# On the Use of *A Priori* Information for Sparse Signal Representations

Oscar Divorra Escoda, Lorenzo Granai and Pierre Vandergheynst
Signal Processing Institute (ITS)
Swiss Federal Institute of Technology in Lausanne (EPFL)
LTS2-ITS-STI-EPFL, 1015 Lausanne, Switzerland
**Technical Report No. 18.2004**

**Abstract**

This report studies the effect of introducing *a priori* knowledge to recover sparse representations when overcomplete dictionaries are used. We focus mainly on Greedy algorithms and Basis Pursuit as for our algorithmic basement, while a priori is incorporated by suitably weighting the elements of the dictionary. A unique sufficient condition is provided under which Orthogonal Matching Pursuit, Matching Pursuit and Basis Pursuit are able to recover the optimally sparse representation of a signal when *a priori* information is available. Theoretical results show how the use of "reliable" *a priori* information can improve the performances of these algorithms. In particular, we prove that sufficient conditions to guarantee the retrieval of the sparsest solution can be established for dictionaries unable to satisfy the results of *Gribonval and Vandergheynst* [1] and *Tropp* [2]. As one might expect, our results reduce to the classical case of [1] and [2] when no *a priori* information is available. Some examples illustrate our theoretical findings.

**Index Terms**

Sparse Representations, Basis Pursuit, Matching Pursuit, Greedy Algorithms, *A Priori* Knowledge.

Contents



Web page: http://lts2www.epfl.ch
The work of Oscar Divorra Escoda is partly sponsored by the IM.2 NCCR
The work of Lorenzo Granai is supported by the SNF grant 2100-066912.01/1

## I. Introduction

In this report we focus on the possibility of finding the exact sparsest representation of a signal over a redundant dictionary $\mathcal{D} = \{g_j\}_{j\in\Omega}$ knowing some a priori information. More precisely, given the signal $f \in \mathbb{R}^n$ we want to recover the exact superposition of $m$ elements of the dictionary such that:

$$f = \sum_{\gamma\in\Gamma} b_\gamma g_\gamma, \tag{1}$$

where $\Gamma \subset \Omega$ has cardinality $m$. Therefore, $f \in span\,(g_\gamma,\ \gamma \in \Gamma)$.

Working with an overcomplete dictionary implies that more than one representation is possible. In many applications - such as compression, de-noising or source separation - a good and efficient signal representation is often characterized by sparsity. We thus wish to identify the sparsest solution, that is the vector $\mathbf{b} \in \mathbb{R}^\Omega$ with the smallest support:

$$\min_{\mathbf{b}} \|\mathbf{b}\|_0 \ \ s.t.\ f = D\mathbf{b}, \tag{2}$$

where $D$ is the synthesis matrix associated to the dictionary $\mathcal{D}$, i.e. every column of $D$ corresponds to an atom in the dictionary.

Solving problem (2) has non-polynomial complexity due to the non convexity of the $\ell_0$ quasi-norm. Two possible approaches that are able to find sparse signal representations over a redundant dictionary are given by the family of greedy algorithms and by Basis Pursuit (BP) [3]. However, only sub-optimal solutions of (2) will normally be recovered with these algorithms. Relevant representatives of the family of greedy algorithms are Matching Pursuit (MP) [4] and Orthogonal Matching Pursuit (OMP) [5]. In these algorithms the atom selection can be influenced by a sub-optimality factor ($\alpha \in (0, 1]$) yielding the well known *Weak*($\alpha$)-MP/OMP [6]. Note that for $\alpha = 1$ these reduce to MP/OMP. The family also includes *Weak*($\alpha$) General MP/OMP in which independent procedures are used for the atom selection and signal approximation [1]. From now on, we will refer to the whole family of *Weak*($\alpha$)-MP/OMP as *Weak*-MP if not otherwise stated.

Very recent results [7], [2], [8], [1] have shown how, under certain conditions, BP and Weak($\alpha$)-MP/OMP are able to recover the optimally sparse solution. Furthermore, sufficient conditions can be established ensuring this optimal behavior if the dictionary is incoherent enough (see Sec. II). However, if the dictionary used has a high coherence $Weak(\alpha)$-MP/OMP and BP will very likely fail to retrieve the optimal representation of $f$.

Note however that BP and pure greedy algorithms are independent of the signal under analysis and do not take its structure into account. Their relationship with the signal is fully driven by the design of the dictionary and waveforms of its atoms. In this report we study the effect of introducing an *a priori* knowledge about the signal in the decomposition. This *a priori* information, that depends on the dictionary as well, can be exploited by considering which atoms of $\mathcal{D}$ are more likely to be used for expanding $f$. A weighting procedure in the previously described algorithms inserts the *a priori* knowledge and leads to the principle of Weighted Basis Pursuit and to a new instance of *Weak*($\alpha$) General Matching Pursuits [1]: Weighted-MP/OMP.

Our main results are:

- The definition of $\mu_1^w(m, \mathcal{D}, f)$, a data dependent measure of the coherence of a dictionary that takes into account the *a priori* information available about the signal. We call this measure *Weighted Babel Function*, for highlighting its relation with the *Babel Function* introduced by *Tropp* in [2].
- The definition of Weighted-BP and Weighted-MP/OMP algorithms. We reformulate classic BP and *Weak*-MP in order to take *a priori* information into account when decomposing the signal.
- A sufficient condition based on $\mu_1^w$, under which Weighted Basis Pursuit and Weighted-MP/OMP find the sparsest signal representation.
- A study of how adapting the decomposition algorithm depending on *a priori* information may help in the recovery of exact sparse representations.
- An analysis of the effects of adding the *a priori* weights on the rate of convergence of *Weak*-MP.

It is important to stress that all the theory we present here can be reduced to the results presented in [2], [8], [1] in the absence of prior information about the signal.

Finally, some examples are shown where the use of the *prior* knowledge is capital for the recovery of the optimal signal representation.

## II. Recovery of Exact Sparse Representations

In this section we summarize very recent theoretical results concerning the possibility for *Weak($\alpha$)*-MP/OMP and BP [1], [2] to exactly recover a given linear combination (1) of $m$ atoms from a redundant dictionary $\mathcal{D} = \{g_j\}_{j\in\Omega}$.

We define $\Gamma$ as the optimal subset of $\Omega$ that indexes the $m$ atoms of the sparse representation (1) and $\overline{\Gamma}$ as the complementary of $\Gamma$ in $\Omega$. Hence, $D_\Gamma$ contains only the linearly independent atoms providing the exact sparsest signal

representation of $f$ and $\mathcal{D} = \mathcal{D}_\Gamma \cup \mathcal{D}_{\overline{\Gamma}}$. We also assume that the cardinality of $\Omega$ is $d$. The dictionary matrix $D$ has size $n \times d$, with $d \geq n$, where $n$ is the size of the input signal $f$.

Since the optimal atoms are not usually known in advance, sufficient conditions for exact recovery are based on the internal coherence of the dictionary. A measure of this coherence is given by the *Babel* function, or cumulative coherence [2], defined as follows:

$$\mu_1(m, \mathcal{D}) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_i, g_\lambda \rangle| , \tag{3}$$

where $\Lambda \subset \Omega$ has size $m$.

Given a signal as in (1), MP/OMP and BP will not necessarily recover the optimal set $\Gamma$. The exact recovery of "correct" atoms will be only ensured if the following *Exact Recovery Condition* (ERC) [2] (also called *Stability Condition* (SC) [1] for MP) is satisfied:

$$\sup_{i \notin \Gamma} \left\| D_\Gamma^+ g_i \right\|_1 < 1, \tag{4}$$

where $(\cdot)^+$ denotes the *Moore-Penrose Pseudoinverse*. In the case of *Weak*-MP [6], the left hand side of (4) is simply replaced by $\alpha$ (see [1], [2]). This bound is indicative of the behavior of general weak greedy algorithms and BP with an over-complete dictionary and a sparse signal. Eq. (4) implies that, in order to recover the optimal functions that expand the signal $f$, these must be different *enough* from any other function of the dictionary not included in $D_\Gamma$. As proved in [1], [2], [8], a second sufficient condition based on the Babel function holds:

***Theorem** 1: (Gribonval and Vandergheynst [1], Tropp [2])* Suppose that $\mu_1$ is the Babel function of $\mathcal{D}$ and $m$ is a positive integer such that

$$\mu_1(m) + \mu_1(m-1) < 1. \tag{5}$$

Then, for any index set $\Gamma$ of size at most $m$ and any $f \in span\,(g_\gamma, \gamma \in \Gamma)$, Eq. (4) holds. This is a sufficient condition for Basis Pursuit to recover the optimal representation of a $(\mathcal{D}, m)$-sparse signal $f$. Moreover, if $\alpha > \mu_1(m)/(1-\mu_1(m-1))$, then *Weak*-MP picks up a correct atom $g \in \Gamma$ at each step.

