Signal Processing 91 (2011) 2754–2766

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Matching pursuit shrinkage in Hilbert spaces

Tieyong Zeng [a,*], François Malgouyres [b]

[a] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[b] Institut de Mathématiques (UMR CNRS 5219), Université Paul Sabati, 31062 Toulouse Cedex 4, France

ARTICLE INFO

Article history:
Received 3 February 2010
Received in revised form
9 March 2011
Accepted 4 April 2011
Available online 14 April 2011

Keywords:
Dictionary
Matching pursuit
Shrinkage
Sparse representation

ABSTRACT

In this paper, we study a variant of the matching pursuit named matching pursuit shrinkage. Similar to the matching pursuit it seeks for an approximation of a datum living in a Hilbert space by a sparse linear expansion in a countable set of atoms. The difference with the usual matching pursuit is that, once an atom has been selected, we do not erase all the information along the direction of this atom. Doing so, we can evolve slowly along that direction. The goal is to attenuate the negative impact of bad atom selections.

We analyze the link between the shrinkage function used by the algorithm and the fact that the result belongs to l^2, l^1 and l^0 space. Experimental results are also reported to show the potential application of the proposed algorithm.

1. Introduction

1.1. Recollection on sparse approximation

Finding a sparse approximation of a datum in a Hilbert space is a recurrent problem in applied science. The problem is to approximate a datum $v \in \mathcal{H}$ ($\mathcal{H}$ is a Hilbert space of finite or infinite dimension) by a linear expansion in a dictionary of known atoms $(\psi_i)_{i\in I}$:

$$v \sim \sum_{i\in I} \lambda_i \psi_i,$$

where $(\lambda_i)_{i\in I} \in \mathbb{R}^I$. The approximation is needed because v is usually corrupted by noise. It is sometimes preferable to search for an approximation which is coarser than the noise requires. By this we favor desired/expected properties of the coordinates $(\lambda_i)_{i\in I}$.

Moreover, the dictionary is usually overcomplete. This offers the freedom to select among all the possible sets of coordinates one of those agreeing with some prior knowledge or desired property of the coordinates. The property receiving most of the attention is sparsity. Heuristically, we select the set of coordinates offering the "simplest" explanation of the datum. Rigorously, for a given accuracy after reconstruction, we want

$$\|(\lambda_i)_{i\in I}\|_0 \stackrel{\text{def}}{=} \#\{i \in I, \lambda_i \neq 0\},$$

to be as small as possible, where # denotes the cardinality of a set.

Unfortunately, problems similar to

$$\begin{cases} \text{minimize } \|(\lambda_i)_{i\in I}\|_0 \\ \text{under the constraint } \left\| \sum_{i\in I} \lambda_i \psi_i - v \right\| \leq \tau, \end{cases} \tag{1}$$

where $\tau > 0$ and $\|\cdot\|$ is the norm associated with the scalar product of the considered Hilbert space, are known to be NP-Hard in general (see [9]).

As a conclusion, solving (1) is both an open and interesting problem. It receives a lot of attention and it would be a challenge to list all the contributions to its resolution. Before describing the most popular techniques, we give in the next section the algorithm studied in this paper. It will then be simpler to motivate our proposal.

* Corresponding author.
E-mail addresses: zeng@hkbu.edu.hk (T. Zeng), malgouy@math.univ-paris13.fr (F. Malgouyres).

0165-1684/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2011.04.010

1.2. The matching pursuit shrinkage

For dictionary composed of atoms with unit-length, the matching pursuit shrinkage (MPS) is very similar to the usual matching pursuit (MP) algorithm (see [27]). The main difference is that it uses a shrinkage[1] function $\theta: \mathbb{R}\to\mathbb{R}$. We describe the algorithm in Table 1.

Several convergence criterions might be considered and for simplicity, we always assume that the algorithm stops whenever $s_n=0$.

Whenever the series are convergent, we can construct coordinates

$$\lambda_i = \sum_{n\in\mathbb{N}:\gamma_n=i} s_n, \quad \forall i\in I \tag{5}$$

from the result of the MPS. We also consider (if the series is convergent)

$$u=\sum_{i\in I}\lambda_i\psi_i=\sum_{n=0}^{+\infty} s_n\psi_{\gamma_n}.$$

Notice that if we sum (3) for $n=0\dots N-1$, we obtain

$$v=\sum_{n=0}^{N-1} s_n\psi_{\gamma_n}+R^N v. \tag{6}$$

This explains the name "residual error" for $R^N v$.

1.3. Other algorithms promoting sparsity

One of the oldest and simplest algorithm for building a sparse approximation is the matching pursuit (MP) [27] or projection pursuit [13]. It corresponds to the algorithm of Table 1 when θ is the identity (i.e. $s_n=M_n$).

In finite dimension (see [27]) and in infinite dimension but under restrictive conditions on the dictionary and the signal (see [19]), the MP is known to converge exponentially. When no hypotheses are made on the dictionary, we only know that the MP converges (see [27]). Some examples show that we cannot expect a "good" convergence rate in the most general settings (see [11]). Yet the MP and the best k-term approximation have a similar convergence, when the dictionary is "quasi-orthogonal" (see [33]).

There exists "fast" variants of the MP (see [17]). Indeed, a real-time implementation of the MP is available for audio signal processing (see [21]). The improved performances are obtained by carefully optimizing the structures, algorithms and their implementation. In particular, the update of $(\langle R^n v,\psi_i\rangle)_{i\in I}$ and the computation of γ_n satisfying (2) (in Table 1) are implemented in a very efficient way. Each iteration of the MP is typically of complexity $O(\log(\#I))$. These optimizations are possible because only one coordinate is updated. If K coordinates (as defined in (5)) are modified at each iteration, we obtain a complexity $O(K+\log(\#I))$. This might be less favorable when K is large. Although its approximation performances are not as good as most modern algorithms such as CoSaMP [29] and similar algorithms or l^1 regularization, these accelerations make the MP a useful algorithm.

[1] The rigorous definition of shrinkage functions is given in Section 2.

