure over the atoms rather than using a uniform measure. Indeed, in its original formulation, the Wasserstein metric is distance between probability density functions. In our case, we consider its simple version where the pdfs are Dirac measures. Several probability priors can be defined. For instance, it is possible in our case to enforce information about the frequency of occurrence of each atom, or their contribution in the signal energy, by modifying the entries of the transportation matrix T leading to metrics aiming at capturing specific behaviors of the dictionaries. Moreover, the d_{W} distance relies on a ground distance which could be adapted to include the desired invariances (sign invariance is a simple example).

Another point of comparison between d_β and d_{W} concerns the extension of dictionary learning to the multivariate model. In this context, the signal is a matrix, thus the atoms are matrices and the dictionaries are tensors [9], which is different from the multichannel model [12]. The natural extension of the d_β distance is to extend the score s using the Frobenius norm. Nonetheless the d_β distance will suffer from the same limitations explained previously. For the distance d_{W}, the only required change concerns the ground distance. An immediate choice is to rely on the Frobenius distance. A good alternative is to compare matrix atoms using principal angles and thus use, for example, the chordal distance [13] as a ground distance (with $p = 2$). Indeed, the resulting distance will be still smooth and differentiable in addition to be invariant to linear transforms, due to the chordal distance which is a distance between subspaces [13].

4. EXPERIMENTS

Experiments were conducted to reproduce state-of-the-art results on real and synthetic datasets and to show how the proposed set-metrics behave compared to the common criteria.

4.1. Convergence evaluation on real dataset

This first experiment is dedicated to show how the set-metrics are applied for an empirical convergence analysis, in the context of an image reconstruction task. We rely on the online dictionary learning (ODL) algorithm [4] to learn a dictionary Φ with $m = 100$ atoms from 8×8 grey level patches sampled from a given set of images (same as those used in [4]). Distances d_β and d_{W} are computed every 10 iterations for 150 total iterations. The experiment is repeated 15 times due to the randomization of examples during ODL process. On Fig. 2, plain lines indicate median values, boxes indicate the quartile and whiskers show extreme values.

Top part of Fig. 2 displays the objective value of Eq. (1) computed on the whole set of patches. Indeed, for a large dataset or online setup, the problem is intractable and one should measure the objective value on a subset of the dataset, degrading the obtained results. Fig. 2 shows the d_β distance and the proposed distance d_{W} on the bottom part. As the ideal dictionary is not known, the displayed values of Fig. 2 are simply evaluated by computing the distance between the learned dictionary at two successive iterations, i.e. $d(\Phi^{k-1}, \Phi^k)$ where k is the current iteration.

The objective value is dependent on the dataset. In our setting it takes values between 0.286 and 0.293, and the variance is higher than the one observed for d_β and d_{W}. Both d_β and d_{W} capture small variations of the dictionary updates and are suitable to evaluate the convergence of DLA. Their values are bounded between 0 and 1. This suggests that d_{W} and d_{H} are good candidates as stopping criterion for the DLA.

4.2. Dictionary recovering

A dictionary Φ of $m = 50$ normalized atoms of $n = 20$ samples is created from white uniform noise. A training dataset Y is generated by combining atoms of Φ. Y contains $q = 1500$ training signals of length n. Each training signal is generated as the sum of $K = 3$ atoms, the coefficients and the atom indices being drawn from a uniform distribution. Gaussian noise is added to training signals, such that SNR has a ratio

