Approximation with highly redundant dictionaries

Rémi Gribonval[a] and Morten Nielsen[b]

[a] IRISA, Rennes, France
[b] Aalborg University, Aalborg, Denmark

Revised on Sept. 5, 2003.

ABSTRACT

Ten years ago, Mallat and Zhang proposed the Matching Pursuit algorithm : since then, the dictionary approach to signal processing has been a very active field. In this paper, we try to give an overview of a series of recent results in the field of sparse decompositions and nonlinear approximation with redundant dictionaries. We discuss sufficient conditions on a decomposition to be the unique and simultaneous sparsest ℓ^τ expansion for all τ, $0 \leq \tau \leq 1$. In particular, we prove that any decomposition has this nice property if the number of its nonzero coefficients does not exceed a quantity which we call the *spread* of the dictionary. After a brief discussion of the interplay between sparse decompositions and nonlinear approximation with various families of algorithms, we review several recent results that provide sufficient conditions for the Matching Pursuit, Orthonormal Matching Pursuit, and Basis Pursuit algorithms to have good recovery properties. The most general conditions are not straightforward to check, but weaker estimates based on the notions of coherence of the dictionary are recalled, and we discuss how these results can be applied to approximation and sparse decompositions with highly redundant incoherent dictionaries built by taking the union of several orthonormal bases. Eventually, based on Bernstein inequalities, we discuss how much approximation power can be gained by replacing a single basis with such redundant dictionaries.

Keywords: Nonlinear approximation, sparse decomposition, dictionary, Matching Pursuit, Basis Pursuit, linear programming, quadratic programming, coherence, spread, spark.

1. INTRODUCTION

In many signal and image processing applications, it is desirable to find a representation where the data is "sparse" [1] : replacing the original data with a sparse approximation can decrease substantially the processing cost. Early data representations were based on orthogonal linear transforms, but natural images and sounds contain superimposed features of very different nature (edges *vs* texture, transients *vs* stationary parts, ...) that do not necessary have a sparse representation in a single basis. Ten years ago, Mallat and Zhang [2] proposed to look for sparse representations using redundant dictionaries and introduced the Matching Pursuit algorithm : since then, the dictionary approach to sparse decompositions and nonlinear approximation has become more and more popular.

In a Hilbert space $\mathcal{H}$ of finite or infinite dimension, a *dictionary* is any family $\mathcal{D} = \{g_k\}$ of unit vectors (called *atoms*) that spans (a dense subspace of) the entire space. Nonlinear approximation consists in approximating a vector $f \in \mathcal{H}$ with m-term approximants of the form

$$f_m = \sum_{k \in I_m} c_k g_k, \text{ where } \mathrm{card}(I_m) = m, \tag{1}$$

Further author information: (Send correspondence to R. Gribonval)

R. Gribonval: E-mail: remi.gribonval@inria.fr, Telephone: (+33) 299 84 25 06, Address: IRISA, METISS Project, Campus de Beaulieu, F-35042 Rennes CEDEX, France.

M. Nielsen: E-mail: mnielsen@math.auc.dk, Telephone (+45) 9635 9984, Address: Aalborg University, Dept. of Mathematical Sciences, Fredrik Bajers Vej 7G, 9220 Aalborg East, Denmark

and the error of best m-term approximation is defined as

$$\sigma_m(f,\mathcal{D}) = \inf_{f_m} \|f - f_m\| \tag{2}$$

where the infimum is taken over all possible choices of coefficients $\{c_k\}$ and index sets I_m with $\mathrm{card}(I_m) = m$.

When $\mathcal{D}$ is an orthonormal basis for $\mathcal{H}$, it is well known how to get the best m-term approximant to any f : the solution is to keep the m atoms of the basis which have the largest inner products $|\langle f, g_k\rangle|$ with f. Moreover, in infinite dimension, the class $A^\alpha(\mathcal{D})$ of vectors f that can be approximated at a given rate $\sigma_m(f,\mathcal{D}) = \mathcal{O}(m^{-\alpha})$ is known to coincide (up to some technicalities [3, 4]) with the class $\mathcal{H}^\tau(\mathcal{D})$ of elements that have sparse expansions in the ℓ^τ sense

$$\sum_k |\langle f, g_k\rangle|^\tau < \infty, \qquad \alpha = 1/\tau + 1/2.$$

When $\mathcal{D}$ is a redundant dictionary, any f has infinitely many possible representations $f = \sum_k c_k g_k$. It is possible to choose *a* representation $\boldsymbol{c} = \{c_k\}$ according to various sparseness measures. Among the most popular ones are the ℓ^τ criteria : for $0 \le \tau \le 2$, when $\mathcal{D}$ is a frame*, one can define the "norms"

$$|f|_\tau^\tau := \min\{\|\mathbf{c}\|_\tau^\tau,\ f = \sum_k c_k g_k\} \tag{3}$$

where $\|\mathbf{c}\|_0^0$ denotes the number of nonzero entries in $\{c_k\}$ and $\|\boldsymbol{c}\|_\tau^\tau = \sum_k |c_k|^\tau$. For each sparseness measure, within the class $\mathcal{H}^\tau(\mathcal{D}) := \{f, |f|_\tau^\tau < \infty\}$, the elements have exact sparse representations [5] (which are not necessarily unique when $\tau \le 1$)

$$\mathbf{c}_\tau(f) := \arg \min_{\mathbf{c},\ f = \sum_k c_k g_k} \|\mathbf{c}\|_\tau^\tau. \tag{4}$$

When $\mathcal{D}$ is redundant and $f \in \mathcal{H}^{\tau_0}(\mathcal{D})$, the sparsest expansion $\mathbf{c}_\tau(f)$ exists for all for $\tau \ge \tau_0$. It changes drastically depending whether $\tau \le 1$ or not [6], but it is not clear in general whether $\mathbf{c}_\tau(f)$ depends on τ for $\tau_0 \le \tau \le 1$. In Section 2 we introduce the notion of *spread* $Z^1(\mathcal{D})$ of a dictionary and prove some conditions on f and $\mathcal{D}$ that guarantee that Eq. (4) defines a *unique* $\mathbf{c}_\tau(f)$ which does not depend on $0 \le \tau \le 1$. From the main results of Section 2 we derive the following Theorem, which generalizes to any $0 \le \tau \le 1$ what was obtained for $\tau \in \{0, 1\}$ in [7, 8]. It is based on the *coherence* of a dictionary

$$M(\mathcal{D}) := \sup_{k \ne k'} |\langle g_k, g_{k'}\rangle| \tag{5}$$

THEOREM 1.1. *Let $\mathcal{D}$ an arbitrary dictionary in a finite of infinite dimensional Hilbert space. If $f := \sum_k c_k g_k$ with*

$$\|\boldsymbol{c}\|_0^0 < \frac{1}{2}\left(1 + 1/M(\mathcal{D})\right) \tag{6}$$

then $\boldsymbol{c}$ is the unique and simultaneous *solution to all ℓ^τ minimization problems,* i.e.

$$\boldsymbol{c}_\tau(f) = \boldsymbol{c},\ \forall\, 0 \le \tau \le 1.$$

An immediate and very nice consequence is that for f with a "sparse enough" representation, linear programming provides simultaneously the solution $\boldsymbol{c}_1(f)$ to all the highly nonlinear ℓ^τ optimization problems $0 \le \tau \le 1$.