Table 1
The matching pursuit shrinkage (MPS).

- Input : A datum v, a dictionary $(\psi_i)_{i\in I}$, a shrinkage function θ and $\alpha\in[0,1]$
- Output : Coordinates $(s_n,\gamma_n)_{n\in\mathbb{N}}$
- The algorithm
  - Initialize $R^0 v=v$
  - Repeat until convergence (loop in n)
    1. Select a well correlated atom ψ_{γ_n} such that
    $$|\langle\psi_{\gamma_n},R^n v\rangle|\geq\alpha\sup_{i\in I}|\langle R^n v,\psi_i\rangle|; \tag{2}$$
    2. Evolve along ψ_{γ_n}
    $$R^n v=s_n\psi_{\gamma_n}+R^{n+1}v, \tag{3}$$
    where
    $$s_n=\theta(M_n) \quad \text{with } M_n=\langle R^n v,\psi_{\gamma_n}\rangle. \tag{4}$$

The accelerations described in [21] can be applied to the MPS, as described in Table 1. The potential advantage of introducing a shrinkage function θ is to attenuate the mistakes in the selection of a coordinate γ_n (or more precisely, coefficient of atoms). In the inverse problem context such as compressed sensing, there is a "ground truth" representation associated to "true atoms", the term "bad selection of atoms" is then well defined. Let us underline that avoiding wrong selection of coordinates is one of the key ingredients of modern variants of the MP such as CoSaMP [29], subspace pursuit [8] and iterative hard thresholding [1]. However, especially when the solution we are looking for is moderately sparse, those algorithms are more computationally intensive.

Let us go back in time. The most famous variant of the MP is the orthogonal matching pursuit (OMP) (see [30]). In Table 1, it replaces the update rule (3) by an orthogonal projection onto the subspace generated by the selected atoms. It is known to provide sparser solutions than the regular MP. From the computational point of view, it has two drawbacks. Firstly, although several attempts have been made to optimize it (see [26,7]), the orthogonal projection is computationally expensive and often requires too much memory. Secondly, every selected coordinate is modified. As a consequence, the adaptation of the optimization performed in [21] would only be efficient when the result is very sparse. Algorithms such as the gradient pursuit (see [2]) approximately solve the OMP at a cost more similar to the cost of the MP. However, at each iteration, they typically update all the selected coordinates. The computational cost of the gradient pursuit is therefore more important than the cost of a fast implementation of the MP, when the solution is moderately sparse.

Finally, the l^1 regularization (also named basis pursuit and basis pursuit denoising, see [5] and the papers citing it) is a very important sparsity promoting model. It consists in minimizing

$$\frac{1}{2}\left\|v-\sum_{i\in I}\lambda_i\psi_i\right\|^2+\beta\sum_{i\in I}|\lambda_i| \tag{7}$$

and it is very efficient for providing sparse approximations of $v \in \mathcal{H}$. However, its resolution remains (and will probably remain in a near future) a challenge for large scale problems. A famous (and representative) solver of the l^1 regularization problem is the iterative soft thresholding (see [10]). It updates all the coordinates at each iteration and often requires many iterations before it reaches a suitable convergence level. It is interesting to notice that, in this context, the impact of the choice of the shrinkage function is well understood (see [6]): Every proximal thresholding function corresponds to a different objective function.

Inspired by the l^1 regularization problem, a "coordinate-wise optimization algorithm" has been proposed in [12]. It performs a soft thresholding, sequentially on each coordinate. The "greedy coordinate descent" proposed in [34] is similar but selects the coordinates according to a criteria similar to the MP. Because they update only one coordinate at each iteration, these algorithms can benefit from the optimization proposed in [21]. Most recently, we also have some other related works [23,24,32].

1.4. Notations

The following notations and hypotheses hold all along the paper.

The datum v belongs to a Hilbert space $\mathcal{H}$. The space $\mathcal{H}$ might be of finite or infinite dimension. For any two elements u and w in $\mathcal{H}$, their scalar product is denoted by $\langle u,w\rangle$. As usual, the norm of $u \in \mathcal{H}$ is defined by $\|u\| \stackrel{\text{def}}{=} \sqrt{\langle u,u\rangle}$. The dictionary $(\psi_i)_{i\in I}$ is made of atoms $\psi_i \in \mathcal{H}$, such that $\|\psi_i\| = 1$, for all $i \in I$. We sometimes denote the dictionary by $\mathcal{D}$. For simplicity, we assume that I is countable. In particular, the supremum in (2) may not be reached. In such a case, the MPS is only defined for $\alpha < 1$. For any $u \in \mathcal{H}$, we denote $\|u\|_{\mathcal{D}} \stackrel{\text{def}}{=} \sup_{i\in I} |\langle u,\psi_i\rangle|$. We denote

$$V \stackrel{\text{def}}{=} \overline{Span\{\mathcal{D}\}} \tag{8}$$

the closed linear span of the elements of $\mathcal{D}$. We denote $V^\perp$ the orthogonal complement of V in $\mathcal{H}$. We denote the orthogonal projection onto V and $V^\perp$ by P_V and $P_{V^\perp}$ respectively.

The sequences $(s_n)_{n\in\mathbb{N}}$, $(\gamma_n)_{n\in\mathbb{N}}$, $(R^n v)_{n\in\mathbb{N}}$ are always defined according to Table 1. The coordinates $(\lambda_i)_{i\in I}$ are according to (5).

We also use the standard notations: $\mathrm{sgn}(t) = 1$, if $t \geq 0$ and -1, if $t < 0$; the symbol # denotes the cardinal of a set; the floor function $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t.