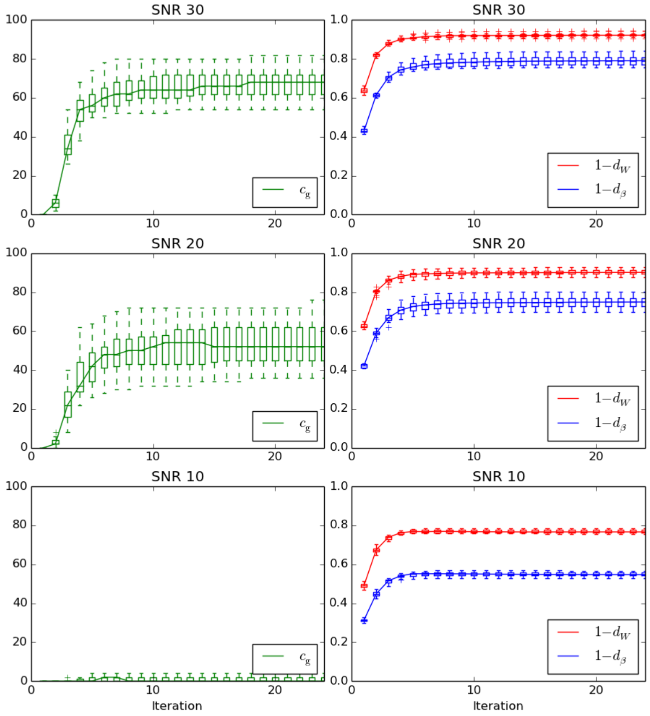

Fig. 3: Detection rate c_{g} for $t = 0.99$ (left) and distances $1 - d_\beta$ and $1 - d_{\mathrm{W}}$ (right) as a function of learning iterations, for a SNR of 30 (top row), 20 (middle) and 10 (bottom).

of 30, 20 or 10, and the experiments are repeated 15 times. A dictionary $\hat{\Phi}$ with at least m atoms is learned from Y using online dictionary learning (ODL). Here, $\hat{\Phi}$ is initialized with random signals from Y and 25 iterations of ODL are performed. The quality of the ODL convergence is assessed by measuring the proportion of atoms in Φ recovered in $\hat{\Phi}$.

The results are presented in Fig. 3: the detection rate c_{g} and the values $1 - d_{\mathrm{W}}$, $1 - d_\beta$ are computed at each iteration. It appears clearly that the dictionary has almost converged after only few iterations. It is thus interesting to investigate how c_{g} and the proposed metric capture the evolution of the dictionary $\hat{\Phi}$ during these first iterations. c_{g} fails to detect any modification ongoing on $\hat{\Phi}$ in the first iterations. Then, after an abrupt increase, the c_{g} shows important variations, due to the threshold effect. In the case of SNR = 10, c_{g} is not able to detect any atoms being recovered. The precision and recall results are almost identical to those of c_{g} and suffer from the same limitations. The pseudo-metrics d_β and d_{W} start with positive values since $\hat{\Phi}$ is initialized with the training signals. This fact is completely neglected by the detection rate-based criterion. d_{W} provides a more accurate evaluation of the convergence, as it is demonstrated by the quartile measures which is smaller than the d_β one. Furthermore, the distance values obtained with d_{W} are not affected by the SNR level while the distance d_β is sensitive to this effect.

5. CONCLUSION

The contributions of this paper are the review of existing criteria for dictionary assessment and the proposition of a new one. We demonstrate the need for criteria with (pseudo-)metrics properties. An experimental setting reveals that our new criterion outperforms prior state of the art in terms of outliers robustness, sensitivity to small variations and efficiency in low SNR. It is a useful tool to compare different dictionary learning algorithms, measuring their recovering ability. Future work will be dedicated to the exploitation of the introduced criterion in a dictionary learning process and in classification tasks, e.g. driver behaviors identification in the context of smart cars.

REFERENCES

[1] S. Mallat, *A Wavelet Tour of signal processing*, 3rd edition, New-York : Academic Press, 2009.

[2] R. Gribonval and K. Schnass, "Dictionary Identification–Sparse Matrix-Factorization via ℓ1-Minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[3] 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.

[4] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J Mach Learn Res*, vol. 11, pp. 19–60, 2010.

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

[6] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, pp. 349–396, 2003.

[7] W. Rucklidge, "Efficiently locating objects using the Hausdorff distance," *IJCV*, vol. 24, pp. 251–270, 1997.

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

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

[10] S. Lesage, *Apprentissage de dictionnaires structurés pour la modélisation parcimonieuse des signaux multicanaux*, Ph.D. thesis, Université de Rennes, 2007.

[11] O. Pele and M. Werman, "Fast and robust Earth Mover's Distances," in *IEEE ICCV*, Kyoto, Japan, 2009, pp. 460–467.

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

[13] S. Chevallier, Q. Barthélemy, and J. Atif, "Subspace metrics for multivariate dictionaries and application to EEG," in *IEEE ICASSP*, Firenze, Italy, 2014.