It is important to stress that Theorem 1 provides a pessimistic bound. There are many cases in which (5) is not respected but indeed MP or BP can find the sparsest solution.

## III. Including *a Priori* Information: Influence on Exact Sparse Representations

The use of redundant dictionaries implies that a signal decomposition is non-unique. This makes it difficult for an algorithm such as *Weak*-MP or BP to recover the sparsest representation. However, as seen in the previous section, this can be theoretically ensured when sufficiently incoherent dictionaries are in use. In this section we prove that, if some valuable *a priori* information about the signal to expand is available, the class of dictionaries where BP and *Weak*-MP are ensured to recover the exact optimal solution can be enlarged. The *a priori* knowledge establishes in advance a likelihood for any atom in the dictionary to appear into the representation of a given signal $f$. This is achieved by suitably weighting the atoms in the dictionary in order to reflect their relevance for the signal $f$.

***Definition** 1:* A weighting matrix $W = W(f, \mathcal{D})$ is a square diagonal matrix of size $d \times d$. Each of the entries $w_i \in (0, 1]$ from the diagonal corresponds to the *a priori* likelihood of a particular atom $g_i \in \mathcal{D}$ to be part of the sparsest decomposition of $f$.

The way $W(f, \mathcal{D})$ should be obtained is particular for each kind of problem and dictionary and will not be treated in this work. In the following we will use $W_\Gamma$ and $W_{\overline{\Gamma}}$ to indicate the diagonal weighting matrices corresponding to $D_\Gamma$ and $D_{\overline{\Gamma}}$ respectively. It is now possible to define a coherence measure equivalent to the Babel function in (3), where *a priori* information is also taken into account: the *Weighted Babel* function. This does not only depend on the dictionary, but also on $f$.

***Definition** 2:* The *Weighted Babel* function of $\mathcal{D}$ is defined as the following data dependent coherence measure:

$$\mu_1^w(m, \mathcal{D}, f) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} | < g_\lambda, g_i > | \cdot w_\lambda \cdot w_i. \tag{6}$$

The *Weighted Babel* function introduces the idea of weighting the correlations among atoms with respect to the *a priori* information we have on $f$. This new coherence measure considers the fact that all functions from the dictionary do not have the same probability to appear in the signal expansion. Indeed, it is of no use to consider in the cumulative coherence measure atoms that are not likely to appear in the representation of a given signal, as they would artificially increase its value.

### A. *Influence of* a Priori *Information on* Weak-*MP*

General Matching Pursuit algorithms [1] iteratively build $n$-terms approximants using a certain rule for the selection of the most appropriate term at every iteration. Every one of these iterations can be seen as a two step procedure:

1) A selection step where an atom $g_{i_k}$ is chosen.
2) An approximation step where an approximant $\hat{f}_n \in span(g_{i_k} \ : \ k \in \{0, ..., n-1\})$ is generated.

The criteria defined to select an atom among all possible candidates is the key point for the retrieval of exact sparse representations. The second step will determine which of the many approximants generated by the selected set of atoms is considered. It is this last step that determines whether MP or OMP is being used.

The selection step can be generally formulated as the maximization of a similarity measure $C(r_n, g_i)$ between the signal to approximate (the residual at the $n^{th}$ iteration: $r_n$) and the atoms from the dictionary:

$$g_{i_n} = \arg\max_{g_i \in \mathcal{D}} C(r_n, g_i). \tag{7}$$

*Weak*-MP uses the scalar product as similarity measure, i.e. $C(r_n, g_i) = |\langle r_n, g_i \rangle|$. This bears some similarity with searching for the atom $g_{i_n}$ with "Maximum Likelihood" given the residual $r_n$: the atom $g_{i_n}$ that maximizes the probability $p(g_i|r_n)$ is selected. Thus, $\langle r_n, g_i \rangle$ is considered as a measure of the conditional probability $p(r_n|g_i)$, and when all $g_i$ are equally probable, maximizing $|\langle r_n, g_i \rangle|$ is equivalent to maximizing $p(g_i|r_n)$ [9].

Let us now consider the case when atoms do not have the same *a priori* probability to appear in the optimal set of atoms $\Gamma$. Indeed, we assume that we have at our disposal a *prior* knowledge about the likelihood of each $g_i$. By means of the Bayes' Rule the probability to maximize becomes

$$p(g_i|r_n) = \frac{p(r_n|g_i)\, p(g_i)}{p(r_n)}, \tag{8}$$

where the denominator is normally characterized by a constant for any signal $r_n$. The selection rule of MP/OMP is thus modified by introducing the multiplication with a weighting factor $w_i \in (0, 1]$ depending on the atom index $i$. We call this family of weighted algorithms *Weighted*-MP/OMP. Let us underline that the approximation step associated to *Weighted*-MP/OMP remains as in the original algorithm. Thus, the residual is updated at every iteration by removing its projection on the selected atom in the case of *Weighted*-MP, and by removing the orthogonal projection of the original signal $f$ on the space generated by the $n$ selected atoms in the case of *Weighted*-OMP. In this work we assume for simplicity that the *prior* knowledge is independent of the iteration of the greedy algorithm.

The following theorem establishes the *Exact Recovery Condition* (ERC) for *Weighted*-MP/OMP.

***Theorem** 2:* Given an *a priori* matrix $W(f, \mathcal{D})$ and a sub-optimality search factor $\alpha \in (0, 1]$, then, for any index set $\Gamma$ such that $f \in span(g_\gamma, \gamma \in \Gamma)$, *Weighted*-MP/OMP will recover a "correct" atom at each iteration if

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| (D_\Gamma W_\Gamma)^+ g_i \cdot w_i \right\|_1 < \alpha. \tag{9}$$

*Proof:* can be interpreted as the probability of a given function to be selected at a that iteration. According to [2], we see that, at every iteration, the following should be satisfied for a $Weak(\alpha)$ greedy algorithm:

$$\rho(r_n) = \frac{\left\| W_{\overline{\Gamma}} \left( D_{\overline{\Gamma}}^T r_n \right) \right\|_\infty}{\left\| W_\Gamma \left( D_\Gamma^T r_n \right) \right\|_\infty} < \alpha, \tag{10}$$

where, as stated previously, $W_\Gamma$, $W_{\overline{\Gamma}}$ are two diagonal sub-matrices of $W(f, \mathcal{D})$ containing the weights $w_i \in (0, 1]$ corresponding to $D_\Gamma$ and $D_{\overline{\Gamma}}$. According to the assumption that $r_n \in \ span(D_\Gamma)$ and that the columns of $D_\Gamma$ are linearly independent, then $r_n = (D_\Gamma W_\Gamma)(D_\Gamma W_\Gamma)^+ r_n = P_\Gamma r_n = P_\Gamma^T r_n$, where $P_\Gamma$ is the orthogonal projector on the space spanned by $D_\Gamma$. This gives:

$$\begin{aligned} \frac{\left\| W_{\overline{\Gamma}} \left( D_{\overline{\Gamma}}^T r_n \right) \right\|_\infty}{\left\| W_\Gamma \left( D_\Gamma^T r_n \right) \right\|_\infty} = \frac{\left\| W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T (D_\Gamma W_\Gamma)(D_\Gamma W_\Gamma)^+ r_n \right\|_\infty}{\left\| W_\Gamma \left( D_\Gamma^T r_n \right) \right\|_\infty} = \\ \frac{\left\| W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left( (D_\Gamma W_\Gamma)^+ \right)^T (D_\Gamma W_\Gamma)^T r_n \right\|_\infty}{\left\| W_\Gamma \left( D_\Gamma^T r_n \right) \right\|_\infty}. \end{aligned} \tag{11}$$

This quantity can be bounded by:

$$\frac{\left\| W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left( (D_\Gamma W_\Gamma)^+ \right)^T (W_\Gamma D_\Gamma^T)\, r_n \right\|_\infty}{\|W_\Gamma (D_\Gamma r_n)\|_\infty} \leq \left\| W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left( (D_\Gamma W_\Gamma)^+ \right)^T \right\|_{\infty,\infty} = \left\| (D_\Gamma W_\Gamma)^+ (D_{\overline{\Gamma}} W_{\overline{\Gamma}}) \right\|_{1,1}. \tag{12}$$

Given that $\|\cdot\|_{1,1}$ is the maximum $\ell_1$ norm of the columns of a matrix, and that the weighting matrices are diagonal, then the ERC is

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| (D_\Gamma W_\Gamma)^+ g_i \cdot w_i \right\|_1 < \alpha, \tag{13}$$

where $w_i$ is the corresponding *a priori* factor of $g_i$ from the diagonal of $W_{\overline{\Gamma}}$.

■

Theorem 2 states, as depicted by (9), that the use of *a priori* weights will help meeting the sufficient condition that guarantees that a Greedy algorithm with a dictionary $\mathcal{D}$ will recover the sparsest representation of $f$. Indeed, as can be observed in (9), given a dictionary and an appropriate $W_\Gamma$ associated to $f$, the weights that multiply each $g_i \in D_{\bar{\Gamma}}$ may help reducing the supremum in (9).