For arbitrary redundant dictionaries, the interplay between the rate of best m-term approximation and the sparseness of the expansions $f = \sum_k c_k g_k$ is not as clear as in the case of a single orthonormal basis, and getting the best m-term approximation from an arbitrary redundant dictionary is a NP-hard problem [9]. In Section 3 we discuss some relations between the sparseness of the expansions of f in $\mathcal{D}$ and the rate of approximation of f with various families of algorithms.

*For more general dictionaries in infinite dimension, a similar norm can be defined but requires a topological definition [4, 5].

Several sub-optimal algorithms have been introduced such as Basis Pursuit [10], Matching Pursuit [2], and variants thereof [11] to compute sparse decompositions or nonlinear approximations of signals with redundant dictionaries. In the last few years, many efforts have been put into understanding what structure should be imposed on f (for a given dictionary), or on the dictionary itself, so that "good" approximants can be obtained with such algorithms [7, 8, 12–21]. In Section 5 we give a survey of a series of recent results that guarantee that some of these algorithms recover the "good" decompositions.

The most general exact recovery conditions expressed in Sections 2 and 4 are not straightforward to check. In Section 5 we recall weaker conditions that are based on estimates of the *spread* and the *spark* of $\mathcal{D}$ (see Section 2) from its *coherence* (see Eq. (5)). We conclude by showing examples of highly redundant incoherent dictionaries where the results presented in this paper can be applied to perform nonlinear approximation and get sparse decompositions, and we discuss how much approximation power can be gained by replacing a single basis with such redundant dictionaries.

2. UNIQUENESS OF SPARSE ℓ^τ EXPANSIONS, $0 \leq \tau \leq 1$

In this section we review some conditions on f and $\mathcal{D}$ such that the ℓ^τ minimization problem

$$\text{minimize} \quad \|\mathbf{c}\|_\tau^\tau \quad \text{subject to } f = \sum_k c_k g_k \tag{7}$$

admits a unique solution, *i.e.* the sparse expansion $\boldsymbol{c}_\tau(f)$ given by Eq. (4) is unique. We also discuss some conditions so that this unique $\boldsymbol{c}_\tau(f)$ does not depend on $0 \leq \tau \leq 1$.

2.1. Definitions

We will us extensively the following definitions. The *support* of a sequence $\boldsymbol{c} = \{c_k\}$ is the set $I(\boldsymbol{c}) := \{k, c_k \neq 0\}$. We will call *kernel of the dictionary* the set $\text{Ker}(\mathcal{D}) := \{\boldsymbol{c}, \sum_k c_k g_k = 0\}$. For $0 \leq \tau \leq 1$ we define

$$P_\tau(I, \mathcal{D}) := \sup_{\boldsymbol{c} \in \text{Ker}(\mathcal{D}), \boldsymbol{c} \neq 0} \frac{\sum_{k \in I} |c_k|^\tau}{\sum_k |c_k|^\tau} \tag{8}$$

where we use the convention $0^0 = 0$ and $c^0 = 1, c > 0$. For notational convenience, we will simply denote $P_\tau(I)$. Finally, we introduce the quantities

$$Z^\tau(\mathcal{D}) := \inf_{\boldsymbol{c} \in \text{Ker}(\mathcal{D}), \|\mathbf{c}\|_\infty = 1} \|\mathbf{c}\|_\tau^\tau \,. \tag{9}$$

Because $Z^\tau(\mathcal{D})$ decreases with τ, the maximal value $Z^0(\mathcal{D})$ (which is called the *spark* of $\mathcal{D}$ in [7]) and the minimal one $Z^1(\mathcal{D})$ (which we call the *spread* of $\mathcal{D}$, because it measures how much the sequences from the kernel of $\mathcal{D}$ can spread –or how "flat" they can be– in the ℓ^1 sense) will play a special role.

2.2. General results

In [8] we proved the following Lemma by refining ideas from [12, 13]. It shows that, given a representation $f = \sum_{k \in I} c_k g_k$, the value of $P_\tau(I)$ can tell us if $\boldsymbol{c}$ is necessarily the unique solution to Eq. (7).

LEMMA 2.1 (GRIBONVAL, NIELSEN). *Let $\mathcal{D}$ a (possibly redundant) dictionary, I an index set and* $0 \leq \tau \leq 1$.

- *If $P_\tau(I) > 1/2$ there exists $\mathbf{c}$ and $\mathbf{d}$ with $\sum_k c_k g_k = \sum_k d_k g_k$ such that $I(\mathbf{c}) \subset I$ and $\|\mathbf{d}\|_\tau < \|\mathbf{c}\|_\tau$.*
- *If $P_\tau(I) < 1/2$ then, for all $\mathbf{c}$ such that $I(\mathbf{c}) \subset I$, $\mathbf{c}$ is* the unique *solution to* (7) *with $f := \sum_k c_k g_k$.*

COROLLARY 1. *If $f = \sum_k c_k g_k$ and*

$$P_\tau(I(\mathbf{c})) < 1/2 \tag{10}$$

then the unique *solution to Eq.* (7) *is $\mathbf{c}_\tau(f) = \mathbf{c}$.*

COROLLARY 2. *If $f = \sum_k c_k g_k$ and*

$$P_\tau(I(\mathbf{c})) < 1/2, \qquad \forall\, 0 \leq \tau \leq 1, \tag{11}$$

then the unique and simultaneous *solution to all ℓ^τ problems Eq.* (7) *is $\mathbf{c}_\tau(f) = \mathbf{c}$*, $\forall\, 0 \leq \tau \leq 1$.

