Learning redundant dictionaries with translation invariance property: the MoTIF algorithm

Philippe Jost, Pierre Vandergheynst
Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Institute
Email: {philippe.jost,pierre.vandergheynst}@epfl.ch
Tel: ++41 +21 693 47 54
Sylvain Lesage, Rémi Gribonval
IRISA (CNRS & INRIA), campus de Beaulieu, 35042 RENNES cedex, FRANCE
Email: {sylvain.lesage,remi.gribonval}@irisa.fr
Tel: +33 2 99 84 74 40

***Abstract*— Sparse approximation using redundant dictionaries is an efficient tool for many applications in the field of signal processing. The performances largely depend on the adaptation of the dictionary to the signal to decompose. As the statistical dependencies are most of the time not obvious in natural high-dimensional data, learning fundamental patterns is an alternative to analytical design of bases and has become a field of acute research. Most of the time, the underlying patterns of a class of signals can be found at any time, and in the design of a dictionary, this translation invariance property should be present. We present a new algorithm for learning short generating functions, each of them building a set of atoms corresponding to all its translations. The resulting dictionary is highly redundant and translation invariant.**

I. Introduction and Motivation

Due to its potential for many tasks in signal processing, such as analysis, denoising, compression or source separation, sparse decomposition using redundant dictionaries is a very active domain [3], [7], [9]. The central problem of sparse approximation is the following: given a discrete signal s of support of size S, pick up N basic elements ϕ_k in a huge collection of signals $\mathcal{D}$, referred to as a dictionary, and combine them to build a good approximation:

$$\tilde{s}_N = \sum_{k=0}^{N-1} c_k \phi_k, \quad \phi_k \in \mathcal{D}\,, \quad \|s - \tilde{s}_N\|_2 \leq \epsilon\,. \qquad (1)$$

The approximant $\tilde{s}_N$ is said sparse when $N \ll S$. Finding the best approximation given a dictionary $\mathcal{D}$ is a largely covered subject [7], [4], [3]. In this article, we focus on the design of dictionaries able to give a satisfying solution to the above problem for a class of signals. A dictionary will be efficient if it closely matches the underlying processes of the signals, and many work has been done to taylor adapted dictionaries [1], [2], [4], [5], [6].

The properties of the signal, dictionary and algorithm, are tightly linked. Often, natural signals have highly complex underlying structures which makes it difficult to explicitly define the link between a class of signals and a dictionary. The rest of the paper presents a learning algorithm that tries to capture the underlying structures under the hypothesis of translation invariance in order to maximize the approximation capabilities.

Learning short generating functions that define a dictionary by applying translations is notably motivated by the fact that natural signals often exhibit statistical properties invariant to translation, and it allows to generate huge dictionaries while using only few parameters. In addition, fast convolution algorithms can be used to compute the scalar products when using pursuit algorithms.

The first section formalizes the learning problem and presents the principle of the algorithm. The next section presents the kind of generating functions obtained when using the proposed algorithm on real data. The last section concludes and discusses the benefits and drawbacks of this new algorithm and proposes some extensions.

II. Principles and algorithms

Formally, the aim is to learn a collection $\mathcal{G} = \{g_i\}_{i=1}^K$ of generating functions g_i such that a highly redundant dictionary $\mathcal{D}$ can be created by applying all possible translations to the generating functions of $\mathcal{G}$.

In this paper, we denote the infinite size signals by low-case letters, e.g. s. If the signal has a support of size S, the restriction to its support is denoted by $\mathbf{s} \in \mathbb{R}^S$. More generally vectors and matrices are written in bold letters. Let T_p be the operator that translates a signal by $p \in \mathbb{Z}$ samples.

Let the set $\{T_p g_i, p \in \mathbb{Z}\}$ contain all possible atoms generated by translating g_i. The dictionary obtained from all the generating functions of $\mathcal{G}$ is $\mathcal{D} = \{T_p g_i, i = 1 \dots K, p \in \mathbb{Z}\}$.

The learning is done on a training set of Q signals $\{f_q\}_{q=1}^Q$ having a support of size S_f. Similarly, the generating functions g_i have a support of size $S_g \leq S_f$.