### *B. Influence of* a Priori *Information on BP*

The BP principle selects the signal representation $\mathbf{b}$ that has minimal $\ell_1$ norm. Formally:

$$\min_{\mathbf{b}} \|\mathbf{b}\|_1 \text{ s.t. } \mathcal{D}\mathbf{b} = f. \tag{14}$$

A variation of this algorithm that takes into account the likelihood matrix $W(f, \mathcal{D})$ is given by the Weighted Basis Pursuit (WBP) principle, introduced in [10]. This method minimizes the $\ell_1$ norm of a weighted vector, leaving the constraints unchanged:

$$\min_{\mathbf{b}} \|W^{-1}\mathbf{b}\|_1 \quad \text{s.t.} \quad \mathcal{D}\mathbf{b} = f. \tag{15}$$

We recall that the entries of $W(f, \mathcal{D})$ are in $(0, 1]$. In this way the atoms with low probability to be selected are penalized by inducing a small weighting factor in $W$. It can be proved that WBP can be equivalently reformulated as a Linear Programming problem [10], just as BP.

The following theorem establishes the ERC for Weighted Basis Pursuit. It basically states which is the sufficient condition such that, given the weights $W(f, \mathcal{D})$, WBP is a correct algorithm for recovering an exact sparse superposition of $m$ atoms from $\mathcal{D}$. Let us just point out that, in the following, we will call $\mathbf{b}_{opt}$ the vector giving the optimal signal representation. It thus contains the coefficients corresponding to the functions in $D_\Gamma$ and its size is $m$.

***Theorem** 3:* Given a dictionary $\mathcal{D}$ and an *a priori* matrix $W(f, \mathcal{D})$, Weighted Basis Pursuit recovers the optimal representation of a sparse signal $f = D_\Gamma \mathbf{b}_{opt}$ if:

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| (D_\Gamma W_\Gamma)^+ g_i \cdot w_i \right\|_1 < 1. \tag{16}$$

*Proof:* Suppose that the optimal representation of $f$ is given by $D_\Gamma \mathbf{b}_{opt}$ and that condition (16) is respected. Suppose also that there exists a different representation $f = D_{alt}\mathbf{b}_{alt}$: there should be at least one atom that belongs to $D_{alt}$ but does not appear in $D_\Gamma$. Let us call it $g_x$. What we want to prove is that

$$\left\| W_\Gamma^{-1} \mathbf{b}_{opt} \right\|_1 < \left\| W_{alt}^{-1} \mathbf{b}_{alt} \right\|_1, \tag{17}$$

where $W_{alt}$ is the square diagonal matrix containing the weights corresponding to the atoms in $D_{alt}$ .

$$\left\| W_\Gamma^{-1} \mathbf{b}_{opt} \right\|_1 = \left\| W_\Gamma^{-1} D_\Gamma^+ D_\Gamma \mathbf{b}_{opt} \right\|_1 = \left\| W_\Gamma^{-1} D_\Gamma^+ D_{alt} \mathbf{b}_{alt} \right\|_1 =$$

$$\left\| W_\Gamma^{-1} D_\Gamma^+ D_{alt} W_{alt} W_{alt}^{-1} \mathbf{b}_{alt} \right\|_1 = \left\| (D_\Gamma W_\Gamma)^+ (D_{alt} W_{alt})\, W_{alt}^{-1} \mathbf{b}_{alt} \right\|_1.$$

If the columns of $M = (D_\Gamma W_\Gamma)^+ (D_{alt} W_{alt})$ do not have identical $\ell_1$ norms, using lemma 3.4 in [2] we can state that:

$$\left\| W_\Gamma^{-1} \mathbf{b}_{opt} \right\|_1 < \|M\|_{1,1} \cdot \left\| W_{alt}^{-1} \mathbf{b}_{alt} \right\|_1,$$

but

$$\|M\|_{1,1} = \sup_{g_i \in D_{alt}} \left\| (D_\Gamma W_\Gamma)^+ g_i w_i \right\|_1.$$

There are now two possibilities: either $g_i \in D_\Gamma$ and so the supremum is $\leq 1$, either $g_i \in D_{\overline{\Gamma}}$ and so the supremum is smaller than 1 thanks to (16). In both cases we obtain that (17) is respected.

On the other hand, if all the columns of $M$ have the same $\ell_1$ norm, this must equal $\left\|(D_\Gamma W_\Gamma)^+ g_x w_x\right\|_1$, where $w_x$ is the weight corresponding to $g_x$. Hypothesis (16) ensures that this norm is strictly smaller than 1. So, we can write:

$$\left\|W_\Gamma^{-1}\mathbf{b}_{opt}\right\|_1 \leq \|M\|_{1,1} \cdot \left\|W_{alt}^{-1}\mathbf{b}_{alt}\right\|_1,$$

but this time $\|M\|_{1,1} < 1$. We can therefore conclude that in both cases (17) is valid and so WBP finds the sparsest solution.

■

We thus have a single sufficient condition, valid for both WBP and *Weighted*-MP/OMP, for recovering the "correct" set of atoms involved in the optimal representation of a signal.

## IV. Exact Recovery Bounds for Weighted Greedy and BP Algorithms

Usually the optimal atoms are not known in advance and so the recovery condition of Theorems 2 and 3 can only be verified *a posteriori*, i.e. once the optimal set of atoms has already been found. The following theorem provides a sufficient recovery condition based on the weighted internal coherence ($\mu_1^w$) of the dictionary.

### A. Sufficient Condition for Exact Expansions Recovery

***Theorem 4:*** Let $W(f,\mathcal{D})$ be the data dependent weighting matrix and let $\epsilon_{max} \triangleq \sup_{\gamma\in\Gamma} \left|1 - w_\gamma^2\right|$. If, for any index set $\Gamma$ of size at most $m$, such that $f = \sum_{\gamma\in\Gamma} b_\gamma g_\gamma$, we have

$$\mu_1^w(m) + \mu_1^w(m-1) < 1 - \epsilon_{max}, \tag{18}$$

then (16) holds and WBP recovers the optimal representation of the sparse signal $f$. Furthermore, if

$$\frac{\mu_1^w(m)}{1 - (\mu_1^w(m-1) + \epsilon_{max})} < \alpha \tag{19}$$

is also enforced, then (9) holds and Weighted-*Weak*($\alpha$) MP will pick up an atom belonging to the optimal set $\Gamma$ at each step. Moreover, Weighted-*Weak*($\alpha$) OMP will exactly recover the sparsest representation of $f$.

The *a priori* information can be considered "reliable" when $\epsilon_{max} \ll 1$. Since $\mu_1^w(m) \leq \mu_1(m)$, one can intuitively see that a "reliable" *a priori* knowledge can help a greedy algorithm or BP when the dictionary does not satisfy the hypothesis of Theorem 1. This will be possible when the weights corresponding to the atoms in $D_{\overline{\Gamma}}$ are sufficiently small.

*Proof:* Theorems 2 and 3 give the conditions under which Weighted *Weak*-MP and WBP recover the optimal set of atoms. In this proof the factor $\alpha$ is conserved independently of the algorithm in use. Note that for the particular results of WBP and Weighted-MP/OMP this value equals 1.

Starting from (13) and following the procedure suggested in [2] an upper bound based on $\mu_1^w$ can be obtained:

$$\sup_{g_i\in D_{\overline{\Gamma}}} \left\|(D_\Gamma W_\Gamma)^+ g_i \cdot w_i\right\|_1 =$$

$$\sup_{g_i\in D_{\overline{\Gamma}}} \left\|\left(\left(D_\Gamma W_\Gamma^T\right)^T \left(D_\Gamma W_\Gamma^T\right)\right)^{-1} \left(W_\Gamma D_\Gamma^T\right) g_i \cdot w_i\right\|_1 \leq \tag{20}$$

$$\left\|\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1}\right\|_{1,1} \cdot \sup_{g_i\in D_{\overline{\Gamma}}} \left\|\left(W_\Gamma D_\Gamma^T\right) g_i \cdot w_i\right\|_1.$$

The first term on the right hand side of the inequality corresponds to the $1,1$-norm of the inverse Gram matrix of the weighed sub-dictionary of optimal functions. This can be expressed as:

$$\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1} = (I + A_w)^{-1}, \tag{21}$$

where $I$ denotes the identity matrix and $A_w$ is a symmetric matrix. Due to the diagonal weight matrices $W_\Gamma$, the matrix $A_w$ is not composed only of the off-diagonal elements. Adding and subtracting the identity matrix, we can rewrite (21) in the following way:

$$(I + A_w)^{-1} = \left(I + \left(\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right) - I\right)\right)^{-1}.$$