2.3. Simplified results

Because the computation of $P_\tau(I)$ is not straightforward, it may not be easy to check conditions (10) or (11). In particular, it is not known if $P_\tau(I)$ varies monotonically with τ or not : if it were the case, the condition (11) would be simplified to $\max(P_0(I), P_1(I)) < 1/2$. In [8] we provided an upper estimate of $P_0(I)$ in terms of $Z^0(\mathcal{D})$, which is easily generalized here to any $0 \leq \tau \leq 1$:

LEMMA 2.2. *For any I,*

$$P_\tau(I) \leq \sup_{\mathbf{c} \in Ker(\mathcal{D}), \mathbf{c} \neq 0} \frac{card(I) \cdot \|\boldsymbol{c}\|_\infty^\tau}{\|\boldsymbol{c}\|_\tau^\tau} = \frac{card(I)}{Z^\tau(\mathcal{D})}. \tag{12}$$

It follows that we have :

COROLLARY 3. *If $f = \sum_k c_k g_k$ and*

$$card(I(\mathbf{c})) < \frac{Z^\tau(\mathcal{D})}{2}, \tag{13}$$

then the unique *solution to Eq.* (7) *is $\boldsymbol{c}_\tau(f) = \mathbf{c}$.*

COROLLARY 4. *If $f = \sum_k c_k g_k$ and*

$$card(I(\mathbf{c})) < \inf_\tau \frac{Z^\tau(\mathcal{D})}{2} = \frac{Z^1(\mathcal{D})}{2}, \tag{14}$$

then the unique and simultaneous *solution to all ℓ^τ problems Eq.* (7) *is $\boldsymbol{c}_\tau(f) = \mathbf{c}$, $\forall\, 0 \leq \tau \leq 1$.*

Corollary 4 shows that if f has a “sparse enough” representation, then it does not matter which ℓ^τ criterion is used to define sparseness. As a consequence, we can use the ℓ^1 criterion, which shows that this very sparse representation can be recovered using linear programming. In practice, given f, one does not know in advance whether f has a sparse enough representation. Nevertheless, linear programming provides *a* minimizer $\boldsymbol{c}_1(f)$, and it is possible to check whether $\boldsymbol{c}_1(f)$ is sparse enough. When the answer is positive, we know that it is *the* ℓ^1 minimizer as well as the ℓ^τ minimizer for $0 \leq \tau \leq 1$. In such a case, it is particularly interesting to notice that linear programming has retrieved in polynomial time the optimum of the ℓ^0 criterion, which usually requires a combinatorial search.

How sparse should be the “sparse enough” representation so that the nice consequences of Corollary 4 hold, this depends on the value of the spread $Z^1(\mathcal{D})$. The following result [7, 8], which is somehow a converse to Corollary 3 for $\tau = 0$, illustrates the role of the spark $Z^0(\mathcal{D})$:

LEMMA 2.3. *If the integer $Z^0(\mathcal{D})$ is even, there exists I with card$(I) = Z^0(\mathcal{D})/2$ and*

$$P_\tau(I) \geq 1/2,\ \forall\, 0 \leq \tau \leq 1.$$

If $Z^0(\mathcal{D})$ is odd, the same is true for some I with card$(I) = (Z^0(\mathcal{D}) + 1)/2$.

We refer the reader to [8] for a proof. Notice that when $Z^0(\mathcal{D})$ is odd, $(Z^0(\mathcal{D}) + 1)/2$ is the smallest integer value that exceeds $Z^0(\mathcal{D})/2$. Hence, if some algorithm provides a representation $\boldsymbol{c}$ of f with card$(I(\boldsymbol{c})) \geq Z^0(\mathcal{D})/2$, we cannot draw conclusions about $\boldsymbol{c}$ from the sole knowledge of card$(I(\boldsymbol{c}))$. In Section 5 we will see that, for dictionaries that are the union of orthonormal bases, there are sufficient conditions that ensure $P_0(I) < 1/2$ (resp. $P_1(I) < 1/2$) and are less restrictive than card$(I) < Z^0(\mathcal{D})$ (resp. card$(I) < Z^0(\mathcal{D})$).

Below is the coarsest but perhaps most general lower estimate on the spread $Z^1(\mathcal{D})$, which is based on the coherence of $\mathcal{D}$ (see Eq. (5)). The estimate coincides with the lower estimate on the spark $Z^0(\mathcal{D})$ obtained simultaneously in [7] and [8]. Its combination with Corollary 4 proves Theorem 1.1 which extends the results from [7, 8] to any $0 \leq \tau \leq 1$ instead of just $\tau \in \{0, 1\}$.

LEMMA 2.4. *Let $\mathcal{D}$ an arbitrary dictionary in finite or infinite dimension. Then*

$$Z^1(\mathcal{D}) \geq 1 + 1/M(\mathcal{D}). \tag{15}$$

Proof. Let $\boldsymbol{c} \in \text{Ker}(\mathcal{D})$, $\boldsymbol{c} \neq 0$. Because $\sum_k c_k g_k = 0$ we can write for any index l

$$|c_l| \quad = \quad \Big| - \sum_{k \neq l} c_k \langle g_k, g_l \rangle \Big| \leq M(\mathcal{D}) \cdot \sum_{k \neq l} |c_k|$$

hence we get $(1 + M) \cdot |c_l| \leq M \cdot \|\boldsymbol{c}\|_1^1$. Taking the supremum of all l we obtain $(1 + M) \cdot \|\boldsymbol{c}\|_\infty \leq M \cdot \|\boldsymbol{c}\|_1^1$. from which we reach the desired result. □

In this section we have focused on general conditions on $\boldsymbol{c}$ such that it solves simultaneously and uniquely Eq. (7) for all $0 \leq \tau \leq 1$. In Section 4 we will discuss other general conditions that ensure exact recovery of a sparse expansion through linear programming, and we will see that a similar recovery can be obtained with several variants of Matching Pursuit instead of linear programming. The general conditions obtained in this section and in Section 4 are not straightforward to check on a given expansion, and in Section 5 we will see how to compute effectively verifiable sufficient conditions, based on the notion of coherence (see Eq. (5)).

3. SPARSE DECOMPOSITIONS AND NONLINEAR APPROXIMATION

For each τ, the sparse representations $\boldsymbol{c}_\tau(f)$ exist (but are not always unique) for every f in the *sparseness class* $\mathcal{H}^\tau(\mathcal{D})$. When $\mathcal{D}$ is an orthonormal basis in infinite dimension, we have already mentioned that these sparseness classes actually coincide with *approximation classes*. Precisely, if we define the norm

$$\|f\|_{A^\alpha(\mathcal{D})} := \|f\| + \sum_{m \geq 1} [m^\alpha \sigma_m(f, \mathcal{D})]^\tau \cdot \frac{1}{m} \qquad \text{with } \alpha = 1/\tau + 1/2, \tag{16}$$

we have $A^\alpha(\mathcal{D}) := \{f, \|f\|_{A^\alpha(\mathcal{D})} < \infty\} = \mathcal{H}^\tau(\mathcal{D})$ with equivalent norms $\|f\|_{A^\alpha(\mathcal{D})}$ and $|f|_\tau$ [3, 4], and best m-term approximants can be obtained through the *thresholding strategy* (also called *greedy strategy*) which consists in two steps :

