MoTIF: AN EFFICIENT ALGORITHM FOR LEARNING TRANSLATION INVARIANT DICTIONARIES

Philippe Jost,Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne
Signal Processing Institute
CH-1015 Lausanne, Switzerland
{philippe.jost,pierre.vandergheynst}@epfl.ch

Sylvain Lesage,Rémi Gribonval

IRISA-INRIA
Campus de Beaulieu
35042 Rennes CEDEX, France
{sylvain.lesage,remi.gribonval}@irisa.fr

ABSTRACT

The performances of approximation using redundant expansions rely on having dictionaries adapted to the signals. In natural high-dimensional data, the statistical dependencies are, most of the time, not obvious. Learning fundamental patterns is an alternative to analytical design of bases and is nowadays a popular problem in the field of approximation theory. In many situations, the basis elements are shift invariant, thus the learning should try to find the best matching filters. We present a new algorithm for learning iteratively generating functions that can be translated at all positions in the signal to generate a highly redundant dictionary.

1. INTRODUCTION AND MOTIVATION

The tremendous activity in the field of sparse approximation [1, 2, 3] is partly motivated by the potential of the related techniques for typical tasks in signal processing such as analysis, dimensionality reduction, de-noising or compression.

Given a signal s of support of size S in a space of infinite size discrete signals, the central problem is the following: compute a good approximation $\tilde{s}_N$ as a linear superposition of N basic elements picked up in a huge collection of signals $\mathcal{D} = \{\phi_k\}$, referred to as a dictionary :

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

The approximant $\tilde{s}_N$ is sparse when $N \ll S$. The main advantage of this class of techniques is the complete freedom in designing the dictionary, which can then be efficiently tailored to closely match signal structures [4, 5, 6, 7, 8].

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. This paper presents a learning algorithm that tries to capture the underlying structures. In our approach, instead of considering atoms ϕ_k having the same support as the signal s, we propose to learn small generating functions, each of them defining a set of atoms corresponding to all its translations. This is notably motivated by the fact that natural signals often exhibit statistical properties invariant to translation, and that using generating functions 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 proposed algorithm learns the generating functions successively and can be stopped when a sufficient number of atoms have been found.

First, we formalize the problem of learning generating functions, and we propose an iterative algorithm to learn successively some adapted atoms, with a constraint on their decorrelation. The following section presents the influence of this constraint on the recovery of underlying atoms, depending on their correlation. A second experiment shows the ability of this learning method to give an efficient dictionary for sparse approximations. We also show that this algorithm recovers the atoms typically learned by other methods on natural images. We then conclude on the benefits of this new approach and list the perspectives we will consider.

2. PRINCIPLE AND ALGORITHM

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

For the rest of the paper, we assume that the signals denoted by lower case letters are discrete and of infinite size. Finite size vectors and matrices are denoted with bold characters. Let T_p be the operator that translates an infinite signal by $p \in \mathbb{Z}$ samples. Let the set $\{T_p g_k\}$ contain all possible atoms generated by applying the translation operator to g_k. The dictionary generated by $\mathcal{G}$ is $\mathcal{D} = \{\{T_p g_k\}, k = 1 \dots K\}$.

The learning is done using a training set $\{f_n\}_{n=1}^{N}$ of N training signals of infinite size and non null on their support of size S_f. Similarly, the size of the support of the generating

functions to learn is S_g such that $S_g \leq S_f$.

The proposed algorithm learns translation invariant filters iteratively. For the first one, the aim is to find g_1 such that the dictionary $\{T_p g_1\}$ is the most correlated in mean with the signals in the training set. Hence, it is equivalent to the following optimization problem:

$$\text{UP} : g_1 = \underset{\|g\|_2=1}{\operatorname{argmax}} \sum_{n=1}^{N} \max_{p_n} \mid \langle f_n, T_{p_n} g \rangle \mid^2 . \quad (2)$$

For learning the next generating functions, the original optimization problem is modified to include a constraint penalizing a generating function if a similar one has already been found. Assuming that $k-1$ generating functions have been learnt, the optimization problem to find g_k can be written as:

$$\text{CP} : g_k = \underset{\|g\|_2=1}{\operatorname{argmax}} \frac{\sum_{n=1}^{N} \max_{p_n} \mid \langle f_n, T_{p_n} g \rangle \mid^2}{\sum_{l=0}^{k-1} \sum_p \mid \langle g_l, T_p g \rangle \mid^2}. \quad (3)$$

Finding the best solution to the unconstrained problem (UP) or the contrained problem (CP) is hard, and we propose to decompose them in two simpler steps that are alternately solved :

- for a given generating function $g_k^{(i)}$, find the best translations $p_n^{(i)}$,
- update $g_k^{(i+1)}$ by solving UP or CP, where the optimal translations p_n are fixed to the previous values $p_n^{(i)}$.