Akin to [2] this can be expanded by means of *Von Neumann* series [11] and, if $\|A_w\|_{1,1}<1$, we have:

$$\left\|(I+A_w)^{-1}\right\|_{1,1} = \left\|\sum_{k=0}^{\infty}(-A_w)^k\right\|_{1,1} \leq \sum_{k=0}^{\infty}\|A_w\|_{1,1}^k = \frac{1}{1-\|A_w\|_{1,1}}.$$

Thus,

$$\left\|\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1}\right\|_{1,1} \leq \frac{1}{1-\|A_w\|_{1,1}}. \tag{22}$$

The $1,1$-norm of $A_w$ can be expressed as:

$$\|A_w\|_{1,1} = \sup_{g_\gamma \in D_\Gamma}\left[\sum_{l\neq\gamma}|<g_l,g_\gamma>|\cdot w_l\cdot w_\gamma + |1-w_\gamma^2|\right], \tag{23}$$

where the summation comes from the off-diagonal elements and the last term comes from the diagonal part. Note that for convergence of the *Von Neumann* series we need $\|A_w\|_{1,1}<1$. This is ensured by hypothesis since $\|A_w\|_{1,1} \leq \mu_1^w(m-1)+\epsilon_{max}$ and

$$\mu_1^w(m-1)+\epsilon_{max}<1$$

by (18) and (19). From (22) it follows that:

$$\left\|\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1}\right\|_{1,1} \leq \frac{1}{1-(\mu_1^w(m-1)+\epsilon_{max})}. \tag{24}$$

Coming back to Eq. (20), the second term can be bounded as

$$\sup_{g_i\in D_\Gamma}\left\|\left(W_\Gamma D_\Gamma^T\right)g_i\cdot w_i\right\|_1 \leq \mu_1^w(m). \tag{25}$$

Finally, from (24) and (25) we obtain

$$\frac{\mu_1^w(m)}{1-(\mu_1^w(m-1)+\epsilon_{max})}<\alpha, \tag{26}$$

and this proves the theorem. ∎

Since $\mu_1^w(m)\leq\mu_1(m)$, we claim that considering "reliable" *a priori* information can help a dictionary unable to satisfy Theorem 1 recover the right set of functions. That is, "reliable" weights allow for using less incoherent dictionaries.

***Corollary** 1:* Given a dictionary $\mathcal{D}$ and the data dependent diagonal matrix $W(f,\mathcal{D})$, where $w_i\in(0,1]$, we can state the following:

- For a Weighted MP/OMP with weakness $\alpha=1$ and WBP a better behavior in the recovery of exact sparse representations is expected with respect to the classical algorithms if:

$$\mu_1^w(m)+\mu_1^w(m-1)<1-\epsilon_{max} \quad \text{and} \quad \mu_1(m)+\mu_1(m-1)\geq 1.$$

- For a Weighted *Weak*-MP a better behavior in the recovery of exact sparse representations is expected with respect to the classical algorithms if:

$$\frac{\mu_1^w(m)}{1-(\mu_1^w(m-1)+\epsilon_{max})}<\alpha \quad \text{and} \quad \frac{\mu_1(m)}{1-\mu_1(m-1)}\geq\alpha.$$

***Corollary** 2:* When no *a priori* information is available (i.e. $W(f,\mathcal{D})=I$), and consequently $\epsilon_{max}=0$ Theorem 4 boils down to the results found by *Gribonval, Vandergheynst* [1] and *Tropp* [2] stated in Theorem 1.

*B. Examples*

*1) A Toy Example for MP in $\mathbb{R}^3$:* Let us consider the following overcomplete dictionary in $\mathbb{R}^3$:

$$D=\begin{pmatrix} 0 & -0.9806 & 0.4472 & -0.5774 \\ 1 & -0.1961 & 0 & 0.5774 \\ 0 & 0 & 0.8944 & -0.5774 \end{pmatrix}. \tag{27}$$

A simple $m$-sparse signal $f$ is considered with $m=2$ and defined as:

$$f=3\cdot D_0+3.059412\cdot D_1, \tag{28}$$

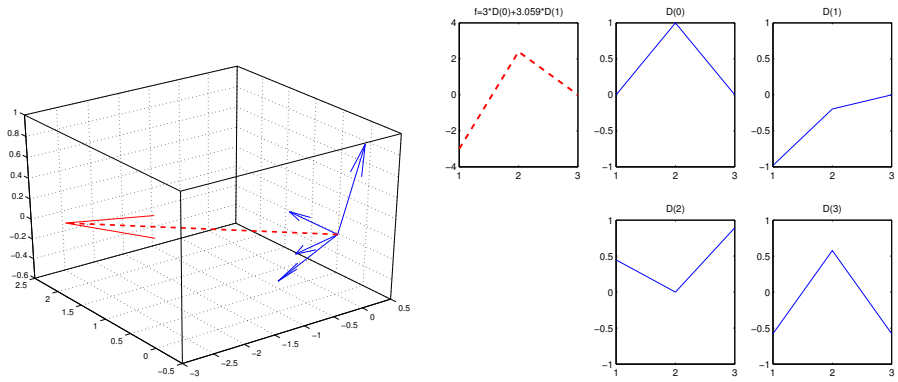


Fig. 1. Left: 3D representation of the overcomplete dictionary (4 components) and the sparse signal $f$ in $\mathbb{R}^3$. Right: Temporal representation of the signal and dictionary atoms.

i.e. the optimal set is $\Gamma = \{D_O, D_1\}$. A general graphical representation of $\mathcal{D}$ and $f$ in $\mathbb{R}^3$ can be observed in Fig. 1 where the non-orthogonality among vectors can be clearly appreciated.

According to the coherence measure $\mu_1$, this dictionary has a high coherence, i.e. $\mu_1(1) = 0.7746$. This turns into a complete failure of the sufficient condition (5). Indeed, $\mu_1(2) + \mu_1(1) = 2.1265$ which is far above the bound with $\alpha = 1$ required to guarantee the recovery of the optimal set of atoms for any $f$.

As a consequence, MP "derails". . The sequence of atoms selected from the dictionary for pure MP is:

```
MP:
Step   1:  select =     3       Step   6:  select =     1
Step   2:  select =     2       Step   7:  select =     2
Step   3:  select =     1       Step   8:  select =     0
Step   4:  select =     0       Step   9:  select =     1
Step   5:  select =     2       Step  10:  select =     2
```

where the selected 0, 1, 2, 3 are the indexes of $D_i$.

Let us now consider the possibility that, by some means, it is feasible to estimate that the signal $f$ has around 60% of chances to be embedded in the $x - y$ plane. This implies that the scalar products by the vectors $D_2$ and $D_3$ can be penalized. Thus, the following weighting matrix can be generated:

$$W(f, \mathcal{D}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.6 \end{pmatrix}.$$

Notice the assumption that our *oracle* does not penalize the two vectors implied in the sparsest representation of $f$ ($w_i = 1 \ : \ i = 0, 1$).

Shifting now to the framework of Weighted-MP, the *Weighted Babel* coherence measure indicates that the effective internal coherence of the dictionary is reduced up to $\mu_1^w(2) = 0.3464$. Moreover the new bound, considering the *a priori*, reads $\mu_1^w(2) + \mu_1^w(1) = 0.9717$, meeting the sufficient requirement to ensure the recovery of the optimal set of vectors $\Gamma$. This time, the sequence of atoms selected is quite different and the Weighted-MP algorithm selects only the atoms belonging to the optimal set:

```
Weight-MP:
Step   1:  select =     1       Step   6:  select =     0
Step   2:  select =     0       Step   7:  select =     1
Step   3:  select =     1       Step   8:  select =     0
Step   4:  select =     0       Step   9:  select =     1
Step   5:  select =     1       Step  10:  select =     0
```

Tests on these examples have been performed with the BP and WBP paradigm as well. For this particular case, however, both are able to recover the optimal set of atoms independently of the fact that for BP the sufficient condition of Theorem 1 was not fulfilled.