1.5. Overview

The sketch of the paper is as follows. In Section 2, we define *shrinkage*, *thresholding* and *gap* functions. We also illustrate these definitions by several examples. Sections 3–6 are then devoted to some important theoretical analysis of the MP shrinkage algorithm which integrates the general shrinkage function with MP. Precisely, in Section 3, we prove that under mild condition, the MP shrinkage algorithm converges. Indeed, as soon as θ is a *shrinkage function*, the residual $(R^n v)_{n\in\mathbb{N}}$ converges and the series $\sum_{n\in\mathbb{N}} s_n \psi_{\gamma_n}$ is convergent. We also prove that $(s_n)_{n\in\mathbb{N}}$ is square summable. In Section 4, we prove that when θ is a *thresholding function*, $(s_n)_{n\in\mathbb{N}}$ is absolutely summable. This implies in particular that $(\lambda_i)_{i\in I}$ exists and is absolutely summable. In Section 5, we prove that when θ is a *gap function*, the sequence $(s_n)_{n\in\mathbb{N}}$ is finite. Again, this implies that $(\lambda_i)_{i\in I}$ exists and is finite. In Section 6, we evaluate $\|\sum_{n\in\mathbb{N}} s_n \psi_{\gamma_n} - P_V v\|_{\mathcal{D}}$, when θ is a general shrinkage function. In Section 7, some experimental results show that in the presence of noise, the new algorithm does not only outperform the regular MP, but also behaves better than some other classical Greedy methods and basis pursuit denoising model when used for detection. Finally, we give some concluding remarks and discussions in Section 8.

2. General shrinkage functions

2.1. Definitions

Definition 1. A function $\theta(\cdot) : \mathbb{R} \to \mathbb{R}$ is called a *general shrinkage function* if and only if it satisfies:

1. $\theta(\cdot)$ is non-decreasing, i.e.

   $$\forall t,t' \in \mathbb{R}, \quad t \leq t' \Longrightarrow \theta(t) \leq \theta(t');$$

2. $\theta(\cdot)$ shrinks the amplitude, i.e.

   $$\forall t \in \mathbb{R}, \quad |\theta(t)| \leq |t|.$$

Notice that this implies

$$\theta(0) = 0,$$

and

$$\theta(-t) \leq 0 \leq \theta(t), \quad \forall t \geq 0. \tag{9}$$

Therefore, for any general shrinkage function $\theta(\cdot)$ and any $t \in \mathbb{R}$, we know that

$$\text{if } t \geq 0,\ 0 \leq \theta(t) \leq t \text{ and } 0 \leq \theta(t)(t-\theta(t)),$$

$$\text{if } t \leq 0,\ 0 \geq \theta(t) \geq t \text{ and } 0 \leq \theta(t)(t-\theta(t)).$$

As a conclusion,

$$\forall t \in \mathbb{R}, \quad \theta(t)(t-\theta(t)) \geq 0. \tag{10}$$

The inequality (9) also guarantees that

$$\forall t \in \mathbb{R}, \quad |t||\theta(t)| = t\theta(t). \tag{11}$$

Definition 2. Let $\theta(\cdot)$ be a general shrinkage function, we call

- the *internal threshold*: $\tau^- \stackrel{\text{def}}{=} \inf_{t:\theta(t)\neq 0} |t|$
- the *external threshold*: $\tau^+ \stackrel{\text{def}}{=} \sup_{t:\theta(t)=0} |t|$.

Moreover, we say that $\theta(\cdot)$ is a *thresholding function* if and only if: $\tau^- > 0$, i.e.

$$\exists \tau > 0, \forall x \in \mathbb{R}, \quad |x| \leq \tau \Rightarrow \theta(x) = 0; \tag{12}$$

otherwise, it will be called *non-thresholding function*.

Note that in the literature, the standard usage is that shrinkage and thresholding are almost interchangeable term. However, throughout our paper, there are important distinctions between general shrinkage functions and more specific thresholding/gap functions.

If $\theta(\cdot)$ is a thresholding function, we trivially have

$$0<\tau^-\le\tau^+.$$

The internal and external thresholds are illustrated in Fig. 1.

Since (10) holds for any general shrinkage function, the following definition is valid.

Definition 3. The *gap* of a general shrinkage function $\theta(\cdot)$ is defined by

$$\mathrm{gap}(\theta)\triangleq\inf_{t:\theta(t)\neq 0}\sqrt{\theta^2(t)+2\theta(t)(t-\theta(t))}. \tag{13}$$

If $\mathrm{gap}(\theta)>0$, we call θ a gap function and if $\mathrm{gap}(\theta)=0$, the function is called a non-gap function.

The following relation exists between the gap and the internal threshold of a general shrinkage function. It proves in particular that any gap function is a thresholding function.

Proposition 1. *For any gap function $\theta(\cdot)$, the following statements hold.*

1. *We have*
$$\mathrm{gap}(\theta)\le\tau^-,$$
where τ^- is the internal threshold of $\theta(\cdot)$.
2. *We have*
$$\inf_{t:\theta(t)\neq 0}t\theta(t)\le\mathrm{gap}(\theta)^2\le 2\inf_{t:\theta(t)\neq 0}t\theta(t)\le 2\tau^+\inf_{t:\theta(t)\neq 0}|\theta(t)|.$$

Proof. The proof is given in Appendix.

For instance, when θ is odd we have

$$\inf_{t:\theta(t)\neq 0}t\theta(t)=\tau^+\inf_{t:\theta(t)>0}\theta(t),$$

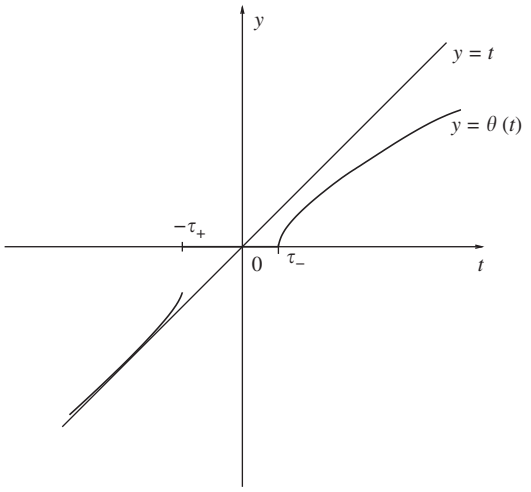

Fig. 1. Example of a thresholding function θ. It is non-gap. Its internal and external thresholds are not equal.

since θ is non-decreasing. Geometrically, $\inf_{t:\theta(t)>0}\theta(t)$ is simply the amplitude of the vertical discontinuity at the boundary of the segment $\{t:\theta(t)\neq 0\}$. It is also clear from the item 2 of Proposition 1 that when the gap function is continuous at the boundary of the segment $\{t:\theta(t)\neq 0\}$, the function is non-gap. These are the reasons leading to the name gap.

