

Sparse decomposition over multi-component redundant dictionaries

Lorenzo Granai and Pierre Vandergheynst
Signal Processing Institute
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland
home page: <http://lts2www.epfl.ch>
email: {lorenzo.grnai,pierre.vandergheynst}@epfl.ch

Abstract—In many applications - such as compression, de-noising and source separation - a good and efficient signal representation is characterized by sparsity. This means that many coefficients are close to zero, while only few ones have a non-negligible amplitude.

On the other hand, real-world signals - such as audio or natural images - clearly present peculiar structures. In this paper we introduce a global optimization framework that aims at respecting the sparsity criterion while decomposing a signal over an overcomplete, multi-component dictionary. We adopt a probabilistic analysis which can lead to consider the signal internal structure. As an example that fits this framework, we propose the Weighted Basis Pursuit algorithm, based on the solution of a convex, non-quadratic problem. Results show that this method can provide sparse signal representations and sparse m -terms approximations. Moreover, Weighted Basis Pursuit provides a faster convergence compared to Basis Pursuit.

I. INTRODUCTION

Over last years there has been a burst of alternatives to traditional signal representations. Instead of representing signals as a superimposition of sinusoids or wavelets, we now have many available alternative dictionaries, most of which are overcomplete. This means that some elements of the dictionary can be described in terms of other ones, therefore a signal decomposition is non-unique. If on one hand this can increase the complexity of the signal analysis, on the other it gives us the possibility of adaptation, selecting among many possible representations the one which (most) fits our purposes.

The dictionaries (\mathcal{D}) we are interested in are large and overcomplete and in particular they are built by the union of $d \geq 2$ sub-dictionaries. Each of them is particularly suitable for describing a certain feature of a structured signal. We call a basis set with these characteristics a Multi-Component Dictionaries (MCD):

$$\mathcal{D} = \bigcup_i \mathcal{D}_i, \text{ with } 1 \leq i \leq d. \quad (1)$$

Signal decompositions we are interested in have to be sparse. The criterion of sparsity has been studied for a long time and in the last few years has become popular in the signal processing community [1], [2], [3]. Applications can be found in analysis and compression of audio [4],

images [5] and video [6]. Sparsity criteria are also present in deconvolution, signal modeling and de-noising, source separation and classification.

The most common measure of sparsity of a vector c , labeled with indexes in Ω , is its l_0 quasi-norm defined as the number of elements different from zero. The smallest is this norm, the sparsest is the vector. Unfortunately this measure is not convex. We can define a p -norm for any positive p :

$$\|c\|_p = \left(\sum_{i \in \Omega} |c_i|^p \right)^{1/p}. \quad (2)$$

It is well known that the smallest p for which Eqn. (2) is convex is 1. For this reason the Basis Pursuit (BP) principle proposed in [7] selects among all the possible representations of a signal $s \in R^n$ the one with the minimal l_1 norm. Formally, BP solves the problem:

$$\min_c \|c\|_1 \text{ s.t. } \mathcal{D}c = s, \quad (3)$$

where \mathcal{D} is the dictionary of cardinality k , here expressed as a matrix of size $n \times k$. The columns of \mathcal{D} are the functions that compose the dictionary (also called atoms). Recently it has been proved that, under certain conditions on the signal and the dictionary, the solution of (3) coincides with the one that minimizes the l_0 norm [1], [2].

It is now possible to introduce a signal decomposition over a MCD:

$$s = \sum_{i=1}^d \sum_{j \in \Lambda_i} c_j \cdot g_j, \quad (4)$$

where Λ_i is a subset of the indexes of the basis functions g in sub-dictionary \mathcal{D}_i and c_j are the coefficients of the functions g_j . The vector c , composed by all the c_j in (4) is then the representation of s over \mathcal{D} , i.e. $s = \mathcal{D}c$.