The first step only consists in finding the location of the maximum correlation between each learning signal f_n and the generating function g.

Let now consider the second step and define $\mathbf{g}_k \in \mathbb{R}^{S_g}$ the restriction of the infinite size signal g_k to its support. As the translation admits a well defined adjoint operator, $\langle f_n, T_{p_n} g_k \rangle$ can be replaced by $\langle T_{-p_n} f_n, g_k \rangle$. Let $\mathbf{F}^{(i)}$ be the matrix (S_f rows, N columns), whose columns are made of the signals f_n shifted by $-p_n^{(i)}$. More precisely, the j^{th} column of $\mathbf{F}^{(i)}$ is $\mathbf{f}_{n,-p_n^{(i)}}$, the restriction of $T_{-p_n^{(i)}} f_n$ to the support of g_k, of size S_g. We denote $\mathbf{A}^{(i)} = \mathbf{F}^{(i)} \mathbf{F}^{(i)^T}$.

With these notations, the second step, for the *unconstrained* problem, can be written :

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

where $.^T$ 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 *constrained* problem, we want to force $g_k^{(i+1)}$ to be as decorrelated as possible from all the atoms in $\mathcal{D}_{k-1}$. This corresponds to minimizing

$$\sum_{l=1}^{k-1} \sum_p |\langle T_{-p} g_l, g \rangle|^2 \quad (5)$$

or, denoting

$$\mathbf{B}_k = \sum_{l=1}^{k-1} \sum_p \mathbf{g}_{l,-p} \mathbf{g}_{l,-p}{}^T, \quad (6)$$

to minimizing $\mathbf{g}^T \mathbf{B}_k \mathbf{g}$. With these notations, the constrained problem can be written :

$$\mathbf{g}_k^{(i+1)} = \underset{||\mathbf{g}||_2=1}{\operatorname{argmax}} \frac{\mathbf{g}^T \mathbf{A}^{(i)} \mathbf{g}}{\mathbf{g}^T \mathbf{B}_k \mathbf{g}} \quad (7)$$

The best generating function $\mathbf{g}_k^{(i+1)}$ is the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem defined in eq. 7. Note that defining $\mathbf{B}_1 = \mathbf{Id}$, we can use CP for learning the first generating function $\mathbf{g}_1$.

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

Algorithm 1 Principle of the learning algorithm (MoTIF)

$k = 0$, training signals set $\{f_n\}$
while not enough generating functions **do**
  $k \leftarrow k+1$, $i \leftarrow 0$
  $\mathbf{B}_k \leftarrow \sum_{l=1}^{k-1} \sum_p \mathbf{g}_{l,-p} \mathbf{g}_{l,-p}{}^T$
  while no convergence reached **do**
    $i \leftarrow i+1$
    for each f_n, find $p_n^{(i)} = \operatorname{argmax}_p \mid \langle f_n, T_p g^{(i)} \rangle \mid$, by locating the maximum correlation between f_n and $g^{(i)}$,
    $\mathbf{A}^{(i)} \leftarrow \sum_{n=1}^{N} \mathbf{f}_{n,-p_n^{(i)}} \mathbf{f}_{n,-p_n^{(i)}}{}^T$
    find $\mathbf{g}_k^{(i+1)} = \operatorname{argmax}_{||\mathbf{g}||_2=1} \frac{\mathbf{g}^T \mathbf{A}^{(i)} \mathbf{g}}{\mathbf{g}^T \mathbf{B}_k \mathbf{g}}$, that is 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

The unconstrained algorithm has been proven to converge in a finite number of iterations to a generating function locally maximizing the unconstrained problem (eq. 2) and we observed on numerous experiments that the constrained algorithm typically converges in few steps to a stable solution independently of the initialization.