The motivation for considering the gap comes from its implication on the convergence of the MPS.

2.2. Examples

Let us illustrate the above definitions through some examples.

1. For $\tau>0$, the soft thresholding function $\rho_\tau(\cdot)$ defined by
$$\rho_\tau(t)=\mathrm{sgn}(t)\cdot\max(|t|-\tau,0) \tag{14}$$
is a thresholding function and it is a non-gap function, i.e. $\mathrm{gap}(\rho_\tau)=0$.
2. For $\tau>0$, the hard thresholding function defined by
$$h_\tau(t)=\begin{cases}t & \text{if } |t|>\tau,\\ 0 & \text{otherwise}\end{cases}$$
is a thresholding function and it is a gap shrinkage function with gap τ.
3. The identity function defined as
$$i(t)=t,\quad \forall t\in\mathbb{R} \tag{15}$$
is not a thresholding function and it is a non-gap function.
4. For $\tau>0$, $p>0$, the general non-negative Garrote threshold function (for the case $p=2$, see [15,4]) defined as
$$\delta^G_\tau(t)=t\max\left(0,\left(1-\frac{|\tau|^p}{|t|^p}\right)\right),\quad \forall t\in\mathbb{R} \tag{16}$$
is a thresholding function and it is non-gap.
5. For $0<\tau_1<\tau_2$, the firm shrinkage function (see [16]) defined as
$$\delta_{\tau_1,\tau_2}(t)=\begin{cases}0 & \text{if } |t|\le\tau_1,\\ \mathrm{sgn}(t)\dfrac{\tau_2(|t|-\tau_1)}{\tau_2-\tau_1} & \text{if } \tau_1<|t|<\tau_2,\\ t & \text{if } |t|\ge\tau_2\end{cases} \tag{17}$$
is a thresholding function and it is non-gap.
6. For $p\in\mathbb{N}$, $\tau>0$, the generalized threshold function (see [36]) defined as
$$\delta^p_\tau(t)=\begin{cases}0 & \text{if } |t|\le\tau,\\ t-\dfrac{\tau^p}{t^{p-1}}(\mathrm{sgn}(t)^p) & \text{if } |t|>\tau\end{cases} \tag{18}$$
is a thresholding function and it is non-gap.

3. Convergence of the MP shrinkage for a general shrinkage function

This section is devoted to prove that under mild condition, the MP shrinkage algorithm converges.

Proposition 2. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v\in\mathcal{H}$ and $\theta(\cdot)$ be a general shrinkage function. For any $M>0$ and any $v\in\mathcal{H}$, the quantities defined in Table 1 satisfy*

$$\|v\|^2=\sum_{n=0}^{M-1}\left(s_n^2+2s_n(M_n-s_n)\right)+\|R^Mv\|^2. \tag{19}$$

As a consequence, we have

$$\|v\|^2\geq\sum_{n=0}^{M-1}s_n^2+\|R^Mv\|^2, \tag{20}$$

$$\sum_{n=0}^{+\infty}s_n^2<+\infty, \tag{21}$$

$$\sum_{n=0}^{+\infty}|s_n||M_n|<+\infty, \tag{22}$$

$$(\|R^nv\|)_{n\in\mathbb{N}}\quad \textit{is non-increasing.} \tag{23}$$

Proof. We can deduce from

$$R^{n+1}v=R^nv-s_n\psi_{\gamma_n},$$

and $\langle\psi_{\gamma_n},\psi_{\gamma_n}\rangle=1$ that

$$\begin{aligned}\|R^{n+1}v\|^2&=\|R^nv\|^2-2s_n\langle R^nv,\psi_{\gamma_n}\rangle+s_n^2\\&=\|R^nv\|^2-2s_n(M_n-s_n)-s_n^2.\end{aligned}$$

Summing these equalities for all $n=0,\dots,M-1$, we obtain after simplification

$$\|R^Mv\|^2=\|R^0v\|^2-\sum_{n=0}^{M-1}(s_n^2+2s_n(M_n-s_n)).$$

We then obtain (19) from $R^0v=v$.

Using (10), we know that

$$s_n(M_n-s_n)=\theta(M_n)(M_n-\theta(M_n))\geq 0.$$

Together with (19) this leads to (20).

Notice that this also provides (23). Moreover, (20) guarantees that $(\sum_{n=0}^{M}s_n^2)_{M\in\mathbb{N}}$ is a bounded increasing sequence. It converges and (21) holds. We also have

$$\begin{aligned}2\sum_{n=0}^{M-1}|s_n||M_n|&=2\sum_{n=0}^{M-1}s_nM_n\quad\text{from (11)}\\&=\|v\|^2-\|R^Mv\|^2+\sum_{n=0}^{M-1}s_n^2\quad\text{from (19)}\\&\leq\|v\|^2+\sum_{n=0}^{+\infty}s_n^2.\end{aligned}$$

This ensures that (22) holds. □

Note that Proposition 2 ensures that

$$\sum_{n=0}^{+\infty}|s_n|^2<+\infty. \tag{24}$$

Now we can prove the convergence of the MPS algorithm. As pointed out by one of the reviewers, it is interesting to notice that though MPS algorithm should be regarded a case of the so-called "Approximate Weak Greedy Algorithms " in [18], it seems that the convergence proofs provided there could not be directly applied to MPS.