*2) A Toy Example for BP in $\mathbb{R}^5$:* Let us now illustrate Theorems 3 and 4 in the Basis Pursuit case. Suppose we have a signal $f = [0, M, A, M, 0]' \in \mathbb{R}^5$ depicted in Figure 2 and we want to decompose it with BP over the following dictionary $\mathcal{D} = \{g_i\}_{i=1,...,10}$:

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}. \tag{29}$$

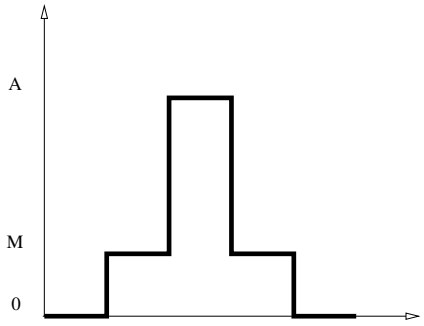


Fig. 2. Signal $f \in \mathbb{R}^5$ to decompose over $\mathcal{D}$

The signal $f$ has, of course, multiple representations over $\mathcal{D}$; let us focus on two of them, setting $M = 1$ and $A = 3$:

$$\begin{aligned} f &= (g_3, g_{10}) \cdot \begin{pmatrix} 3 \\ \sqrt{2} \end{pmatrix} = D_\Gamma \cdot \mathbf{b}_{opt} \\ &= (g_3, g_7, g_8) \cdot \begin{pmatrix} 1 \\ \sqrt{2} \\ \sqrt{2} \end{pmatrix} = D_{alt} \cdot \mathbf{b}_{alt}. \end{aligned} \tag{30}$$

Computing (5) for $m = 2$ we obtain a value of around $3.1 > 1$. Hence, Theorem 1 does not apply and we have no guarantee that BP will find the sparsest solution. In fact, BP selects the second representation in (30) which has a smaller $\ell_1$ norm. If we insert now a diagonal weighting matrix $W$ with the non-zero elements equal to $w_i, i = \{1, ..., 10\}$, it is easy to verify that WBP selects the sparsest solution with the following probability weights:

$$w_i = \begin{cases} 1 & \text{if } i = 3, 10 \\ v < \frac{2}{1+\sqrt{2}} & otherwise \end{cases}. \tag{31}$$

This solution gives $\epsilon_{max} = 0$. Note that setting $v$ slightly smaller than $\frac{2}{1+\sqrt{2}}$ the value of $\mu_1^w(2) + \mu_1^w(1)$ is bigger than 1, nevertheless the solution found is the optimal one. This clearly shows that the sufficient condition offered by Theorem 4 is pessimistic. On the other hand if $v$ is smaller (e.g. 0.45) we obtain $\mu_1^w(2) + \mu_1^w(1) \simeq 0.95 < 1$. In this new case the hypothesis of Theorem 4 is respected.

A last remark about this toy example can be done observing that it is not necessary that the weights corresponding to the optimal basis functions are exactly one. For example WBP is able to find the sparsest solution even with the following weights:

$$w_i = \begin{cases} r \leq 1 & \text{if } i = 3, 10 \\ v < r \cdot \frac{2}{1+\sqrt{2}} & otherwise \end{cases}. \tag{32}$$

For example, setting $r = 0.9$ and $v = 0.4$ we obtain $\epsilon_{max} = 0.19$ and $\mu_1^w(2) + \mu_1^w(1) \simeq 0.76 < 1 - \epsilon_{max}$. Therefore the hypothesis of Theorem 4 is again respected. Note that, for this particular example, both MP and weighted-MP are able to recover the optimal subset of functions, even if the former does not satisfy the sufficient condition of Theorem 1.

## V. Rate of Convergence of Weighted-MP/OMP

### A. Theoretical Rate of Convergence

To find a bound on the rate of convergence of Weighted-MP/OMP we follow the path of [1] and [12] where the respective authors look for an equivalent result for the case of *Weak*-MP in the former, and for the particular case of a block based dictionary in the latter. Simply looking to the results found in [1] it is intuitively clear that the knowledge

of some *a priori* information should allow for a better bound on the rate of convergence of the representation. Indeed, in the convergence of the *Weak*-MP algorithm the *Babel* function [2] appears as a determining factor that drives the speed of exponential decay. Given the fact that $\mu_1^w(m) \leq \mu_1(m)$, we consider that having some *a priori* knowledge contributes to determine a lower bound on the exponentially decaying rate of convergence associated to Weighted-MP/OMP.

***Theorem** 5:* Let $W(f, \mathcal{D})$ be the data dependent weighting matrix that introduces *a priori* knowledge in $\mu_1^w(m)$. Let $m$ be an integer such that:

$$\frac{\mu_1^w(m)}{1 - (\mu_1^w(m-1) + \epsilon_{max})} < \alpha. \tag{33}$$

Then for any subset $\mathcal{D}_\Gamma \subset \mathcal{D}$ with $|\mathcal{D}_\Gamma| \leq m$, and any $f \in span(\mathcal{D}_\Gamma)$, Weighted-MP/OMP picks up only "correct" atoms at each step and

$$\|r_{n+1}\|^2 \leq \|f\|^2 \left(1 - \alpha^2 \frac{(1 - \mu_1^w(m-1) - \epsilon_{max})}{m}\right)^{n+1}. \tag{34}$$

Let us consider first two preliminary lemmas that will be used in the proof of this theorem. These correspond to those used in the methodology appearing in [1]. However, particular considerations are taken into account to adapt them to the case where *a priori* information is used.

***Lemma** 1:* Consider an optimal set $\Gamma$ with $|D_\Gamma| = m$ associated to the exact sparse expansion of signal $f$ and the "reliable" *a priori* knowledge weighting matrix $W(f, \mathcal{D})$ (and the $W_\Gamma$ sub-matrix). Then, the square singular values $(\sigma^2_{min_w})$ of the matrix $(D_\Gamma W_\Gamma)$ are such that:

$$\sigma^2_{min_w} \geq 1 - \mu_1^w(m-1) - \epsilon_{max}. \tag{35}$$

*Proof:* Consider the gram matrix $G \triangleq (D_\Gamma W_\Gamma)^T (D_\Gamma W_\Gamma)$, then the singular values $\sigma^2_{k_w}$ are the eigenvalues $(\lambda_k)$ of $G$. From the Geršgorin Disk Theorem [11] an upper bound on the eigenvalues of $\lambda_k$ can be drawn in the way performed in [1], [12], [2], [7], [13], [14]. This shows that every eigenvalue of $G$ lies in one of the m disks

$$\Delta_k = \left\{ z : |G_{kk} - z| \leq \sum_{j \neq k} |G_{jk}| \right\}. \tag{36}$$

Hence, since $\sum_{j \neq k} |G_{jk}| = \sum_{j \neq k} |< w_j \cdot g_j, w_k \cdot g_k >| \leq \mu_1^w(m-1)$ then:

$$|G_{kk} - \lambda_k| \leq \mu_1^w(m-1), \tag{37}$$

where $G_{kk} \geq 1 - \epsilon_{max}$ and since $\mu_1^w(m-1) + \epsilon_{max} < 1$,

$$\sigma^2_{min_w} \geq 1 - \epsilon_{max} - \mu_1^w(m-1). \tag{38}$$

■

Note that if $\epsilon_{max} \ll 1$, then $\sigma^2_{min_w} \gtrsim 1 - \mu_1^w(m-1)$, which mimics the result of classic *Weak*-MP [1]

***Lemma** 2:* For any index set $\Gamma$ of $|\Gamma| = m$, the corresponding data dependent weighting matrix $W_\Gamma$ and a coefficients vector $\mathbf{b}$,

$$\sup_{\gamma \in \Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >| \geq \frac{\|D_\Gamma \mathbf{b}\|^2}{\left\|W_\Gamma^{-1} \mathbf{b}\right\|_1}, \tag{39}$$

and given a residual $r_n = f - f_n$ such that $r_n \in span\,(g_\gamma, \gamma \in \Gamma)$ and the smallest square singular value of $D_\Gamma$ $(\sigma^2_{min_w})$ then,

$$\frac{\sup_{\gamma \in \Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >|}{\|r_n\|} \geq \sqrt{\frac{\sigma^2_{min_w}}{m}}. \tag{40}$$

For the sake of clarity of the section, the proof of this lemma is included in the Appendix.

Using the previous lemmas, let us finally prove the result depicted in Theorem 5:

*Proof:* Let $r_{n+1} = f - f_n$ be the residual of the Weighted-MP/OMP algorithm at the $n$th iteration, then it is known that:

$$\|r_{n+1}\|^2 \leq \|r_n\|^2 - \left|< r_n, g_{\gamma_{n+1}} >\right|^2, \tag{41}$$

where the inequality applies for OMP, while for the case of MP the equality holds. In our case the selection of $g_{\gamma_{n+1}}$ is driven by $W(f,\mathcal{D})$, i.e.

$$\left|\left\langle r_n, g_{\gamma_{n+1}}\right\rangle\right| = \alpha \cdot \frac{1}{w_\gamma} \sup_\gamma \left|\langle r_n, g_\gamma \cdot w_\gamma\rangle\right|. \tag{42}$$

Hence,

$$\begin{aligned} \|r_{n+1}\|^2 \leq & \quad \|r_n\|^2 - \alpha^2 \cdot \frac{1}{w_\gamma^2} \sup_\gamma |< r_n, g_\gamma \cdot w_\gamma >|^2 \\ & \leq \quad \|r_n\|^2 \left(1 - \alpha^2 \frac{\frac{1}{w_\gamma^2} \sup_\gamma |< r_n, g_\gamma \cdot w_\gamma >|^2}{\|r_n\|^2}\right). \end{aligned} \tag{43}$$

Then, from Eqs. (40), (35) and given $w_\gamma \leq 1$, it follows:

$$\begin{aligned} \|r_{n+1}\|^2 \leq & \quad \|r_n\|^2 \left(1 - \alpha^2 \frac{\sigma^2_{min_w}}{w_\gamma^2 m}\right) \\ & \leq \quad \|r_n\|^2 \left(1 - \alpha^2 \frac{\sigma^2_{min_w}}{m}\right) \\ & \leq \quad \|r_n\|^2 \left(1 - \alpha^2 \frac{1 - \mu_1^w(m-1) - \epsilon_{max}}{m}\right) \\ & \leq \quad \|f\|^2 \left(1 - \alpha^2 \frac{1 - \mu_1^w(m-1) - \epsilon_{max}}{m}\right)^{n+1}. \end{aligned} \tag{44}$$

■

Thus, since $\mu_1^w(m-1) \leq \mu_1(m-1)$ and assuming $\epsilon_{max}$ to be small enough, a faster rate of convergence is reached.