3. EXPERIMENTAL RESULTS

The first experiment consists in exploring the ability of the algorithm to recover correctly a set $\mathcal{G}^O = \{g_k^O\}_{k=1}^K$ of known

generating functions refereed to as the original set of functions. Starting from this set, a sparse coefficient vector c is randomly created. It defines a signal :

$$s = \sum_{k=0}^{N-1} c_k \phi_k, \quad \phi_k \in \mathcal{D} = \{\{T_p g_k^O\}, k = 1..K\}.$$

The training set $\{f_n\}$ is obtained by taking the maximal number of non overlapping parts of the signal s. The size of the patches f_n is such that $supp(f_n) = 2 * supp(g_k^O) - 1$, where $supp$ denotes the size of the support. These patches are used by the MoTIF algorithm to learn a set $\mathcal{G}$ of translation invariant generating functions. A function g_i^O from the original set $\mathcal{G}^O$ is said to be recovered if $\max_{g \in \mathcal{G}} |\langle g_i, g \rangle| > \delta$.

We created 3000 original sets of generating functions made of 3 Gabor atoms with random normalized frequency between 0 and 0.5. The size of their spatial support is 16. Each of these generating functions was present 10 times in a signal of size 1600 with a random amplitude between 0 and 1. The number of patches f_n used was 298. For each set of generating function, we run the algorithm 10 times on 10 different signals.

Figure 1 illustrates the recovery ability of the MoTIF algorithm. It presents the mean number of generating functions recovered as a function of the minimal correlation of the original set $\mathcal{G}^O$ computed as $\min_{i,j} \max_p |\langle T_p g_i^O, g_j^O \rangle|$, which means that the correlation between other atoms can only be higher. The equivalence limit δ for two generating functions was fixed to 0.8.

For the same settings, in more than 2 cases out of 3, the first generating function found by MoTIF is one from the original set. To recover the next atoms, the constrained optimization problem (CP) has to be solved. Thus, the next functions are constrained to be as uncorrelated as possible with the past found functions, which is clearly not the case when the original set of functions is highly coherent. This leads to a poor rate of recovery when the minimal coherence is higher than 0.6. Recovering is easier when dealing with rather uncorrelated set of functions. Indeed, for very small values of the minimal correlation, in mean, nearly two functions out of three are recovered. In between these two extreme cases (minimal coherence between 0.2 and 0.6), the algorithm's behavior is rather constant and recovers more than half of the functions.

The second experiment studies the ability of a dictionary learnt on real data to sparsely approximate signal of the class of the learning data, compared to a classical dictionary like Gabor atoms. The class of signals we consider is music. The first half of a signal has been used for learning whilst the second part is kept to test the approximation ability. The learning part has been divided into 10422 patches of size 4095 in order to learn generating functions with a spatial support of size 2048. The learnt set $\mathcal{G}_L$ has a cardinality of 50. The reference set $\mathcal{G}_R$ of generating functions is made of multi-scale Gabor atoms with 50 different normalized frequencies spread between 0 and 0.5 and 5 different scales. Thus the cardinality of $\mathcal{G}_R$ is 250.

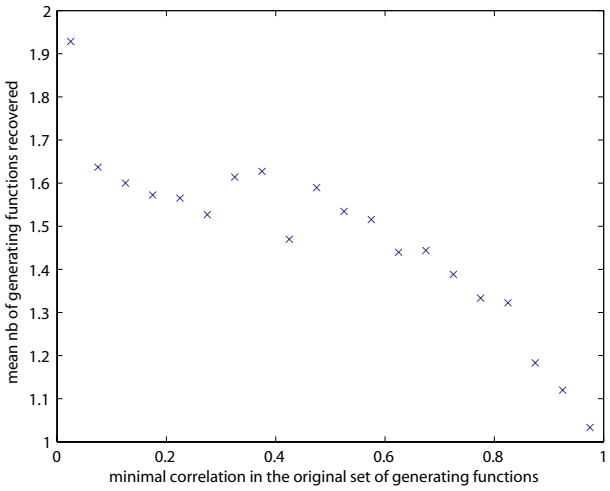

Fig. 1. Mean number of recovered generating functions as a function of the minimal coherence of the reference dictionary.