Theorem 1. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v\in\mathcal{H}$ and $\theta(\cdot)$ be a* general shrinkage function. *The sequences defined in* (4) *satisfy*

$(R^nv)_{n\in\mathbb{N}}$ *converges.*

As a consequence, the limit

$$\sum_{n=0}^{+\infty}s_n\psi_{\gamma_n}\quad\textit{is finite.}$$

We denote the limit of $(R^nv)_{n\in\mathbb{N}}$ by $R^{+\infty}v$ and we trivially have

$$v=\sum_{n=0}^{+\infty}s_n\psi_{\gamma_n}+R^{+\infty}v.$$

Proof. The proof is based on Jones' proof for the convergence of projection pursuit regressions (see [14]) and the proof of Theorem 1 in [27].

First notice that the statement of the proposition is trivial for $v=0$. We further assume that $v\neq 0$.

In order to prove the theorem, we prove that the sequence $(R^nv)_{n\in\mathbb{N}}$ is a Cauchy sequence. Before doing so, let us start with some preliminaries.

Notice first that for all $w_1,w_2\in\mathcal{H}$, we have

$$\begin{aligned}\|w_1-w_2\|^2&=\|w_1\|^2-\|w_2\|^2-2\langle w_2,w_1-w_2\rangle\\&\leq\|w_1\|^2-\|w_2\|^2+2|\langle w_2,w_1-w_2\rangle|.\end{aligned} \tag{25}$$

Moreover, for $N_2>N_1\geq 0$, from (6) we have

$$R^{N_1}v-R^{N_2}v=\sum_{n=N_1}^{N_2-1}s_n\psi_{\gamma_n}. \tag{26}$$

Finally, for any $n\geq 0$ and any $m\geq 0$,

$$\begin{aligned}|\langle R^mv,s_n\psi_{\gamma_n}\rangle|&=|s_n||\langle\psi_{\gamma_n},R^mv\rangle|\leq|s_n|\sup_{i\in I}|\langle\psi_i,R^mv\rangle|\\&\leq\frac{1}{\alpha}|s_n||M_m|.\end{aligned} \tag{27}$$

Let us now consider $N_2>N_1\geq 0$. Using (25)–(27), we obtain

$$\begin{aligned}\|R^{N_1}v-R^{N_2}v\|^2&\leq\|R^{N_1}v\|^2-\|R^{N_2}v\|^2+2\left|\left\langle R^{N_2}v\sum_{n=N_1}^{N_2-1}s_n\psi_{\gamma_n}\right\rangle\right|\\&\leq\|R^{N_1}v\|^2-\|R^{N_2}v\|^2+\frac{2}{\alpha}|M_{N_2}|\sum_{n=N_1}^{N_2-1}|s_n|.\end{aligned} \tag{28}$$

Using (23) of Proposition 2, we know that the sequence $(\|R^nv\|)_{n\in\mathbb{N}}$ is non-negative and non-increasing. Therefore, it converges to some value R_∞ and for any $\varepsilon>0$, there exists $K>0$ such that for all $m>K$,

$$R_\infty^2\leq\|R^mv\|^2\leq R_\infty^2+\varepsilon^2.$$

As a consequence, for any $N_2 > N_1 \geq K$,

$$\|R^{N_1}v - R^{N_2}v\|^2 \leq \varepsilon^2 + \frac{2}{\alpha}|M_{N_2}| \sum_{n=N_1}^{N_2} |s_n|. \tag{29}$$

Using (22), we know that $\sum_{n=0}^{+\infty} |M_n||s_n| < +\infty$. Moreover, $0 \leq |s_n| \leq |M_n|$ for all $n \in \mathbb{N}$. So Lemma 2 (see Appendix) can be applied with $x_n \equiv |s_n|$ and $y_n \equiv |M_n|$. Two situations might occur:

- The first one is that: $\sum_{n=0}^{+\infty} |s_n| < +\infty$. In this case, we know that there is $K' > 0$ such that for any $N_2 > N_1 \geq K'$

  $$\sum_{n=N_1}^{N_2} |s_n| \leq \frac{\alpha}{2\|v\|}\varepsilon^2.$$

  Moreover, from (20) and Cauchy–Schwartz inequality, we know that

  $$|M_{N_2}| = |\langle R^{N_2}v, \psi_{\gamma_{N_2}}\rangle| \leq \|R^{N_2}v\| \leq \|v\|.$$

  So (29) becomes for any $\varepsilon > 0$ there are K and $K' > 0$ such that for any $N_2 > N_1 \geq \max(K, K')$

  $$\|R^{N_1}v - R^{N_2}v\|^2 \leq \varepsilon^2 + \varepsilon^2.$$

  Hence, in the first case, $(R^n v)_{n\in\mathbb{N}}$ is a Cauchy sequence.
- The second one is that: $\liminf_{q\to+\infty} |M_q| \sum_{n=0}^{q} |s_n| = 0$. In this case, let $\varepsilon > 0$ and let $p > 0$ be an integer. We are going to estimate $\|R^m v - R^{m+p}v\|$, for $m > K$ (K is such that (29) holds).
  First, there is $q > m+p$ such that

  $$|M_q| \sum_{n=0}^{q} |s_n| \leq \frac{\alpha}{2}\varepsilon^2. \tag{30}$$

  Moreover, we can decompose

  $$\|R^m v - R^{m+p}v\| \leq \|R^m v - R^q v\| + \|R^{m+p}v - R^q v\|.$$

  Applying (29) with $N_1 = m$ and $N_2 = q$ and using (30) we obtain

  $$\|R^m v - R^q v\|^2 \leq \varepsilon^2 + \varepsilon^2.$$

  Similarly, applying (29) for $N_1 = m+p$ and $N_2 = q$ and using (30) we obtain

  $$\|R^{m+p}v - R^q v\|^2 \leq \varepsilon^2 + \varepsilon^2.$$

  Hence, we finally obtain

  $$\|R^m v - R^{m+p}v\| \leq 2\sqrt{2}\varepsilon,$$

  which proves that $(R^n v)_{n\in\mathbb{N}}$ is a Cauchy sequence in the second case.