Signals we are interested in are structured, i.e. they present peculiar features that, if taken into account, can improve their representation and/or approximation. These structures can be for example transients and stationary parts in audio signals or edges, textures and smooth parts in natural images. This explains our decision to work with MCDs. The problem that we are addressing aims at representing a signal with few and significant components. While the meaning of the first requisite we are looking

for is well explained in terms of l_0 norm, for the second one we have to point out that a component is considered significant if it catches one of the basic features we have pointed out.

II. PROPOSED FRAMEWORK

Here we introduce a decomposition framework which considers together a sparsity criterion and a probabilistic model. Let \mathcal{D} be a dictionary with cardinality k and composed by unit norm functions, it is possible to define a new vector d , such that

$$d_i = \frac{c_i}{p_i}, \quad \text{with } i = 1, \dots, k. \quad (5)$$

In (5) the weights p_i are computed after a probabilistic analysis of the signal and indicate how much the i^{th} function, with its parameters and its location, fits the structure of the signal. The problem we propose to solve now, instead of (3), is:

$$\min_{d \in R^k} \|d\|_1 \quad \text{s.t.} \quad \mathcal{D}c = s. \quad (6)$$

In the following we will denote this problem as Weighted Basis Pursuit (WBP). Note that here the minimization is made on d , but this vector is related to c by Eqn. (5). The following theorem proves that, although WBP involves non-linear optimization, it can be equivalently reformulated as a Linear Programming (LP) problem [8].

Theorem 1: The problem expressed by equations (5) and (6) can be solved by Linear Programming techniques, reducing it in the standard form:

$$\min_x f^T x \quad \text{s.t.} \quad Ax = b \quad \text{and} \quad x \geq 0. \quad (7)$$

Proof: Let us begin this constructive proof by defining a new $n \times k$ matrix Φ such that $\Phi_{j,i} = D_{j,i} \cdot p_i$, with $j = 1, \dots, n$ and $i = 1, \dots, k$. The problem in equation (6) can be now rewritten as:

$$\min_{d \in R^k} \|d\|_1 \quad \text{s.t.} \quad \Phi d = s. \quad (8)$$

The equivalence between this new problem and (7) is well know (e.g. see [8]) and it is also the one that allows to solve BP with LP.

It is important to highlight that the solution of the LP problem will be the vector of the weighted coefficients d ; to go back to the coefficients with respect to the dictionary \mathcal{D} one just have to multiply each entry by the corresponding probability, inverting Eqn. (5).

WBP principle, while leaving unchanged the constraints, modifies the vector whose norm has to be minimized. This can lead to respecting both sparsity and meaningfulness of the representation. A signal is no longer treated as a pure numerical array, but as an entity provided by an internal structure. This is true upon condition that a good dictionary is used and the weights p_i in (5) are properly chosen (the latter choice depends also on the former).

III. RESULTS IN 1D

In this section we present some results obtained by applying the WBP principle to 1D signals. The problem of Eqn. (6) has been solved recurring to a Linear Programming technique based on the interior-point method [9]. The dictionary used is the union of two sub-dictionaries: the first one is a Cosine Packet dictionary, aimed at representing the oscillating part of the signal. The second dictionary is composed by all the possible translation of a second derivative of a Gaussian: these functions are able to efficiently represent sudden variations and transients. The probabilities p_i are computed on the basis of a simple analysis of the finite difference of the input signal. This method can highlight the regions where the signal presents sudden variations and so can give an hint on the sub-dictionary which has higher probability to be used at that time. This way of computing the weights is only an example and can be surely improved in the future.