The proposed algorithm learns the generating functions iteratively. The first one is intended to generate the dictionary $\{T_p g_1, p \in \mathbb{Z}\}$ that is the most correlated to the learning signals. Hence, it is equivalent to the following optimization problem:

$$\text{UP: } g_1 = \underset{\|g\|_2=1}{\operatorname{argmax}} \sum_{q=1}^{Q} \max_p |\, \langle f_q, T_p g \rangle \,|^2\,. \qquad (2)$$

In order not to recover several times the same generating function, a constraint forcing all the generating functions to be as decorrelated possible is added. Assuming that $k-1$ functions have been learnt, g_k is the solution of the following constrained optimization problem:

$$\text{CP: } g_{=} \underset{\|g\|_2=1}{\operatorname{argmax}} \frac{\sum_{q=1}^{Q} \max_p \mid \langle f_q, T_p g \rangle \mid^2}{\sum_{l=0}^{k-1} \sum_p \mid \langle g_l, T_p g \rangle \mid^2}. \tag{3}$$

Finding the best solution to the unconstrained (UP) or constrained problem (CP) is difficult. It is decomposed into two tractable steps :

- for a given generating function $g_k^{(i)}$, find the best translations $p_q^{(i)}$ on each training signal f_q,
- update $g_k^{(i+1)}$ by solving UP or CP, where the optimal translations p_q are replaced by using the previously found translations $p_q^{(i)}$.

The first step only consists in finding the location of the maximal correlation between $g_k^{(i)}$ and each training signal f_q.

Let now consider the second step. As the translation admits a well defined adjoint operator, $\langle f_q, T_p g \rangle$ can be replaced by $\langle T_{-p} f_q, g \rangle$. Let $\mathbf{F}^{(i)}$ be the matrix whose q^{th} column is $\mathbf{f}_{q,-p_q^{(i)}}$, the restriction of the signal $T_{-p_q^{(i)}} f_q$ to the support of the generating function g_k, of size S_g. Denoting $\mathbf{g}$ the restriction of g to its support, the second step of the unconstrained problem can be written:

$$\mathbf{g}_k^{(i+1)} = \underset{\|\mathbf{g}\|_2=1}{\operatorname{argmax}} \mathbf{g}^T \mathbf{A}^{(i)} \mathbf{g}, \tag{4}$$

where $\mathbf{A}^{(i)} = \mathbf{F}^{(i)} \mathbf{F}^{(i)^T}$, and $.^T$ denotes the transposition. The best generating function $\mathbf{g}_k^{(i+1)}$ is the eigenvector associated with the biggest eigenvalue of $\mathbf{A}^{(i)}$.

For the second step of the constrained problem, denoting:

$$\mathbf{B}_k = \sum_{l=1}^{k-1} \sum_p \mathbf{g}_{l,-p} \mathbf{g}_{l,-p}^T, \tag{5}$$

the decorrelation constraint consists in minimizing $\mathbf{g}^T \mathbf{B}_k \mathbf{g}$, and the second step becomes:

$$\mathbf{g}_k^{(i+1)} = \underset{\|g\|_2=1}{\operatorname{argmax}} \frac{\mathbf{g}^T \mathbf{A}^{(i)} \mathbf{g}}{\mathbf{g}^T \mathbf{B}_k \mathbf{g}}. \tag{6}$$

The best generating function $\mathbf{g}_k^{(i+1)}$ is the eigenvector associated with the biggest eigenvalue λ of the generalized eigenvalue problem:

$$\mathbf{A}^{(i)} \mathbf{g} = \lambda \mathbf{B}_k \mathbf{g}. \tag{7}$$

In order to use the constrained problem formulation for all the iterations, we define $\mathbf{B}_1 = \mathbf{Id}$ for learning the first generating function.

The algorithm, which we call MoTIF, for Matching of Time Invariant Filters, is summarized in **Algorithm 1**.

