Learning redundant dictionaries with translation invariance property: the MoTIF algorithm
Philippe Jost, Pierre Vandergheynst, Sylvain Lesage, Rémi Gribonval

To cite this version:

HAL Id: hal-00816804
https://inria.hal.science/hal-00816804
Submitted on 23 Apr 2013
Learning redundant dictionaries with translation invariance property: the MoTIF algorithm

Philippe Jost, Pierre Vanderdaghyst
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 3 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 \( \phi_k \) in a huge collection of signals \( D \), referred to as a dictionary, and combine them to build a good approximation:

\[
\tilde{s}_N = \sum_{k=0}^{N-1} c_k \phi_k, \quad \phi_k \in D, \quad \|s - \tilde{s}_N\|_2 \leq \epsilon. \tag{1}
\]

The approximant \( \tilde{s}_N \) is said sparse when \( N \ll S \). Finding the best approximation given a dictionary \( D \) 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 \( G = \{g_i\}_{i=1}^K \) of generating functions \( g_i \) such that a highly redundant dictionary \( D \) can be created by applying all possible translations to the generating functions of \( G \).

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 \( s \in \mathbb{R}^S \). 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 \( G \) is \( D = \{T_p g_i, i = 1 \ldots K, p \in \mathbb{Z}\} \).

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 \( S_y \leq S_f \).

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:

\[
\text{UP: } g_1 = \arg \max_{\|g\|_2 = 1} \sum_{q=1}^Q \max_p |\langle f_q, T_p g \rangle|^2. \tag{2}
\]
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:

\[
CP: \quad g_k = \arg\max_{\|g\|_2=1} \sum_{q=1}^{Q} \max_p | \langle f_q, T_p g \rangle |^2 /
\]

\[
\sum_{q=1}^{Q} \sum_p | \langle g_k, T_p g \rangle |^2 .
\]

(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 \( p_q^{(i)} \) 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 \( F^{(i)} \) be the matrix whose \( q^{th} \) column is \( f_{q-p}^{(i)} \), the restriction of the signal \( T_{-p} f_q \) to the support of the generating function \( g_k \). of size \( Q \). Denoting \( g \) the restriction of \( g \) to its support, the second step of the unconstrained problem can be written:

\[
g_k^{(i+1)} = \arg\max_{\|g\|_2=1} g^T A^{(i)} g ,
\]

(4)

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

For the second step of the constrained problem, denoting:

\[
B_k = \sum_{l=1}^{k-1} \sum_p g_{l,-p} T_p g_{l,-p}^T ,
\]

(5)

the decorrelation constraint consists in minimizing \( g^T B_k g \), and the second step becomes:

\[
g_k^{(i+1)} = \arg\max_{\|g\|_2=1} g^T A^{(i)} g / g^T B_k g .
\]

(6)

The best generating function \( g_k^{(i+1)} \) is the eigenvector associated with the biggest eigenvalue \( \lambda \) of the generalized eigenvalue problem:

\[
A^{(i)} g = \lambda B_k g .
\]

(7)

In order to use the constrained problem formulation for all the iterations, we define \( B_1 = \text{Id} \) 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)**

1: \( k = 0 \), training signals set \( \{f_q\} \)
2: while not enough generating functions do
3: \( k \leftarrow k + 1 \), \( i \leftarrow 0 \), initialize \( g_k^{(0)} \) (e.g. randomly)
4: \( B_k \leftarrow \sum_{i=1}^{k-1} \sum_p g_{i,-p} T_p g_{i,-p}^T \)
5: while no convergence reached do
6: \( i \leftarrow i + 1 \)
7: for each training signal \( f_q \), find \( p_q^{(i)} = \arg\max_p | \langle f_q, T_p g_k^{(i)} \rangle | \), corresponding to the location of the maximal correlation between \( f_q \) and \( g_k^{(i)} \)
8: \( A^{(i)} \leftarrow \sum_{q=1}^{Q} \langle f_q - p_q^{(i)} f_q T_{-p_q^{(i)}} \rangle^T \)
9: find \( g_k^{(i+1)} = \arg\max_{\|g\|_2=1} g^T A^{(i)} g / g^T B_k g \), corresponding to the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem \( A^{(i)} g = \lambda B_k g \).
10: end while
11: 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 (electroencephalo-graph) signals. Finding fundamental time-invariant patterns
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 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 functions 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