The first signal we analyze is a part of “earthquake”, a signal taken from the dataset provided in [10]. We have $n = 256$ and $k = |\mathcal{D}| = 1792$. In order to compare a BP decomposition (c_{BP}) with a WBP one (c_{WBP}), we first threshold coefficients having a very small amplitude ($< 0.001\%$ of the maximum). BP guarantees to minimize the l_1 norm, but, in general, not the l_0 . So, as expected $\|c_{BP}\|_1 < \|c_{WBP}\|_1$. But we can also observe that c_{WBP} is sparser than c_{BP} and nevertheless the achieved error is smaller. This result does not depend on the adopted threshold, in fact it can be observed that the error given by the m -terms approximations of the signal, (approximations computed considering only the m values of c having the biggest absolute value) is always smaller for the WBP method. Note that neither BP nor WBP are approximation algorithms and the m -terms approximations we refer to are obtained by thresholding from the representation coefficients. The problem of extending these algorithms into an approximation framework will be faced in next section.

Let suppose that a signal has an exact representation c over a certain dictionary \mathcal{D} having coherence μ ; it is theoretically proved that if

$$\|c\|_0 \leq \frac{1}{2} \cdot (\mu^{-1} + 1), \quad (9)$$

BP will recover all the optimal atoms and their correct coefficients [2]. Our example is one of the many cases in which this bound is not respected, having the dictionary a high coherence. Therefore it is possible to find another representation sparser than the one given by BP. This is in practice done by our WBP method.

Further results are achieved decomposing synthetic signals generated by a superposition of waveforms belonging to the dictionary. Here the two methods present the same

performances in terms of sparsity, but WBP turns out to be faster, being its choice driven by the weights.

IV. EXTENSION TO 2D CASE

From its beginning the proposed framework has been thought to be extended to the bi-dimensional case. Natural images present different features that we want to represent using different sub-dictionaries, following the MCD model. Our 2D dictionary is composed by a sub-dictionary \mathcal{D}_{edge} aimed at representing the edges and one \mathcal{D}_{smooth} for representing the smooth parts. The first one is generated by translating (\vec{b}), anisotropically scaling ($\vec{a} = (a_{x_1}, a_{x_2})$) and rotating (θ) a 2D mother function

$$g1(x_1, x_2) = (4x_1^2 - 2) \cdot e^{-(x_1^2 + x_2^2)} \quad (10)$$

on the model of the dictionary presented in [5]. With respect to this dictionary, for the sake of simplicity, the number of rotations and scaling factor is reduced and the bending avoided. \mathcal{D}_{edge} can be parametrized as follows:

$$\mathcal{D}_{edges} = \{g1_{(\vec{b}, \theta, \vec{a})}(x_1, x_2)\}_{\vec{b}, \theta, \vec{a}} \quad (11)$$

The sub-dictionary \mathcal{D}_{smooth} is composed by translated and (isotropically) scaled versions of a bi-dimensional Gaussian function ($g2$) and can be expressed as follows:

$$\mathcal{D}_{smooth} = \{g2_{(\vec{b}, a)}(x_1, x_2)\}_{\vec{b}, a} \quad (12)$$

This multi-component dictionary is just a general prototype aimed at testing the WBP algorithm and the new probability model in the case of natural images. Moreover it can be enriched introducing a third sub-dictionary for textures.

As previously said, the choice of the probability weights p_i of Eqn. (5) has a crucial importance. The algorithm we adopt for computing them is based on the analysis of the dual local autocovariance matrix [11]. Differences in the eigenvalues of an autocovariance matrix indicate directions at which the local Fourier power spectrum of a function is slowly decreasing. It is therefore possible to assign a certain degree of “edginess” to any location in an image looking at the relationship between these eigenvalues. Thanks to a slight variation of the method described in [11] we define an edginess $e(x_1, x_2)$ with values in $[0, 1]$ which tends to 1 when the point (x_1, x_2) has high probability to be on an edge (see figure 2 too):

$$e(x_1, x_2) = 1 - \left(\frac{\lambda_{min}}{\lambda_{max}} \right)^2 \quad (13)$$

λ_{min} and λ_{max} are respectively the minimum and maximum eigenvalue of the dual local autocovariance matrix centered in (x_1, x_2) .