1. compute *the* representation $\boldsymbol{c}(f)$ such that $f = \sum_k c_k g_k$;
2. threshold $\boldsymbol{c}(f)$ to keep the index set I_m of the m largest coefficients, and build the approximant

$$f_m := \sum_{k \in I_m} c_k(f) g_k. \tag{17}$$

When $\mathcal{D}$ is a Riesz basis, but not necessarily orthonormal, we still have $A^\alpha(\mathcal{D}) = \mathcal{H}^\tau(\mathcal{D})$ with equivalent norms [22]. The thresholding strategy no longer provides the best m-term approximants, but is not far from it [23]. In fact, it reaches the optimum rate of approximation : we have $A^\alpha(\mathcal{D}) = T^\alpha(\mathcal{D})$ with equivalent norms, where $T^\alpha(\mathcal{D})$ is a new *thresholding approximation class* defined similarly to $A^\alpha(\mathcal{D})$ [22] by replacing, in Eq. (16), $\sigma_m(f, \mathcal{D})$ with $\|f_m - f\|$, f_m being obtained through the thresholding strategy.

When $\mathcal{D}$ is redundant, it is possible to generalize the notion of thresholding approximation class [5]. Without going into the details, in the first step of the thresholding strategy, several representations of f are possible: the thresholding approximation class is based on the one which provides the highest rate of approximation (not necessarily the sparsest expansion). It is not known in general whether $A^\alpha(\mathcal{D}) = T^\alpha(\mathcal{D}) = \mathcal{H}^\tau(\mathcal{D})$ with equivalent norms, but as soon as $\mathcal{D}$ is a frame† we have [4, 5] the continuous embeddings $\mathcal{H}^\tau(D) \subset T^\alpha(\mathcal{D}) \subset A^\alpha(\mathcal{D})$ with norm inequalities $c\|f\|_{A^\alpha(\mathcal{D})} \leq \|f\|_{T^\alpha(\mathcal{D})} \leq C|f|_\tau$. The converse embedding $A^\alpha(\mathcal{D}) \subset \mathcal{H}^\tau(\mathcal{D})$, which corresponds (up to some nontrivial technicalities that involve interpolation of function spaces [24]) to a so-called Bernstein inequality

$$|f_m|_\tau \leq C m^\alpha \|f_m\|, \text{ for all } m\text{-term expansions } f_m, \tag{18}$$

is known for some special redundant dictionaries [25, 26]. Bernstein inequalities are generally hard to prove, and it is still an open problem to characterize the class of dictionaries for which $A^\alpha(\mathcal{D}) \subset \mathcal{H}^\tau(\mathcal{D})$.

For a given f, assuming we know that $f \in \mathcal{H}^\tau(\mathcal{D})$, a practical version of the thresholding strategy consists in thresholding a sparsest ℓ^τ representation $\boldsymbol{c}_\tau(f)$ computed by some optimization algorithm. In general, as the value of τ is not known in advance, it is difficult to choose which ℓ^τ criterion should be optimized. However we have seen in Section 2 that for elements f that have "sparse enough" representations, $\boldsymbol{c}_\tau(f) = \boldsymbol{c}_1(f)$ does not depend on τ and can be computed using linear programming.

†weaker conditions than being a frame are sufficient [5], but their expression requires mathematical details that neither fit in this margin nor in this paper.

4. EXACT RECOVERY WITH SOME ALGORITHMS

Computing best m-term approximant from an arbitrary redundant dictionary is a NP-hard problem [9], and several algorithms have been proposed to get sub-optimal approximants. In this section we recall the definition of the main algorithms and discuss some of their essential properties. In particular we review a series of recent results on *exact recovery conditions* that indicate some cases where these algorithms are (close to being) optimal. In the next section we will discuss how these exact recovery conditions can be checked on specially structured dictionaries that may be highly redundant.

4.1. Basis Pursuit

Basis Pursuit (BP) [10] is a principle rather than an algorithm. It consists in computing the sparsest representation of f from $\mathcal{D}$ based on the ℓ^1 sparseness criterion. The implementations proposed in [10] combine recent methods of linear programming with *ad hoc* fast transforms based on the structure of the dictionary to speed up the computations as much as possible, but the computational complexity remains quite large and the implementation with poorly structured dictionaries is tricky. Recent approaches based on Minimum Fuel Neural Networks seem a promising direction [27]. A variant is Basis Pursuit Denoising (BPD) which uses quadratic programming to compute

$$\boldsymbol{c}(f,\lambda) := \arg\min_{\boldsymbol{c}} \left(\frac{1}{2}\|f - \sum_k c_k g_k\|^2 + \lambda \|\boldsymbol{c}\|_1 \right) \tag{19}$$

When the dictionary is a union of orthonormal bases (which may have common elements), the Block Coordinate Relaxation method [28] seems a good strategy to perform the optimization.

Many numerical experiments on synthetic data [10] have lead to the observation that if f happens to have a very sparse expansion in $\mathcal{D}$ and if $\mathcal{D}$ is well structured, then the sparse expansion is perfectly recovered through BP. The observation was turned into a series of proved theorems, first when $\mathcal{D}$ is the union of two mutually incoherent bases [12–14], then for the union of several mutually incoherent bases and for more general (quasi)incoherent dictionaries [7, 8, 15–17, 19].

One of the most general and most subtle results in this series is perhaps the following Theorem obtained by Fuchs [15–17].

THEOREM 4.1 (FUCHS). *Assume f has an expansion $f = \sum_k c_k g_k$ in an arbitrary dictionary. If there exists $h \in \mathcal{H}$ such that*

$$\langle h, g_k \rangle = \mathit{sign}(c_k), \quad k \in I(\boldsymbol{c}) \tag{20}$$

$$|\langle h, g_k \rangle| < 1, \quad k \notin I(\boldsymbol{c}) \tag{21}$$

then BP exactly recovers the expansion, i.e. $\boldsymbol{c} = \boldsymbol{c}_1(f)$.

Fuchs obtained a similar result for BPD under a slightly stronger condition.

THEOREM 4.2 (FUCHS). *Assume f has an expansion $f = \sum_k c_k g_k$ in an arbitrary dictionary. For $I = I(\boldsymbol{c})$, denote T_I the linear operator $T_I : \boldsymbol{d} \mapsto T_I \boldsymbol{d} := \sum_{k \in I} d_k g_k$. If λ is small enough and*

$$\left| \left\langle (T_I^\dagger)^\star \mathit{sign}(\boldsymbol{c}), g_k \right\rangle \right| < 1, \quad k \notin I, \tag{22}$$

where $(\cdot)^\star$ is the adjoint and $(\cdot)^\dagger$ the pseudo-inverse, then BPD "recovers" the expansion : each nonzero component of $\boldsymbol{c}(f,\lambda)$ corresponds to an index $k \in I$.