Overall, $(R^n v)_{n\in\mathbb{N}}$ converges in both cases. The second statement directly follows from (6). □

4. l^1 norm bounds specific to thresholding functions

In general, when $\mathcal{H}$ is an infinite dimensional space, we have no guarantee that

$$\sum_{n=0}^{+\infty} |s_n| < +\infty. \tag{31}$$

A simple counterexample consists in considering $(\psi_i)_{i\in I}$ an orthonormal or Riesz basis (for definition, see [25]) of $\mathcal{H}$, $v = \sum_{i\in I} s_i\psi_i \in \mathcal{H}$ such that $\sum_{i\in I}|s_i|$ diverges and $\theta(t) \equiv t$.

Below, we prove that (31) exists, whatever $v \in \mathcal{H}$ and whatever the dictionary, as soon as θ is a thresholding function.

Proposition 3. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a* thresholding function. *The quantities defined in Table1 satisfy*

$$\sum_{n=0}^{+\infty} |s_n| \leq \frac{\|v\|^2 - \|R^{+\infty}v\|^2}{\tau^-} \leq \frac{\|v\|^2}{\tau^-}, \tag{32}$$

where $\tau^- > 0$ denotes the internal threshold as defined in Definition2.

Proof. Let $M \in \mathbb{N}$ fixed. Using (20), we know that

$$\sum_{n=0}^{M-1} s_n^2 \leq \|v\|^2 - \|R^M v\|^2.$$

Together with (19), this leads to

$$\sum_{n=0}^{M-1} s_n M_n = \frac{1}{2}\left(\|v\|^2 + \sum_{n=0}^{M-1} s_n^2 - \|R^M v\|^2\right) \leq \|v\|^2 - \|R^M v\|^2.$$

Using (11) and the fact that $\theta(\cdot)$ is a thresholding function, for any $n \in \mathbb{N}$, we have

$$s_n M_n = |s_n||M_n| \geq \tau^- |s_n|,$$

where the last inequality is obtained after discussing the two cases: $s_n = 0$ or $s_n \neq 0$.

As a conclusion for all $M \in \mathbb{N}$ we have

$$\sum_{n=0}^{M-1} |s_n| \leq \frac{\|v\|^2 - \|R^M v\|^2}{\tau^-}. \tag{33}$$

Letting M go to infinity, we obtain (32). □

Remark 1. The above upper bound does not depend on the dictionary $(\psi_i)_{i\in I}$. It holds for any $v \in \mathcal{H}$. We therefore do not expect this bound to be tight in any dedicated or applicative context.

Remark 2. As a side effect, the above proposition guarantees that the coordinates λ_i exist for all $i \in I$ (see (5)). We even know that

$$\sum_{i\in I} |\lambda_i| < +\infty.$$

5. l^0 bounds specific for gap functions

If $\theta(\cdot)$ is a gap function then the MP shrinkage stops automatically after a finite number of iterations.

Proposition 4. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v \in \mathcal{H}$ and $\theta(\cdot)$ be a* gap function (*i.e.* $\mathrm{gap}(\theta) > 0$). *The sequence $(s_n)_{n\in\mathbb{N}}$ defined in Table1 satisfies*

$$\#\{n | s_n \neq 0\} \leq \left\lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2} \right\rfloor.$$

Proof. Suppose that the sequence $(s_n)_{n\in\mathbb{N}}$ contains M nonzero terms. Observing Definition 3, for each $s_n \neq 0$, we

have

$$s_n^2+2s_n(M_n-s_n)\geq \mathrm{gap}(\theta)^2,$$

where we recall that $M_n=\langle R^n v,\psi_{\gamma_n}\rangle$, $s_n=\theta(M_n)$.

From (19), we know that

$$\|v\|^2\geq \sum_{n\in\mathbb{N}:s_n\neq 0}\left(s_n^2+2s_n(M_n-s_n)\right)\geq M\cdot \mathrm{gap}(\theta)^2.$$

Noting that M is integer, we have

$$M\leq \left\lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2}\right\rfloor. \qquad \square$$

Remark 3. An interesting consequence of the proposition is that

$$\#\{i\in I,\lambda_i\neq 0\}\leq \left\lfloor \frac{\|v\|^2}{\mathrm{gap}(\theta)^2}\right\rfloor.$$

In other words, v is approximated with less than $\lfloor \|v\|^2/\mathrm{gap}(\theta)^2\rfloor$ non-zero coordinates.

6. Bound on the residual error

In this section, we are interested in the residual error norm. The result concerns general shrinkage functions. Before stating the result, let us give the following lemma:

Lemma 1. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v\in\mathcal{H}$ and $\theta(\cdot)$ be a* general shrinkage function. *The sequence $(M_n)_{n\in\mathbb{N}}$ defined in Eq.* (4) *satisfies*

$$\limsup_{n\to+\infty} M_n\leq \sup_{t:\theta(t)=0} t \tag{34}$$

and

$$\inf_{t:\theta(t)=0} t\leq \liminf_{n\to+\infty} M_n. \tag{35}$$

Proof. Let us prove the first statement. If $\sup_{t:\theta(t)=0} t=+\infty$ the statement is trivial. We therefore focus on the case $\sup_{t:\theta(t)=0} t<+\infty$. Let us assume that (34) does not hold. Then there exists $\varepsilon>0$ and an increasing sequence $(k_n)_{n\in\mathbb{N}}\in\mathbb{N}^{\mathbb{N}}$ such that

$$M_{k_n}\geq \sup_{t:\theta(t)=0} t+\varepsilon>0, \quad \forall n\in\mathbb{N}.$$

So there exists an increasing sequence $(k_n)_{n\in\mathbb{N}}\in\mathbb{N}^{\mathbb{N}}$ such that

$$s_{k_n}=\theta(M_{k_n})\geq \theta\left(\sup_{t:\theta(t)=0} t+\varepsilon\right)>0.$$

This means that

$$\limsup_{n\to+\infty} s_n>0.$$

The latter statement is impossible since, from (21), we know that $\lim_{n\to+\infty} s_n=0$. This proves (34).