### *B. A Toy Example for Weighted-MP and MP*

To illustrate the theoretical result found in this section, we go back to the toy example presented in sec. IV-B.1 where an overcomplete coherent Dictionary in $\mathbb{R}^3$ is used. As can be expected from Theorem 4 and Theorem 5 and observed in Fig. 3, the rate of convergence of Weighted-MP shows a much faster decay of the error energy than classical MP. Indeed, as guaranteed by the sufficient condition (18) and as illustrated in Sec. IV-B.1, the Weighted-MP algorithm gets trapped selecting over and over only vectors from the optimal set $\Gamma$. This avoids introducing spurious terms in the signal expansion and allows a faster exponential convergence than in the pure greedy case.

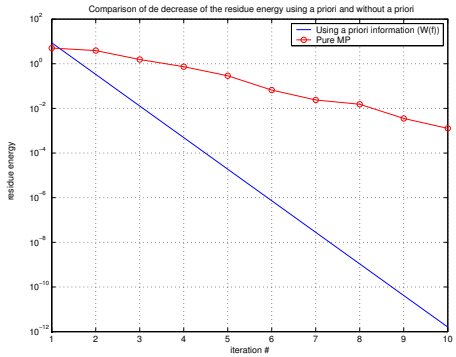


Fig. 3. Convergence of the approximation error of the example of Fig. 1. The respective rates with and without using weights are compared. The use of weights enhances the asymptotic rate of convergence.

## VI. Examples

This section just offers a more complex example than those appearing in previous sections. Some experiments on retrieving the sparsest signal representation and the sparsest approximation using a redundant dictionary are shortly presented. Both weighted and classical approaches are used.

### *A. Heuristics in a coherent dictionary: Use of Footprints*

Let us explore the representation of piecewise-smooth signals and the use of dictionaries composed by the mixture of an orthonormal wavelet basis and a family of wavelet footprints (see *Dragotti* in [15]). Wavelet footprints are the functions composed by all wavelet coefficients that a given singularity generates on a orthonormal basis or frame as illustrated in Fig. 4.

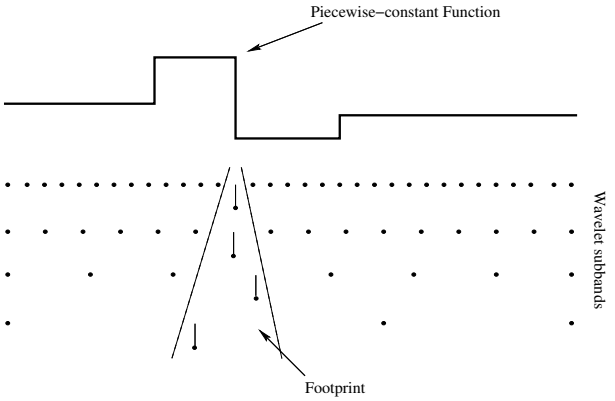


Fig. 4. Wavelet Footprints description scheme for a piecewise-constant signal [15].

In this example, $f$ is a 1-D signal with 128 samples such that this can be sparsely represented describing the singularities by means of footprints. We assume that the family of wavelets in use has a sufficiently high number of vanishing moments such that polynomial parts of the signal are efficiently represented by the coefficients of the scaling functions. Moreover the set of discontinuities appearing in the signal are also contained in the dictionary in the form of footprints. For the sake of simplicity, we consider a piecewise constant signal $f$ (see Fig. 8). The dictionary is defined by the union of an orthonormal basis defined by the *symmlet*-4 family of wavelets [16] and the respective family of footprints for all possible translations of the Heaviside function. The later is used to model the piecewise constant discontinuities. The graphical representation of the dictionary matrix can be seen in Fig. 5 where the columns are the waveforms that compose the dictionary.

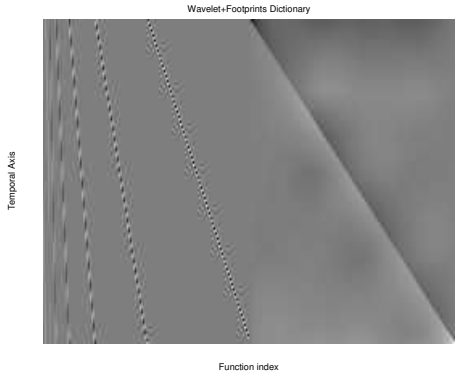


Fig. 5. Dictionary formed by the Symmlet-4 [16] (left half) and its respective footprints for piecewise constant singularities (right half).

The overcompleteness of the dictionary is evident: the number of atoms is twice the dimension of the signal. In spite of its simplicity, the dictionary presents a very high coherence factor $\mu_1(1) = 0.9606$. It is indeed very difficult for such a dense dictionary to fulfill the bounds of Theorem 1. For example, for m=3, $\mu_1(3) + \mu_1(2) = 4.7664$, which is already quite far from the required upper bound. In this example the optimal subset that represents the signal $f$ has size $m = 9$.

The signal $f$ has been selected such that footprint components are close enough to strongly interact. If they were not overlapping, then any pure greedy or BP algorithm would be able to recover the good representation without problems, given their orthogonality.

The weights of $W(f, D)$ are estimated from the data. This is done following a simple procedure inspired from [15]. This somehow tries to estimate the location of footprints and to penalize those wavelets that overlap with the footprints location. The detailed procedure is depicted in Algorithm 1.

The resulting vector of weights from the diagonal of $W(f, D)$ is shown in Fig. 6. Notice the four spikes in the right part of Fig. 6. These point the index of the footprint functions that are more likely to be components of $f$. All the

**Algorithm 1** $W(f, D)$ estimation

**Require:** $\mathcal{D} = \mathcal{D}_{Symmlet} \cup \mathcal{D}_{Footprints}$, define a threshold $\lambda$ , define a penalty factor $\beta$

1: $f_{diff} = D^{+}_{Footprints} \cdot f$ {Footprints location estimation (edge detection)}
2: Threshold $f_{diff}$ by $\lambda$ putting greater values to 1 or $\beta$ otherwise.
3: $W^{diag}_{footprints} = f_{diff}$ {Diagonal of the sub-matrix of $W(f, D)$ corresponding to footprints.}
4: Create $W^{diag}_{wave}$ s.t. all wavelets intersecting the found footprints locations equal $\beta$, set to 1 otherwise.
5: $W(f, D) = diag\left(\begin{bmatrix} W^{diag}_{wave} & W^{diag}_{footprints} \end{bmatrix}\right)$;

spikes in the left part correspond to the wavelet function indexes that interact with the location of the more probable footprint functions. This weights are obtained setting the values of the parameters $\lambda$ and $\beta$ in Algorithm 1 to 0.7 and 0.6 respectively.

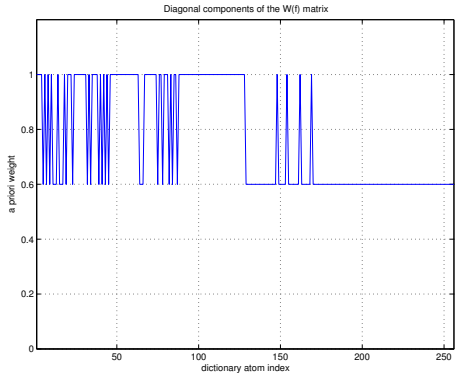


Fig. 6. Weights involved on introducing the a priori information to drive OMP.

The effect of applying the weights is reflected in the Gramm matrix representations of $D$ and $D \cdot W$ in Fig. 7. A reduction on the strength of interference between the dictionary atoms can be observed in the Gramm matrix of the weighted dictionary.

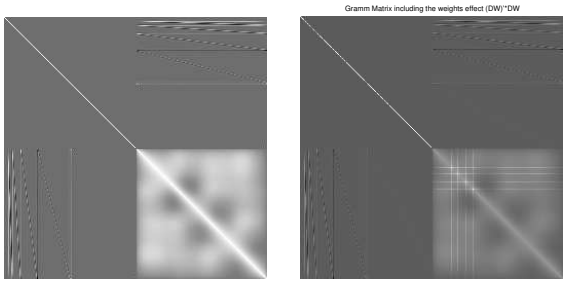


Fig. 7. Left: Representation of the Gramm matrix (i.e. $D^T \cdot D$) of the combined wavelet-footprints dictionary of Fig. 5. It clearly depicts the cross products between the different atoms. The upper left side perfectly describes the orthogonality of the symmlet basis. At the bottom right a sketch of the high coherence among the footprints. Right: Representation of the Gramm matrix after applying weights. Notice the reduction of cross-interferences.