Being our dictionary composed by functions well located in the space we can assign $e(x_1, x_2)$ as weight to the functions of \mathcal{D}_{edge} centered in (x_1, x_2) . Consequently the



Fig. 1. Two 16x16 images of the test set. On the left an artificial image, on the right a crop from cameraman

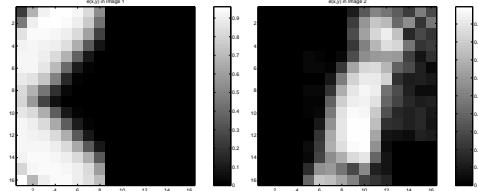


Fig. 2. Edginess of the previous images. White: $e = 1$; Black: $e = 0$

corresponding functions in \mathcal{D}_{smooth} will be weighted by $1 - e(x_1, x_2)$.

In many cases we are not interested in exact representations of images but on approximations that can allow a certain error. Very recently (e.g. see again [3]) first theoretical attempts are made to study the approximation properties of the convex relaxed problem with error constraints. Here we try to approximate an image using the WBP principle, extending it by considering a trade-off between the error and the l_1 norm of the weighted coefficients. Using the same notations as in Eqns. (5) and (6) we formulate the new problem in the following way:

$$\min_c (\|d\|_1 + \lambda \|s - \mathcal{D}c\|_\rho) \quad (14)$$

In this new problem λ weights the error in the minimization process and so controls the trade-off.

Following the model of the proof of theorem 1 we can also show that, setting $\rho = 1$, it is possible to solve this problem by Linear Programming methods. Of course, one can prefer the use of l_2 norm for measuring the error, but in this case the solution of (14) requires more complex quadratic programming techniques.

We now briefly show some results obtained computing approximations of 16x16 images using the WBP method with the error penalization of equation (14). As can be observed in figure 1 these images present strong geometrical characteristics and are quite adapted to be represented by the MCD previously illustrated. However more complex behavior can be caught by bigger and more complete and elaborate dictionaries [5]. The approximations are computed considering only the m terms of c having the biggest absolute value. As can be seen in table I the error achieved by WBP is smaller than the one of BP.

TABLE I

Ratio between the errors obtained by m -terms approximations computed with WBP and BP

ratio between error(WBP) and error(BP)		
m	Image 1	Image 2
200	$7 \cdot 10^{-3}$	$3 \cdot 10^{-3}$
150	$7 \cdot 10^{-2}$	$3 \cdot 10^{-2}$
100	0.40	0.22

V. DISCUSSION

In this work we present a new framework for signal approximation over a multi-component dictionary, based on a probabilistic analysis of the input signal. Since natural signals usually offer a mixture of different features we think that MCD represent a suitable tool for exploiting this diversity. Moreover the proposed probabilistic approach drive the subset selection. Proper weights help in choosing functions in the “correct” sub-dictionary and speed up the search procedure, avoiding the algorithm to search within the basis functions having no probability to be selected.

WBP method has been extended in section IV to the case where the function to minimize includes an error penalty term. In order to measure the error, in the reported results we made use of the l_1 norm, much less sensible to “wild” pixels in the image. This choice has been taken mainly for a question of simplicity. However, the convexity of the problem in (14) is assured for any $\rho \geq 1$. If the sub-dictionaries are orthogonal the function can be minimized also avoiding LP. In fact, just not considering the weights, Eqn.(14) turns out to be equivalent to the problem faced in [12] recurring to the faster Block Coordinate Relaxation (BCR) method [13]. Weights coming from the probabilistic analysis can therefore also be seen as a way to improve the sub-problem of the function selection inside an orthogonal block in the BCR algorithm. Even if BCR (without any weights) is proved to converge only when the blocks (the sub-dictionaries, in our case) are orthogonal, it worths mentioning that this algorithm has been recently used removing the orthogonality constraint and nevertheless it still gives empirical promising results [14].