Algorithm 1 Principle of the learning algorithm (called MoTIF)

$k = 0$, training signals set $\{f_q\}$
while not enough generating functions **do**
  $k \leftarrow k+1$, $i \leftarrow 0$, initialize $g_k^{(0)}$ (e.g. randomly)
  $\mathbf{B}_k \leftarrow \sum_{l=1}^{k-1} \sum_p \mathbf{g}_{l,-p} \mathbf{g}_{l,-p}^T$
  while no convergence reached **do**
    $i \leftarrow i+1$
    for each training signal f_q, find $p_q^{(i)} = \operatorname{argmax}_p \mid \langle f_q, T_p g^{(i)} \rangle \mid$, corresponding to the location of the maximal correlation between f_q and $g^{(i)}$,
    $\mathbf{A}^{(i)} \leftarrow \sum_{q=1}^{Q} \mathbf{f}_{q,-p_q^{(i)}} \mathbf{f}_{q,-p_q^{(i)}}^T$
    find $g_k^{(i+1)} = \operatorname{argmax}_{\|g\|_2=1} \frac{\mathbf{g}^T \mathbf{A}^{(i)} \mathbf{g}}{\mathbf{g}^T \mathbf{B}_k \mathbf{g}}$, corresponding to the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem $\mathbf{A}^{(i)} \mathbf{g} = \lambda \mathbf{B}_k \mathbf{g}$.
  end while
end while

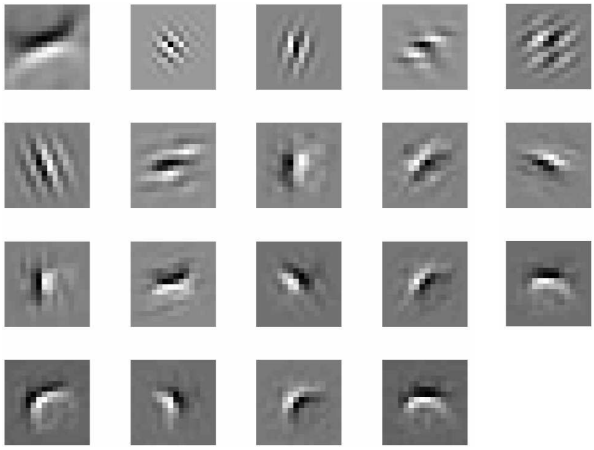

Fig. 1. 19 generating functions learnt on natural images

III. Experimental Results

The first experiment is done on the set of natural images used by Olshausen et al. in [8] (the same pre-conditioning was applied). The size of the patches f_q is 31x31 pixels, whereas the generating functions are 16x16 images. For the computation of eigenvectors, the two-dimensional patches are reshaped into vectors. The search of the optimal positions is done directly on the patches. Figure 1 shows the 19 first generating functions that have been learnt by MoTIF. They are spatially localized and oriented. They are oscillating in a direction different from the orientation and at different frequencies. The first generating functions are Gabor atoms, the second series contains line edge detectors, and the last are curved edge detectors. The two first categories were already observed in [2] and the third completes the range of natural features. Figure 2 presents a closer view to a learnt generating function behaving as a curved edge detector.

The second experiment deals with EEG (electroencephalograph) signals. Finding fundamental time-invariant patterns

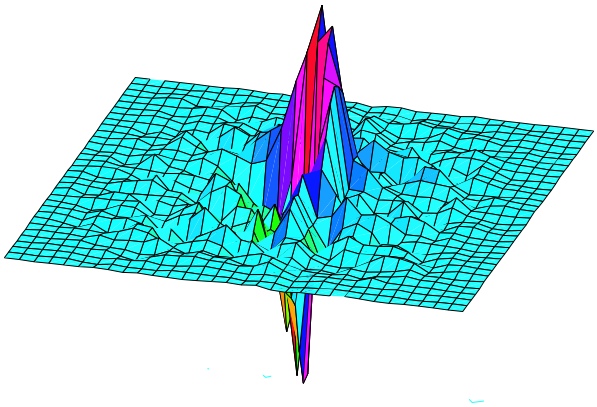

Fig. 2. Learnt curved edge detector generating function.