The proof of (35) is similar. $\square$

In particular, if the external threshold of $\theta(\cdot)$ is zero (i.e. $\tau^+=0$),

$$\lim_{n\to+\infty} M_n=0,$$

since $\sup_{t:\theta(t)=0} t=\inf_{t:\theta(t)=0} t=0$.

Recall that we have defined the semi-norm on $\mathcal{H}$ as

$$\|u\|_{\mathcal{D}}\overset{\mathrm{def}}{=}\sup_{i\in I}|\langle u,\psi_i\rangle|, \quad \forall u\in\mathcal{H}.$$

Notice that $\|\cdot\|_{\mathcal{D}}$ is a norm as soon as $\mathcal{D}$ generates $\mathcal{H}$.

The $\|\cdot\|_{\mathcal{D}}$ is important because it penalizes the error made in the direction of the elements of the dictionary much more than in other directions. Geometrically, its level sets

$$\{u\in\mathcal{H},\|u\|_{\mathcal{D}}\leq\tau\}$$

is a polyhedron, for $\tau\geq 0$. When the elements of the dictionary correspond to structures that one expects in the data, by this norm we can have a strong constraint in the direction of these structures and a small constraint in other directions (such as noise). Some basic properties for this semi-norm under discrete settings were given in [35]. We also refer the reader to [3,22,28] for more results along this direction.

Recall that in (8) we denote $V\overset{\mathrm{def}}{=}\overline{\mathrm{Span}((\psi_i)_{i\in I})}$, the closure of vector space spanned by the dictionary $(\psi_i)_{i\in I}$, $V^\perp$ its orthogonal complement and we denote the orthogonal projection onto V and $V^\perp$ by P_V and $P_{V^\perp}$ respectively.

Proposition 5. *Let $(\psi_i)_{i\in I}$ be a normed dictionary, $v\in\mathcal{H}$ and $\theta(\cdot)$ be a* general shrinkage function. *The limits defined in Theorem*1 *satisfy*

$$\left\|\sum_{n=0}^{+\infty} s_n\psi_{\gamma_n}-P_V v\right\|_{\mathcal{D}}=\left\|R^{+\infty}v-P_{V^\perp}v\right\|_{\mathcal{D}}\leq\frac{\tau^+}{\alpha},$$

where τ^+ is the external threshold of $\theta(\cdot)$, as defined in Definition 2.

Proof. Let $\varepsilon>0$, from Lemma 1, we know that for any $k\geq 0$ there is $n_k\geq k$

$$\inf_{t:\theta(t)=0} t-\varepsilon\leq M_{n_k}\leq \sup_{t:\theta(t)=0} t+\varepsilon.$$

Given the definition of τ^+, we therefore know that

$$-\tau^+-\varepsilon\leq M_{n_k}\leq\tau^++\varepsilon.$$

We rewrite

$$|M_{n_k}|\leq\tau^++\varepsilon.$$

Moreover, using the property of projection and given the construction of M_{n_k}, we know that

$$|M_{n_k}|\geq\alpha\sup_{i\in I}|\langle R^{n_k}v,\psi_i\rangle|=\alpha\sup_{i\in I}|\langle R^{n_k}v,P_V(\psi_i)\rangle|=\alpha\sup_{i\in I}|\langle P_V(R^{n_k}v),\psi_i\rangle|.$$

Therefore, for all $i\in I$,

$$|\langle P_V(R^{n_k}v),\psi_i\rangle|\leq\frac{\tau^+}{\alpha}+\frac{\varepsilon}{\alpha}.$$

Since $(R^{n_k}v)_{k\in\mathbb{N}}$ converges to $R^{+\infty}v$ (see Theorem 1), we finally have

$$|\langle P_V(R^{+\infty}v),\psi_i\rangle|\leq\frac{\tau^+}{\alpha}+\frac{\varepsilon}{\alpha},$$

for all $i \in I$. Since the above inequalities hold for any $\varepsilon > 0$, we obtain

$$\left\|P_V(R^{+\infty}v)\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

Moreover, using Theorem 1, we know that

$$P_{V^\perp}(R^{+\infty}v) = P_{V^\perp}(v) - P_{V^\perp}\left(\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n}\right) = P_{V^\perp}(v).$$

We therefore obtain

$$\|R^{+\infty}v - P_{V^\perp}v\|_{\mathcal{D}} = \|P_V(R^{+\infty}v)\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

Using Theorem 1 (again), we also know that

$$\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} = P_V\left(\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n}\right) = P_V(v) - P_V(R^{+\infty}v).$$

Therefore,

$$\left\|\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} - P_V(v)\right\|_{\mathcal{D}} = \left\|P_V(R^{+\infty}v)\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

This finishes the proof of the theorem. □

Remark 4. In the above proposition, if the external threshold τ^+ is zero (this is the case for the matching pursuit), we deduce from the proposition that $\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} - P_V v \in V^\perp$. Therefore, since $\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} - P_V v \in V$, we finally obtain that $\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} = P_V v$. We might have $P_V v \neq v$. However, when $P_V v \neq v$, we cannot expect to obtain a better approximation of v since $\sum_{n=0}^{+\infty} s_n \psi_{\gamma_n} \in V$ and $P_V v$ minimizes the Euclidean distance between v and V.

When $\tau^+ \neq 0$, the algorithms generally do not recover v. The benefit of this approximation is to obtain a decomposition with less significant coordinates (see Propositions 3 and 4).

Remark 5. A consequence of the above proposition is that when the MPS is used with a *thresholding function*, it provides a feasible point for the "Dantzig selector" (see [3]). The "Dantzig selector" consists of the optimization problem:

$$\min_{(\lambda_i)_{i\in I}} \sum |\lambda_i| \quad \text{subject to} \quad \left\|\sum_{i\in I} \lambda_i \psi_i - P_V v\right\|_{\mathcal{D}} \leq \frac{\tau^+}{\alpha}.$$

From Proposition 3, we know that the MPS provides a set of coordinates $(\lambda_i)_{i\in I}$ (see (5)) such that

$$\min_{(\lambda_i)_{i\in I}} \sum |\lambda_i|$$

is finite. Proposition 5 guarantees that the constraint is satisfied.