How small should be λ is a function of the index set I and, most of all, of the magnitude of the smallest components in $\boldsymbol{c}$. Note that when Eq.(22) is satisfied, Eqs. (20) and (21) are also satisfied with $h := (T_I^\dagger)^\star \mathrm{sign}(\boldsymbol{c})$, hence Eq. (22) implies that plain BP allows a perfect recovery.

The results of Fuchs are subtle in the sense that they take into account not only the support $I(\boldsymbol{c})$ of the representation, but also its sign sign$(\boldsymbol{c})$. If we are not willing to go into such details, noticing that

$$\sup_{\boldsymbol{c}, I(\boldsymbol{c})=I} \sup_{k \notin I} \left| \left\langle (T_I^\dagger)^\star \mathrm{sign}(\boldsymbol{c}), g_k \right\rangle \right| = \sup_{\boldsymbol{c}, I(\boldsymbol{c})=I} \sup_{k \notin I} \left| \left\langle \mathrm{sign}(\boldsymbol{c}), T_I^\dagger g_k \right\rangle \right| = \sup_{k \notin I} \|T_I^\dagger g_k\|_1$$

we recover the Exact Recovery Condition of Tropp [19] [Theorem 3.3] as a corollary of Fuchs' results.

THEOREM 4.3 (TROPP). *Assume $f = \sum_k c_k g_k$ in an arbitrary dictionary, and let $I = I(\boldsymbol{c})$. If*

$$\sup_{k \notin I} \|T_I^{\dagger} g_k\|_1 < 1 \tag{23}$$

then BP exactly recovers the expansion, i.e. *$\boldsymbol{c} = \boldsymbol{c}_1(f)$ is the unique ℓ^1-sparsest expansion of f.*

Because of Lemma 2.1, we have

$$\sup_{k \notin I} \|T_I^{\dagger} g_k\|_1 < 1 \qquad \Longrightarrow \qquad P_1(I) < 1/2, \tag{24}$$

but it is not clear whether the converse is true. Another question is whether (23) implies $P_\tau(I) < 1/2$ for some $0 \leq \tau < 1$.

4.2. Matching Pursuit and its variants

Matching Pursuit (MP) [2] is an iterative *greedy* algorithm that builds m-term approximants f_m and residuals $r_m = f - f_m$ by adding one term at a time in the approximant. At the beginning we set $f_0 = 0$ and $r_0 = f$; assuming f_{m-1} and r_{m-1} are defined, we set

$$k_m = \arg\max_k |\langle r_{m-1}, g_k \rangle| \tag{25}$$

$$f_m = f_{m-1} + \langle r_{m-1}, g_{k_m} \rangle \, g_{k_m} \tag{26}$$

$$r_m = f - f_m. \tag{27}$$

MP is also known as Projection Pursuit in the statistics community [29, 30] and as a Pure Greedy Algorithm [31] in the approximation community. In infinite dimension, it might not be possible to attain the maximum in Eq. (25) so "weak" variants are possible where we only require

$$|\langle r_{m-1}, g_{k_m} \rangle| \geq \alpha \sup_k |\langle r_{m-1}, g_k \rangle| \text{ with } 0 < \alpha < 1. \tag{28}$$

Both plain MP and its weak variants are known to converge [32], *i.e.* $\|f_m - f\| \to 0$, as soon as the dictionary spans (a dense subspace of) the entire space $\mathcal{H}$. In finite dimension, the convergence is even exponential [2] but there is in general no control on the convergence rate compared to best m-term approximation.

Orthonormal Matching Pursuit (OMP) [11] is a variant of MP where the update step given by Eq. (26) is replaced with

$$f_m = P_m f \tag{29}$$

where P_m is the orthonormal projector onto the linear span of $\{g_{k_1}, \ldots, g_{k_m}\}$. Just as plain MP, Orthonormal Matching Pursuit is convergent [31].

Several counter-example have been built that show that, in infinite dimension, it is possible to find "bad" dictionaries and "good" functions for which MP provides "bad" approximants. In [4] a dictionary is built where for some special $f = c_1 g_1 + c_2 g_2$, we have $\|f - f_m\| \geq C m^{-1/2}$. In [33], another construction provides $f \in \mathcal{H}^1(\mathcal{D})$, with a convergence as slow as $\|f - f_m\| \geq C m^{-0.27}$. Orthonormal Matching Pursuit has a better behaviour in terms of convergence rate, in particular $\|f - f_m\| \leq C |f|_1 m^{-1/2}$ [4]. Such counter-examples illustrate the need for conditions on $\mathcal{D}$ and/or f to ensure a good behaviour of MP. The first positive result in this direction was obtained when Villemoes [34] proved that for Walsh wavelet packets, MP on "good functions" ($f = c_i g_i + c_j g_j$ any sum of any two wavelet packets) picked up one of the two "correct" wavelet packets at each step and was exponentially convergent with $\|f_m - f\|^2 \leq (3/4)^m \|f\|^2$.

Nothing much happened in this direction until Gilbert *et al.* [18, 21] proved that (a variant of) Orthonormal Matching Pursuit recovers sufficiently sparse expansions over incoherent dictionaries. As shown by Tropp [19], the "sufficiently sparse expansion" condition turned out to be actually given by the Exact Recovery Condition, Eq. (23). Thus, the Exact Recovery Condition is not only sufficient for BP to exactly recover an expansion, it also works for OMP, and it is a sharp condition. Later on the result was extended [20] to plain MP as follows. Note that, when Theorem 4.4 is applied to $\mathcal{D}$ the Walsh wavelet packet dictionary and $f = c_1 g_{k_1} + c_2 g_{k_2}$, we recover the result of Villemoes [34].

THEOREM 4.4 (TROPP, GRIBONVAL-VANDERGHEYNST). *Assume $f = \sum_k c_k g_k$ in an arbitrary dictionary and let $I = I(\boldsymbol{c})$ (which need not be a finite set).*

1. *If*

$$\sup_{k \notin I} \|T_I^\dagger g_k\|_1 < 1 \tag{30}$$

then plain (resp. Orthogonal) Matching Pursuit recovers the expansion, in the sense that at each step, a "correct" atom is chosen : $\forall m \geq 1,\ k_m \in I$.

2. *If* (30) *is not satisfied, there exists* $f = \sum_{k \in I} c_k g_k$ *for which* $k_1 \notin I$.