To the contrary of what the reader could expect now, we are not able to say that given an *a priori* information the sufficient conditions defined previously in this paper are satisfied. Indeed, the signal singularities are so close that their optimal atoms are not incoherent enough to allow the summation $\mu^w_1(m) + \mu^w_1(m-1)$ to be smaller than one. Despite that, we are be able to say that the use of Weighted-OMP (MP and Weighted-MP fail in any case) and Weighted Basis Pursuit helps recovering the optimal representation. This illustrates the intuitive idea that *a priori* information may help the signal representation even if the sufficient conditions of Theorem 4 are not satisfied.

The comparative results of representation by means of OMP and Weighted-OMP can be seen Fig. 8. The effect of the *a priori* knowledge to recover the optimal representation is obvious (first picture on the left). The high coherence of the dictionary makes the non-weighted algorithm select wavelet bases when it should not.

A global view of the impact of using the *a priori* information is presented in Fig. 9. Weighting is able to keep OMP on the track for the recovery of the exact-sparse representation unlike classical OMP.

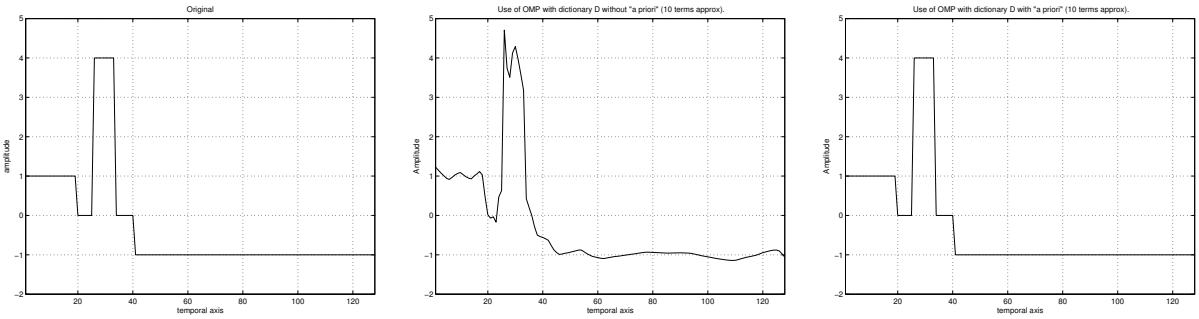


Fig. 8. Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 5). Left: Original signal. Middle: blind OMP approximation. Right: (only 9 terms are different from 0) OMP with prior knowledge of the footprints location.

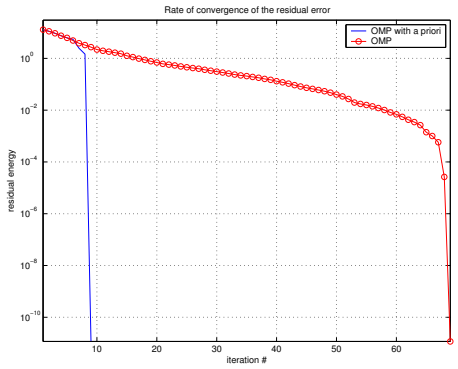


Fig. 9. Rate of convergence of the error with respect to the iteration number in the experiment of fig. 8

Decomposing the same input signal with BP and WBP we observe that the latter is able to recover the four footprints and the wavelet scaling functions participating in the signal expansion, while the former can only recover three out of four footprints and uses the wavelet functions to represent the other discontinuity. Fig. 10 shows the coefficients of the signal representation $\mathbf{b}$ obtained by BP ($\mathbf{b}_{BP}$) and WBP ($\mathbf{b}_{WBP}$). This situation illustrates again how the use of the weights can help in recovering the sparsest signal decomposition. Note that while $\|\mathbf{b}_{BP}\|_1 < \|\mathbf{b}_{WBP}\|_1$, we have that $\|\mathbf{b}_{BP}\|_0 > \|\mathbf{b}_{WBP}\|_0$.

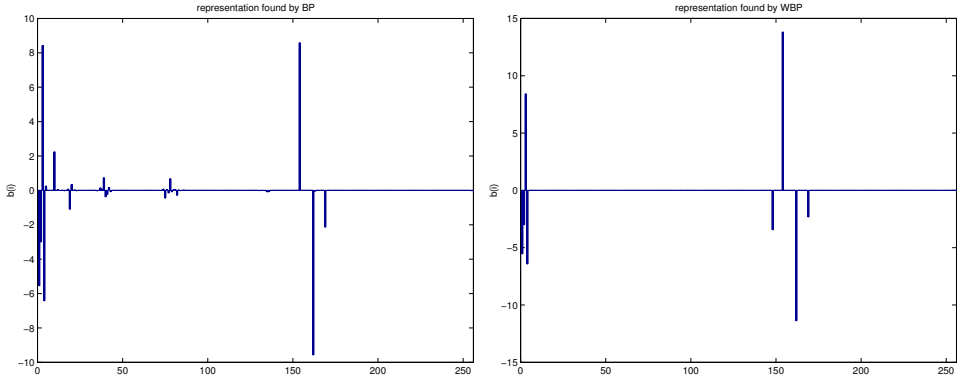


Fig. 10. Coefficients of the signal representations obtained by BP and WBP. The index $i$ ranges from 1 to the size of the dictionary. The first half of the $\mathcal{D}$ is composed by wavelets, the rest by footprints, as in Figure 5

### *B. Heuristics in a coherent dictionary: Use of Footprints and $\epsilon$-Sparse Approximations.*

In natural signals, it is unlike to find examples where exact sparse representations are possible: as pointed out in [2] the set of such signals has measure zero. Thus, some additional theoretical considerations are required for the problems related to sparse approximations of signals. In Figs. 11 and 12 the problem formulated in the previous subsection is reconsidered for approximation of piecewise-smooth signals with higher order polynomials (higher than the 4 vanishing moments of the *symmlet*-4 in use for our experiences. The use of weights is definitively helpful for the rate of convergence

of the signal approximation with few coefficients and a considerable gain in the reduction of the approximation error is achieved. The approximation properties of Weighted-MP/OMP and of Weighted-BP will be discussed in a forthcoming paper [17].

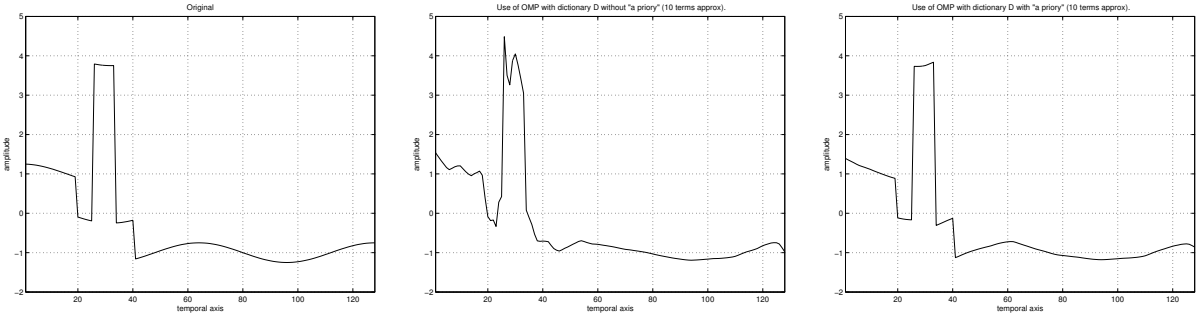


Fig. 11. Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 5). Left: Original signal. Middle: "blind" OMP approximation. Right: OMP with prior knowledge of the footprints location.

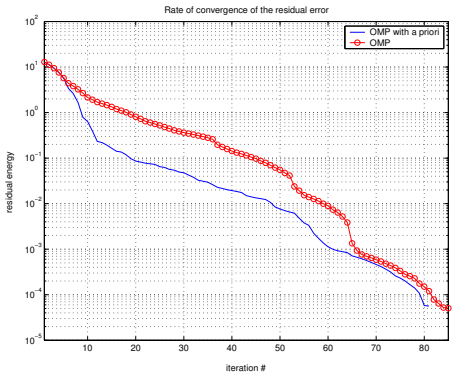


Fig. 12. Rate of convergence of the error with respect to the iteration number in the experiment of Fig. 11

## VII. Conclusions

Suppose one wants to decompose a signal over a redundant dictionary and aims at finding its sparsest representation. Suppose the signal does present a structure that can help in the decomposition and has to be preserved. It is intuitive that exploiting the knowledge we have about the signal may be helpful for its analysis, conditioned to the fact that the information we have is correct. This information can be exploited, for example by assigning a kind of "probability" to atoms based on their ability to catch some important features of the signal. This is precisely what is done by the weighting matrix $W$.