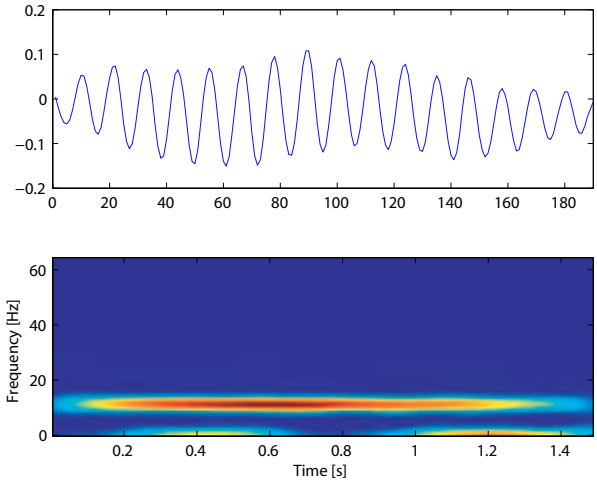

Fig. 3. Learnt generating function on EEG signals having the same frequency as a standard Alpha wave (8-12 Hz).

in EEG is nowadays of big interest as the research in the field of brain computer interfaces is becoming more and more popular. The constituting elements of EEG signals are often described in terms of characteristic frequency bands. We are interested in knowing if MoTIF is able to recover generating functions corresponding to some of these frequencies. Figure 3 presents the first generating function learnt by MoTIF and the corresponding time-frequency representation. The dominant frequency is centered at 11.5 Hz, which makes this function a highly probable candidate to represent the Alpha waves present in the signal. Figure 4 presents another generating function learnt whose frequencies are typical from the Beta waves.

IV. Conclusions

We have presented a new method for learning a set of translation-invariant functions adapted to a class of signals. At every iteration, the algorithm produces the waveform that is the most present in the signals and adds all its shifted versions to the dictionary. A constraint in the objective function forces the learnt waveforms to have low correlation, such that no atom is picked several times. The main drawback of this method is that the generating functions found just after the first one may exhibit features not necessarily present in the

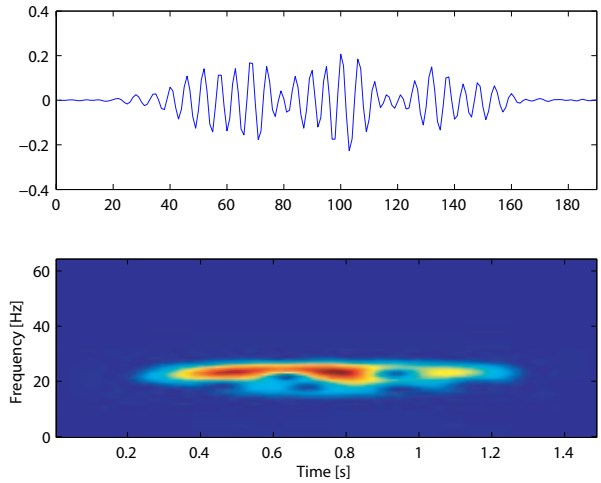

Fig. 4. Learnt generating function on EEG signals having the same frequency as a standard Beta wave (13-30 Hz).

signal. It is due to the decorrelation constraint that is too strong if the underlying generating functions are all similar. However, the constrained algorithm seems to capture the underlying processes quite well, notably when they are really decorrelated. On real data like images, the learnt generating functions are edge detectors (spatially local and oriented) as previously found by Bell and Sejnowski. On EEG, the algorithm recovers the classical waves present in the signal.

In the future, some possible extensions of this algorithm will be studied as learning multichannel generating fuctions from multichannel training signals. The potential of the proposed algorithm for applications as source separation will also be explored. Using the properties of the scalar product, we also plan to explore invariance for different transformations that admit a well defined adjoint (e.g. translations + rotations for images).

References

[1] S.A. Abdallah and M.D. Plumbley. If edges are the independent components of natural images, what are the independent components of natural sounds? In *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, pages 534–539, december 2001.

[2] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[3] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scientific Comp.*, 20:33–61, 1999.

[4] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, 2003.

[5] M.S. Lewicki and B. Olshausen. A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America*, 1999.

[6] M.S. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[7] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, Dec 1993.

[8] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[9] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.