3. *If card*$(I) < \infty$ *the convergence to* f *with OMP occurs in a finite number of steps. With MP, the convergence is exponential, and if card*$(I) < 1 + 1/M(\mathcal{D})$ *we have the estimate*

$$\|f_m - f\|^2 \leq \|f\|^2 \cdot \left(\left(1 - \frac{1}{\mathit{card}(I)}\right)(1 + M(\mathcal{D}))\right)^m. \tag{31}$$

Let us give the flavour of the proof. Denoting $\mathcal{H}_I$ the linear span of $\{g_k, k \in I\}$, Eq. (30) ensures for all $h \in \mathcal{H}_I$ that $\sup_{k \in I} |\langle h, g_k\rangle| > \sup_{k \notin I} |\langle h, g_k\rangle|$. As $r_0 = f \in \mathcal{H}_I$ and $f_0 = 0$, we get inductively that $k_m \in I$, $f_m \in \mathcal{H}_I$ (see Eqs. (26)- (29) and $r_m \in \mathcal{H}_I$, hence the first part of the result : $\forall m \geq 1,\ k_m \in I$. The second part is almost trivial: for OMP, each index $k \in I$ can be chosen only once because the residuals r_m are orthogonal to the atoms selected at previous steps; for MP, we simply notice that the pursuit is performed in the finite dimensional space $\mathcal{H}_I$, hence the exponential convergence [2].

As noted by Tropp, a similar exact recovery condition $\sup_{k \notin I} \|T_I^\dagger g_k\|_1 < \alpha$ ensures (sharply) the good behaviour of weak versions of MP and OMP, where the selection of the atom at each step if performed according to Eq. (28).

When f does not satisfy the assumptions of Theorem 4.4, it may happen that its best m-term approximation $f_m = \sum_{k \in I_m} c_k^m g_k$ is better conditioned in the sense that I_m satisfies Eq. (30). In this case, Tropp [19] proved that OMP will pick up "correct" atoms until it reaches a good enough approximant. MP [20] as well as other combinations of variants of MP [18, 21] share the same "provably good behaviour".

5. HIGHLY REDUNDANT INCOHERENT DICTIONARIES

So far we have mentioned a series of conditions that ensure the good behaviour of a variety of approximation or sparse decomposition algorithms. These conditions (Eqs. (10)-(14) and (23)) are generally not straightforward to check on a given representation $\boldsymbol{c}$. Using the coherence of $\mathcal{D}$ or its so-called *Babel function*, several estimates that can actually be checked have been proposed [7, 18, 19, 21], one of which is given by Lemma 2.4. In this section we focuss on sufficient conditions for exact recovery based on the coherence, and we show that there exists highly redundant incoherent dictionaries where the exact recovery conditions are easily met.

5.1. Exact recovery conditions for unions of bases

The first condition given in Lemma 2.4 was a lower bound on the spread $Z^1(\mathcal{D})$ based on the coherence $M(\mathcal{D})$. In [8] we proved other sufficient conditions for dictionaries that are the union of orthonormal bases, *i.e.* $\mathcal{D} = \cup_{l=1}^L \mathcal{B}_l$ where $\mathcal{B}_l$ is an orthonormal basis, $1 \leq l \leq L$. The results, which generalize those obtained previously for pairs of bases [12–14] are gathered in the following theorems.

THEOREM 5.1. *Let* $\mathcal{D}$ *a union of* L *orthonormal bases,* $f = \sum_k c_k g_k$ *and* $I = I(\boldsymbol{c})$. *Let* p_l *the number of indexes in* I *that correspond to atoms of the* l*-th basis. Without loss of generality assume the bases have been numbered so that* $0 \leq p_1 \leq \ldots \leq p_L$. *If*

$$\sum_{l \geq 2} \frac{Mp_l}{1 + Mp_l} < \frac{1}{2(1 + Mp_1)} \tag{32}$$

then $\mathbf{c}_1(f) = \boldsymbol{c}$ *is the unique solution to the* ℓ^1 *minimization problem.*

THEOREM 5.2. *Let $\mathcal{D}$ a union of L orthonormal bases, $f = \sum_k c_k g_k$ and $I = I(\mathbf{c})$. If*

$$card(I) < \left(\frac{1}{2} + \frac{1}{2(L-1)}\right)\frac{1}{M} \tag{33}$$

then $\mathbf{c}_0(f) = \mathbf{c}$ is the unique solution to the ℓ^0 minimization problem. If the stronger condition

$$card(I) < \left(\sqrt{2} - 1 + \frac{1}{2(L-1)}\right)\frac{1}{M} \tag{34}$$

is met, then $\mathbf{c}_0(f) = \mathbf{c}_1(f) = \mathbf{c}$ is the unique and simultaneous solution to both the ℓ^0 and the ℓ^1 minimization problems.

Condition (33) corresponds to a lower estimate on the spark $Z^0(\mathcal{D}) \geq (1 + 1/(L-1))/(2M)$, and we have by Lemma 2.2

$$\text{card}(I) < \left(\sqrt{2} - 1 + \frac{1}{2(L-1)}\right)\frac{1}{M} \implies P_0(I) < 1/2. \tag{35}$$

The second condition Eq. (34) is obtained from Theorem 5.1 by noticing that Eq. (34) implies Eq. (32). Moreover, Tropp proved that Eq. (32) implies the exact ℓ^1 recovery condition (Eq. (23)), so in summary we have

$$\text{card}(I) < \left(\sqrt{2} - 1 + \frac{1}{2(L-1)}\right)\frac{1}{M} \Longrightarrow \sum_{l \geq 2} \frac{Mp_l}{1 + Mp_l} < \frac{1}{2(1 + Mp_1)} \Longrightarrow \sup_{k \notin I} \|T_I^\dagger g_k\|_1 < 1 \Longrightarrow P_1(I) < 1/2. \tag{36}$$

The quantity $2\left(\sqrt{2} - 1 + \frac{1}{2(L-1)}\right)\frac{1}{M}$ is a natural candidate as a lower estimate on the spread $Z^1(\mathcal{D})$, but this is still a conjecture. For $0 \leq \tau < 1$ it is not clear whether a condition similar to Eq. (32) can be derived which ensures $P_\tau(I) < 1/2$.

5.2. Existence of highly redundant incoherent dictionaries

In finite dimension N, as soon as the dictionary contains an orthonormal basis, it is not difficult to check that the value of the coherence M is at least $1/\sqrt{N}$. Thus, a dictionary which contains an orthonormal basis is said to be (perfectly) incoherent if its coherence is minimum $M = 1/\sqrt{N}$. For $L = 2$, a well known pair of bases is incoherent :

EXAMPLE 1 (DIRAC-FOURIER). *Let $\mathcal{B}_1$ the Dirac basis in $\mathcal{H} = \mathbb{C}^N$ and $\mathcal{B}_2$ the complex Fourier basis. Then $\mathcal{D} = \mathcal{B}_1 \cup \mathcal{B}_2$ is incoherent.*

The Dirac-Fourier dictionary was indeed the first for which exact recovery conditions were proved [12]. Other examples of pairs of incoherent bases can be built, *e.g.*, based on the Haar and the Walsh systems on $L^2(0,1)$ (see [35, 36]). Let us see that it is actually possible to have unions of several orthonormal bases that are incoherent. The proof of Theorem 5.3 can be found in [37, 38].