The problem of how to find a reliable *a priori* information is not addressed in this work. We present theoretical results that show when and how this information can help recovering the optimal, sparsest representation of a $m$-sparse signal. Weights computation is an open question and strongly depend on the dictionary and the class of signals to represent. It is closely related to estimation and signal analysis. An example of how to compute the weights for some class of signal is given in [10], where the WBP algorithm is used.

Based on the experimental results, we claim as well that even if sufficient conditions on the recoverability of a $m$-sparse signals are not fulfilled, *a priori* information can still contribute to the recovery of a better representation.

The use of *a priori* information can be seen as a way of adaptively reducing the size of a very redundant dictionary removing undesired candidates in the selection process. In this optic we can understand how weighted algorithms can allow the use of more coherent dictionaries. We want to underline that the *a priori* must not be necessarily applied under the form of weights. In this work weights have been introduced for simplicity in the theoretical calculations. *A priori* information can equally appear as the application of a model or the inclusion of additional constraints on the desired solution [18].

It is important to observe that all the theoretical results we present here reduce to the classical case of sparse recovery when $W = I$, i.e. no *a priori* information is available.

Finally, the results we presented in this work concern only the exact recovery of a signal. An interesting evolution of this study can be the extension to the approximation case, as briefly shown in section VI-B. This is the central topic of [17].

## Appendix

Proof of Lemma 2 of Section V.

*Proof:* To prove this lemma, we just need to follow the procedure appearing in [1], [12] which uses results from *DeVore and Temlyakov* [19].

$$\begin{aligned} \|D_\Gamma \mathbf{b}\|^2 = & \left\langle D_\Gamma \mathbf{b}, D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b} \right\rangle \\ = & \sum_{\gamma \in \Gamma} \frac{b_\gamma}{w_\gamma} \left\langle g_\gamma \cdot w_\gamma, D_\Gamma \mathbf{b} \right\rangle \\ \leq & \sum_{\gamma \in \Gamma} \left| \frac{b_\gamma}{w_\gamma} \right| \left| \langle D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma \rangle \right| \\ \leq & \left\| W_\Gamma^{-1} \mathbf{b} \right\|_1 \sup_{\gamma \in \Gamma} \left| \langle D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma \rangle \right|. \end{aligned} \tag{45}$$

For the final proof of the lemma two additional results are needed.

- By the *Jensen's Inequality* [20] $\left\| W_\Gamma^{-1} \mathbf{b} \right\|_1$ can be bounded as

$$\left\| W_\Gamma^{-1} \mathbf{b} \right\|_1^2 \leq m \cdot \left\| W_\Gamma^{-1} \mathbf{b} \right\|_2^2. \tag{46}$$

  In fact:

$$\begin{aligned} \left\| W_\Gamma^{-1} \mathbf{b} \right\|_1^2 = & \left( \sum_{i=0}^{m-1} \left| \frac{b_i}{w_i} \right| \right)^2 = m^2 \left( \sum_{i=0}^{m-1} \frac{|b_i|}{m \cdot w_i} \right)^2 \\ \leq & \; m^2 \sum_{i=0}^{m-1} \left| \frac{b_i}{w_i} \right|^2 \frac{1}{m} \leq m \cdot \left\| W_\Gamma^{-1} \mathbf{b} \right\|_2^2. \end{aligned} \tag{47}$$

- By means of the *Singular Value Decomposition* [21] any $\left\| W_\Gamma^{-1} \mathbf{b}_n \right\|_2^2$ (where $n$ indicates the iteration number) can be bounded as

$$\frac{\|r_n\|^2}{\sigma_{min_w}^2} \geq \left\| W_\Gamma^{-1} \mathbf{b}_n \right\|_2^2. \tag{48}$$

  This is proved by:

$$\begin{aligned} \|r_n\|^2 = & \left\| D_\Gamma \mathbf{b}_n \right\|^2 = \mathbf{b}_n^T \left( D_\Gamma W_\Gamma W_\Gamma^{-1} \right)^T D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_n \\ = & \; \mathbf{b}_n^T \left( W_\Gamma^{-1} W_\Gamma D_\Gamma^T \right) D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_n \\ = & \; \mathbf{b}_n^T W_\Gamma^{-1} \left( D_\Gamma W_\Gamma \right)^T D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_n \\ = & \; \mathbf{b}_n^T W_\Gamma^{-1} \left( U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T \right)^T \left( U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T \right) W_\Gamma^{-1} \mathbf{b}_n \\ = & \; \mathbf{b}_n^T W_\Gamma^{-1} \left( V_{\Gamma_W} \Sigma_{\Gamma_W}^T U_{\Gamma_W}^T \right) \left( U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T \right) W_\Gamma^{-1} \mathbf{b}_n, \end{aligned} \tag{49}$$

  where $U_{\Gamma_W}$ and $V_{\Gamma_W}$ are orthonormal matrices and $\Sigma_{\Gamma_W}$ is a diagonal matrix such that

$$diag\left( \Sigma_{\Gamma_W} \right) = \left( \sigma_{0_w}, \sigma_{1_w}, ..., \sigma_{k_w}, ..., \sigma_{m_w} \right).$$

  From now on consider $\mathbf{y} = V_{\Gamma_W}^T W_\Gamma^{-1} \mathbf{b}_n$. Therefore,

$$\begin{aligned} \|r_n\|^2 = & \; \mathbf{b}_n^T W_\Gamma^{-1} V_{\Gamma_W} \Sigma_{\Gamma_W}^2 V_{\Gamma_W}^T W_\Gamma^{-1} \mathbf{b}_n \\ = & \; \mathbf{y}^T \Sigma_{\Gamma_W}^2 \mathbf{y} = \sum_{k=0}^{m-1} \sigma_{k_w}^2 \cdot y_k^2 \\ \geq & \; \sigma_{min_w}^2 \|\mathbf{y}\|^2 = \sigma_{min_w}^2 \left\| W_\Gamma^{-1} \mathbf{b}_n \right\|^2. \end{aligned} \tag{50}$$

Thus, finally from (46) and (48) it follows

$$\frac{\|r_n\|^2}{\sigma^2_{min_w}} \geq \frac{\|W_\Gamma^{-1}\mathbf{b}_n\|_1^2}{m}, \tag{51}$$

that jointly with (45) gives the result stated by Lemma 2:

$$\frac{\sup_{\gamma\in\Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >|}{\|r_n\|} \geq \sqrt{\frac{\sigma^2_{min_w}}{m}}. \tag{52}$$

■

## Acknowledgments

We would like to thank Rosa M. Figueras i Ventura and Lorenzo Peotta for fruitful discussions.

## References

[1] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," IRISA, Rennes, France, Tech. Rep., 2004.
[2] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," ICES, University of Texas at Austin, Austin, USA, Tech. Rep., 2003.
[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
[4] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Proc.*, vol. 41, no. 12, pp. 3397–3415, December 1993.
[5] Y. Pati, R. Reziifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition."
[6] V. N. Temlyakov, "Weak greedy algorithms," Department of Mathematics, University of South Carolina, Columbia, Tech. Rep., 1999.
[7] D. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via $\ell_1$ minimization," *Proc. Nat. Aca. Sci.,*, vol. 100, no. 5, pp. 2197–2202, March 2003.
[8] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," ICES, University of Texas at Austin, Austin, USA, Tech. Rep., 2004.
[9] V. Van Trees, *Detection, Estimation, and Modulation Theory - Part I.* Jonh Wiley and Sons, 1967.
[10] L. Granai and P. Vandergheynst, "Sparse decomposition over multi-component redundant dictionaries," in *Proc. of Multimedia Signal Processing, Workshop on. MMSP04*, September 2004, to appear on.
[11] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge Univ. Press, 1985.
[12] L. Peotta and P. Vandergheynst, "MP in block quasi-incoherent dictionaries," Signal Processing Institute, EPFL, Lausanne, Switzerland, Tech. Rep., 2003.
[13] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit." *IEEE Trans. in Signal Processing.*, vol. 1, no. 51, January 2003.
[14] A. Gilbert, S. Muthukrishnan, and M. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *14th ACM-SIAM Symposium on Discrete Algorithms (SODA'03)*, January 2003.
[15] P. Dragotti and M. Vetterli, "Wavelet footprints: Theory, algorithms and applications," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1306–1323, May 2003.
[16] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.
[17] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of *a Priori* information for sparse signal approximations," ITS/LTS-2 EPFL, Tech. Rep., in Preparation.
[18] J. L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," CEA-Saclay, DAPNIA/SEDI-SAP, Tech. Rep., 2004.
[19] R. DeVore and V. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 2-3, no. 5, pp. 173–187, 1996.
[20] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products.* San Diego, CA: Academic Press, 2000.
[21] G. Golub and C. V. Loan, *Matrix Computations.* Baltimore,MD: Johns Hopkins University Press, 1996.