To compare the approximation performances of both sets, we used Matching Pursuit [1] with the dictionaries $\mathcal{D}_L$ and $\mathcal{D}_R$ generated respectively by $\mathcal{G}_L$ and $\mathcal{G}_R$. Figure 2 presents the obtained results. The length of the test signal is 50000 and 500 iterations of Matching Pursuit were performed for the approximation. Thus, this approximation is more than 100 times sparse. Even if the cardinality of the learnt dictionary is 5 times smaller than the reference dictionary, the decay of the mean square error (MSE) is faster. The learnt dictionary is adapted to the considered signal and contains meaningful features present in the test signal.

The third experiment is done on a set of natural images. The two-dimensional patches are reshaped in vectors for the computation. The size of the training signals f_n is 31×31 pixels, whereas the generating functions are 16×16 images. We learn 19 generating functions with the constrained algorithm. They are shown on figure 3. The generating functions are spatially localized and oriented. They are oscillating in a different direction from the orientation, at different frequencies depending on the atoms. The generating functions #2 to #5 are mainly high frequency due to the decorrelation constraint with the first atom. Whereas 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 [4] and the third ones complete the range of natural features.

4. 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

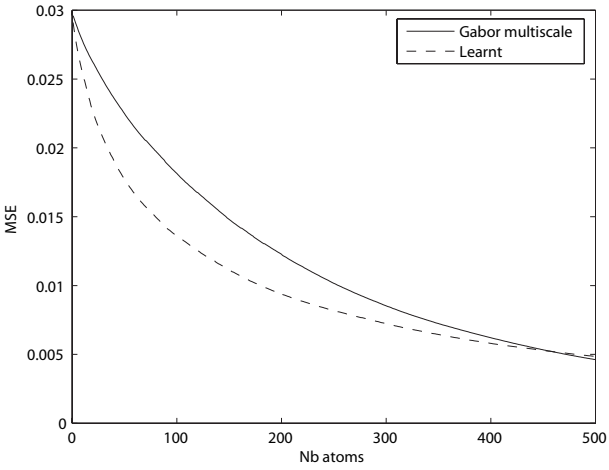

Fig. 2. Approximation abilities of a learnt set of generating functions regarding a reference set of generating functions containing multi-scale Gabor atoms.

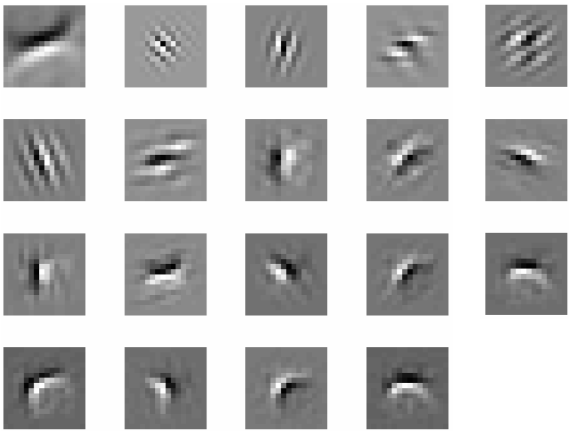

Fig. 3. 19 generating functions learnt on natural images.

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 the fact that the few generating functions following the first one are mainly due to the decorrelation constraint, more than the attachment to the signal. Despites, this constrained algorithm seems to capture the underlying processes quite well, notably when they are really decorrelated. The learnt dictionaries show ability to sparse decompose the corresponding signals. On real data like images, the learnt generating functions are edge detectors (spatially local and oriented) as previously found by Bell and Sejnowski. Some extensions of this algorithm are considered, as learning multichannel atoms on multichannel signals. Using this type of learning, some applications in multichannel source separation can be expected. Another extension, based on the properties of the inner product, is to replace the translation invariance by the invariance to a whole set of transformations that admit a well defined adjoint (e.g. translations + rotations for images).

5. REFERENCES

[1] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

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

[3] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[4] A.J. Bell and T.J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[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*, vol. 12, no. 2, pp. 337–365, 2000.

[7] 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*, vol. 15, pp. 349–396, 2003.

[8] 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*, december 2001, pp. 534–539.