THEOREM 5.3. *Let $N = 2^{j+1}$, $j \geq 0$ and consider $\mathcal{H} = \mathbb{R}^N$. There exists a dictionary $\mathcal{D}$ in $\mathcal{H}$ consisting of the union of $L = 2^j = N/2$ orthonormal bases for $\mathcal{H}$, such that for any pair $u, v \in \mathcal{D}, u \neq v$: $|\langle u, v\rangle| \in \{0, N^{-1/2}\}$.*

For $N = 2^j$, $j \geq 0$ and $\mathcal{H} = \mathbb{C}^N$, one can find a dictionary $\mathcal{D}$ in $\mathcal{H}$ consisting of the union of $L = N + 1$ orthonormal bases for $\mathcal{H}$, again with the perfect incoherence property: $u, v \in \mathcal{D}, u \neq v \Rightarrow |\langle u, v\rangle| \in \{0, N^{-1/2}\}$. The dictionaries given by Theorem 5.3 are called Grassmannian frames due to the connection between their construction and the Grassmannian packing problem [38]. They can be highly redundant: in the complex case, $\mathcal{D}$ can contain as many as $N(N+1)$ atoms.

Given an arbitrary orthonormal basis $\mathcal{B}_1$, it is not difficult to check from the above theorem that it is possible to complete $\mathcal{B}_1$ with other bases $\mathcal{B}_l$ which are incoherent. However, it does not seem clear whether such a completion is still possible when the first two mutually incoherent bases are fixed. In the case the Dirac and Fourier bases, the Chirp basis

$$g_k[n] = \frac{1}{\sqrt{N}} \exp\left(2i\pi \frac{n^2 + k \cdot n}{N}\right)$$

can be added to get three incoherent basis, but additional work is needed to know if the construction can go further. The reader interested in the construction of quasi-incoherent dictionaries (with $M \leq C/\sqrt{N}$) is referred to [38] and [18] where some other constructions are proposed.

6. CONCLUSION

We have discussed several issues related with nonlinear approximation and sparse decompositions with redundant dictionaries. We have proved that, under some general conditions, the "sparsest" representation of an element is unique and does not depend on the choice of the sparseness measure ℓ^τ, $0 \leq \tau \leq 1$. These results complete a series of recent advances which show that Basis Pursuit is a provably good strategy for sparse decomposition of signals over redundant dictionaries.

We have reviewed and made connections between different "exact recovery conditions" that turn out to ensure simultaneously the good behaviour of Basis Pursuit and Matching Pursuit (and variants thereof). Because these general conditions are not always straightforward to check, we have provided easily verifiable conditions based on the coherence of a dictionary. Besides general dictionaries, we have given specific results for unions of several orthonormal bases.

Besides the now classical pairs of mutually incoherent bases such as the Dirac-Fourier bases, we have shown that there exists highly redundant incoherent dictionaries built by taking the union of several mutually incoherent bases. A key question, back to applied signal/image processing, is whether and how one can build "useful" and "meaningful" families of incoherent bases, and in particular whether it is possible to do it incrementally. We are currently investigating such matters. Besides algebraic construction methods [37, 38], we believe a promising direction consists in "learning" the bases from training datasets [39–43].

Let us conclude this paper by mentioning that for incoherent dictionaries, a Bernstein inequality can be proved [35]. Instead of Eq. (18) with $\alpha = 1/\tau - 1/2$ we have for incoherent dictionaries

$$|f_m|_\tau \leq C m^\gamma \|f_m\|, \text{ for all } m\text{-term expansions } f_m,$$

with $\gamma = 2(1/\tau - 1/2)$, hence the approximation class $A^\gamma(\mathcal{D}) = A^{2\alpha}(\mathcal{D})$ introduced in Section 3 is, up to technicalities that involve interpolation spaces, continuously embedded into the sparseness class $\mathcal{H}^\tau(\mathcal{D})$. We conjecture that this result is sharp for unions of incoherent bases. This conjecture is optimistic : as $A^{2\alpha}(\mathcal{D}) \subsetneq A^\alpha(\mathcal{D})$, it implies

$$A^\alpha(\mathcal{D}) \not\subset \mathcal{H}^\tau(\mathcal{D})$$

which means that we gain in approximation power when combine several incoherent bases. To see it, notice [5] that

$$\mathcal{H}^\tau(\mathcal{D}) = \sum_l \mathcal{H}^\tau(\mathcal{B}_l) = \sum_l A^\alpha(\mathcal{B}_l),$$

so the sharpness of the Bernstein inequality would show that there exists elements $f \in \mathcal{H}$ that can be approximated at a good rate from $\mathcal{D}$ (they are in the approximation class $A^\alpha(\mathcal{D})$) even if they have no decomposition as a sum $f = \sum_l f_l$ of elements $f_l \in A^\alpha(\mathcal{B}_l)$, which can individually be approximated at the same good rate. In a sense, taking unions of incoherent bases would double the approximation power for some elements. The negative aspect of the conjecture is that thresholding approximation algorithms as well as Matching Pursuits will probably have a limited ability to approximate such elements, and other strategies will have to be investigated to get effectively computable good approximants. It may very well be that these elements for which the approximation gain is potentially substantial are those where the NP-hardness [9] of m-term approximation eventually shows up.

ACKNOWLEDGMENTS

The authors thanks R. DeVore for initiating their interest in this research and D. Donoho for suggesting –at a meeting of the Wavelet IDR Center in Florham Park in fall 1999– to look for optimality results for Basis Pursuit with more than two orthonormal bases. We also want to thank Joel Tropp and Jean-Jacques Fuchs for pointing out some chronological inaccuracies in the first version of this paper.

REFERENCES

1. B. D. Rao, "Signal processing with the sparseness constraint," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pp. 1861–1864, (Seattle), may 1998.
2. S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.* **41**, p-p. 3397–3415, Dec. 1993.