The framework we introduce is composed by three main points: the MCD design, the probabilistic analysis and the function selection procedure. Concerning the last task here we take inspiration from the Basis Pursuit principle. It can be interesting to study the performances of such approach adopting a different approximation method. The Matching Pursuit algorithm already proved to possess good approximation properties [5] and recently new interesting theoretical results about its convergence were proved [2], [3], [15]. It so can be an appropriate candidate for substituting BP in the proposed framework.

Concerning the probabilistic analysis of an image we

can make a step further. The dual local autocovariance matrix we used to determine the edginess through Eqn. (13) can give also information about the orientation of the edge. It is in fact proved (again in [11]) that the eigenvector of the larger eigenvalue will be normal to the edge (of course when such a normal exists). We can then imagine a more general way to assign probability weights to the basis functions that compose a dictionary (or a sub-dictionary) on the model of (11) and (12). So if we have a sub-dictionary parametrized as follows

$$\mathcal{D}_j = \{\text{atom}_{(\vec{b}, \theta, \vec{a}, \dots)}(x_1, x_2)\}_{\vec{b}, \theta, \vec{a}, \dots}, \quad (15)$$

we can also consider the probability weights as:

$$P = P(\vec{b}, \theta, \vec{a}, \dots). \quad (16)$$

In this case the probabilities do not only depend on the location of the atoms but potentially on all its parameters. This model can certainly expand the horizon of the framework and improve the power of the WBP approach.

REFERENCES

- [1] D. L. Donoho and X. Huo, “Uncertainty principles and ideal atom decomposition,” *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov 2001.
- [2] J. Tropp, “Greed is good : Algorithmic results for sparse approximation,” Texas Institute for Computational Engineering and Sciences, Tech. Rep., 2003.
- [3] —, “Just relax: Convex programming methods for subset selection and sparse approximation,” Texas Institute for Computational Engineering and Sciences, Tech. Rep., 2004.
- [4] R. Gribonval and E. Bacry, “Harmonic decomposition of audio signals with matching pursuit,” *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 101–111, Jan 2003.
- [5] L. Peotta, L. Granai, and P. Vanderghyest, “Very low bit rate image coding using redundant dictionaries,” in *Proc. of 48th SPIE annual meeting*, vol. 2, San Diego, USA, August 2003. [Online]. Available: <http://its2www.epfl.ch/publications.html>
- [6] R. Neff and A. Zakhor, “Very low bit-rate video coding based on matching pursuit,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 158–171, Feb 1997.
- [7] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33 – 61, 1999.
- [8] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Belmont, Mass.: Athena Scientific, 1997.
- [9] Y. Zhang, “Solving large-scale linear programs bt interior-point methods under matlab enviroment,” *Optimization Methods and Software*, vol. 10, pp. 1 – 31, 1998.
- [10] <http://www-stat.stanford.edu/~atomizer/>.
- [11] W. K. Czaja and M. V. Wickerhauser, “Singularity detection in images using dual local autocovariance,” *Applied and Computational Harmonic Analysis*, vol. 13, no. 1, pp. 77–88, July 2002.
- [12] S. Sardy, P. Tseng, and A. Bruce, “Robust wavelet denoising,” *IEEE Trans. Signal Processing*, vol. 49, no. 6, 2001.
- [13] S. Sardy, A. Bruce, and P. Tseng, “Block coordinate relaxation methods for nonparametric wavelet denoising,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361–379, 2000.
- [14] J. L. Starck, D. L. Donoho, and E. J. Candes, “Image decomposition via the combination of sparse representations and a variational approach,” *IEEE Trans. Image Processing*, 2004, submitted to.
- [15] R. Gribonval and P. Vanderghyest, “On the exponential convergence of matching pursuit in quasi-inchoherent dictionaries,” *IEEE Trans. Inform. Theory*, 2004, submitted to.