3. S. B. Stechkin, “On absolute convergence of orthogonal series,” *Dok. Akad. Nauk SSSR* **102**, pp. 37–40, 1955.
4. R. A. DeVore and V. N. Temlyakov, “Some remarks on greedy algorithms,” *Adv. Comput. Math.* **5**(2-3), pp. 173–187, 1996.
5. R. Gribonval and M. Nielsen, “Nonlinear approximation with dictionaries. I. Direct estimates.,” *J. Fourier Anal. and Appl.* , 2003. to appear.
6. K. Kreutz-Delgado, B. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, “Convex/schur-convex (csc) log-priors and sparse coding,” in *6th Joint Symposium on Neural Computation*, pp. 65–71, (Institute for Neural Computation), May 1999.
7. D. Donoho and M. Elad, “Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization,” *Proc. Nat. Aca. Sci.* **100**, pp. 2197–2202, Mar. 2003.
8. R. Gribonval and M. Nielsen, “Sparse decompositions in unions of bases,” Tech. Rep. 1499, IRISA, Nov. 2002. submitted to IEEE Trans. Inf. Th.
9. G. Davis, S. Mallat, and M. Avellaneda, “Adaptive greedy approximations,” *Constr. Approx.* **13**(1), pp. 57–98, 1997.
10. S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing* **20**, pp. 33–61, Jan. 1999.
11. Y. Pati, R. Rezaiifar, and P. Krishnaprasad, “Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition,” in *Proceedings of the* 27^{th} *Annual Asilomar Conf. on Signals, Systems and Computers*, Nov. 1993.
12. D. Donoho and X. Huo, “Uncertainty principles and ideal atomic decompositions,” *IEEE Trans. Inform. Theory* **47**, pp. 2845–2862, Nov. 2001.
13. M. Elad and A. Bruckstein, “A generalized uncertainty principle and sparse representations in pairs of bases,” *IEEE Trans. Inform. Theory* **48**, pp. 2558–2567, Sept. 2002.
14. A. Feuer and A. Nemirovsky, “On sparse representations in pairs of bases,” *IEEE Trans. Inform. Theory* , Nov. 2002. accepted.
15. J.J. Fuchs, “Une approche à l’estimation et l’identification simultanées,” Actes du seizième colloque GRETSI, Grenoble,vol. 2, pp. 1273-1276, sept. 1997.
16. J.J. Fuchs, “Detection and estimation of superimposed signals”. IEEE ICASSP, vol. III, pp. 1649-1652, Seattle, 1998.
17. J.-J. Fuchs, “On sparse representations in arbitrary redundant bases,” tech. rep., IRISA, Dec. 2002. submitted to IEEE Trans. Inf. Th.
18. A. Gilbert, S. Muthukrishnan, and M. Strauss, “Approximation of functions over redundant dictionaries using coherence,” in *The 14th ACM-SIAM Symposium on Discrete Algorithms (SODA’03)*, Jan. 2003.
19. J. Tropp, “Greed is good : Algorithmic results for sparse approximation,” tech. rep., Texas Institute for Computational Engineering and Sciences, 2003. In preparation.
20. R. Gribonval and P. Vandergheynst, “Exponential convergence of Matching Pursuit in quasi-incoherent dictionaries,” tech. rep., IRISA, 2003. in preparation.
21. A. Gilbert, S. Muthukrishnan, M. Strauss, and J. Tropp, “Improved sparse approximation over quasi-incoherent dictionaries,” in *Int. Conf. on Image Proc. (ICIP’03)*, (Barcelona, Spain), sep 2003.
22. R. Gribonval and M. Nielsen, “Some remarks on nonlinear approximation with Schauder bases,” *East J. Approx.* **7**(3), pp. 267–285, 2001.
23. S. Konyagin and V. Temlyakov, “A remark on greedy approximation in Banach spaces,” *East J. Approx.* **5**(3), pp. 365–379, 1999.
24. R. A. DeVore, “Nonlinear approximation,” in *Acta numerica, 1998*, pp. 51–150, Cambridge Univ. Press, Cambridge, 1998.
25. R. Gribonval and M. Nielsen, “On approximation with spline generated framelets,” *Constr. Approx.* , 2003. to appear.
26. R. Gribonval and M. Nielsen, “On the quasi-greedy property and uniformly bounded orthonormal systems,” Tech. Rep. R-2003-09, Aalborg Univ., Dept of Math., Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg East, Denmark, 2003. submitted to J. Approx. Th.
27. Z. S. Wang, J. Y. Cheung, Y. S. Xia, and J. D. Z. Chen, “Minimum fuel neural networks and their application to overcomplete signal representations,” *IEEE Trans. Circuits and Systems–I: Fundamental Theory and Applications* **47**, pp. 1146–1159, aug 2000.

28. S. Sardy, A. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Journal of Computational and Graphical Statistics* **9**(2), 2000.
29. J. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Stat. Assoc.* **76**, pp. 817–823, 1981.
30. P. Huber, "Projection pursuit," *The Annals of Statistics* **13**(2), pp. 435–475, 1985.
31. V. Temlyakov, "Weak greedy algorithms," *Advances in Computational Mathematics* **12**(2,3), pp. 213–227, 2000.
32. L. Jones, "On a conjecture of Huber concerning the convergence of PP-regression," *The Annals of Statistics* **15**, pp. 880–882, 1987.
33. E. Livschitz and V. Temlyakov, "Two lower estimates in greedy approximation," *Constr. Approx.* **19**, 2003.
34. L. Villemoes, "Nonlinear approximation with walsh atoms," in *Proceedings of "Surface Fitting and Multiresolution Methods", Chamonix 1996*, A. Le M'ehaut'e, C. Rabut, and L. Schumaker, eds., pp. 329–336, Vanderbilt University Press, 1997.
35. R. Gribonval and M. Nielsen, "Nonlinear approximation with dictionaries. II. Inverse estimates," Tech. Rep. R-02-2019, Aalborg Univ., Dept of Math., Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg East, Denmark, Nov. 2002. submitted to Constructive Apprimation.
36. B. Golubov, A. Efimov, and V. Skvortsov, *Walsh series and transforms*, Theory and applications, Translated from the 1987 Russian original by W. R. Wade, Kluwer Academic Publishers Group, Dordrecht, 1991.
37. A. R. Calderbank, P. J. Cameron, W. M. Kantor, and J. J. Seidel, "Z_4-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets," *Proc. London Math. Soc. (3)* **75**(2), pp. 436–480, 1997.
38. T. Strohmer and R. Heath, "Grassmannian frames with applications to coding and communications," tech. rep., 2002. Preprint, submitted to Appl.Comp.Harm.Anal.
39. D. Field and B. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, pp. 607–609, 1996.
40. A. Bell and T. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research* **37**(23), pp. 3327–3338, 1997.
41. A. Bell and T. Sejnowski, "Learning the higher-order structure of a natural sound," *Network* **7**(2), 1996.
42. M. Lewicki, "Efficient coding of natural sounds," *Nature Neuosci.* **5**(4), pp. 356–363, 2002.
43. S. Abdallah and M. Plumbley, "If edges are the independent components of natural images, what are the independent components of natural sounds?," in *Proc. of the Int. Conf. on Indep. Component Anal. and Blind Signal Separation (ICA2001)*, pp. 534–539, (San Diego, California), Dec. 2001.