7. Experimental results

This section is devoted to the comparison of the MP shrinkage algorithm with some classical sparse representation methods: the regular MP, OMP and BPDN. In all the experiments, the predefined constant α equals 1. For simplicity, the only shrinkage function considered in these experiments is the soft-thresholding function (see (14)).

7.1. Denoising with a translation invariant wavelet dictionary

We report the experiments on the Pepper image with pixel values in [0,255]. The dictionary contains all the translations of the Daubechies _3 wavelets decomposed at the level 4. In 2D, this makes 13 convolution kernels. The original image is contaminated by Gaussian noise of standard variation $\sigma = 20$. Both the original and noisy images are displayed in Fig. 2.

In this experiment, we run the MPS (again the shrinkage function is the soft thresholding function) for $\tau \in \{0,10,50,100\}$. We remind that $\tau = 0$ corresponds to the usual MP. The iterative process is stopped once one of the following two conditions is met: (a) the l^2-norm of the residual is smaller than σ; (b) the length of the forward step is negligible: $|s_n| \leq 10^{-6}$.

We display the graph of the sequence $n \to |s_n|$ on the left side of Fig. 3. We see on these curves that, as expected, when τ increases $|s_n|$ decays to 0 more rapidly. Hence, numerically we observe that the MP shrinkage with τ rather big converges much faster than MP.

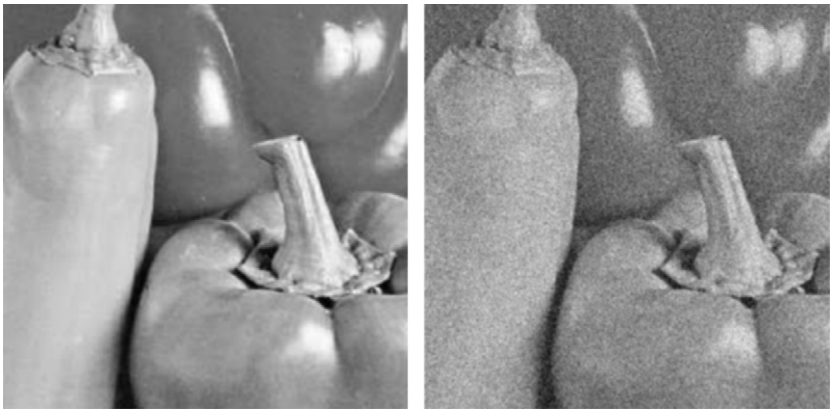

Fig. 2. Experiments on Pepper with wavelet dictionary. Left: clean image; right: noisy image, $PSNR = 22.10$.

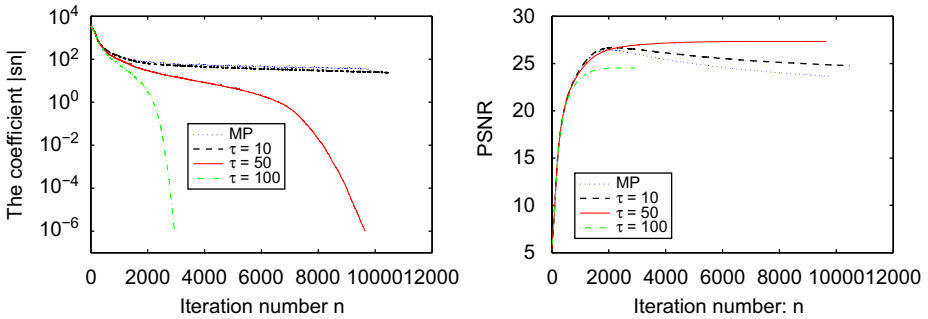

Fig. 3. The MP shrinkage with the soft thresholding function for $\tau=0$ (i.e. MP), $\tau=10$, 50 and 100. Left: the curves $n\to|s_n|$. Right: the PSNR along the iterative process.

This behavior illustrates Proposition 5 since, indeed

$$\left\|\sum_{i=0}^{n-1} s_i\psi_{\gamma_i}-P_V v\right\|_{\mathcal{D}}=\left\|\sum_{i=0}^{n-1} s_i\psi_{\gamma_i}-v\right\|_{\mathcal{D}}=|M_n|\le|s_n|+\tau.$$

The right side of Fig. 3 displays, for the different values of τ, the PSNR between the image obtained at the iteration n and the original clean image. This quantity is shown for all the values of n.

During the first iterations (say when $n<1000$) the PSNR are similar for all the values of τ. However, later in the iterative process (say for $n>2500$), the curves corresponding to $\tau=10$ and 50 are above the other curves. Notice also that, a wisely chosen stopping criterion based on the residual norm could lead to an earlier stop, with a stronger denoising effect. Indeed, on the right of Fig. 3, the supreme of the PSNR curve corresponding to $\tau=0$ (i.e. the MP) is 26.45 (for $n=2057$). This would correspond to the result obtained if an oracle had told us when to stop the algorithm. However, it is still smaller than the PSNR obtained with the MP shrinkage at convergence, for $\tau=50$. The PSNR, in this case is indeed 27.33. For fairness, let us highlight that the choice of τ could be critical for the MPS.

Overall, we numerically observed that the MP shrinkage with positive threshold converges faster than MP and meanwhile returns a better denoised result.

7.2. Detection of letters

We now consider a detection problem in a very noisy case. The original image and the noisy image are on the top row of Fig. 5. The values of the pixels of the original image (Fig. 5(a)) range in [0,255]. On the original image appears several letters whose shape is known (see Fig. 4). The noisy image (Fig. 5(b)) is obtained by adding to this original image a Gaussian noise of standard deviation $\sigma=150$.

The purpose of the detection is to recover the letters. In order to do so, we build a translation invariant dictionary by translating the letters displayed in Fig. 4. Notice that though some couples of elements of the dictionary are very correlated, this problem resemble a compressed sensing since it is an inverse problem (and not just an approximation problem) with a ground truth representation which might be recovered exactly. Interestingly, this problem

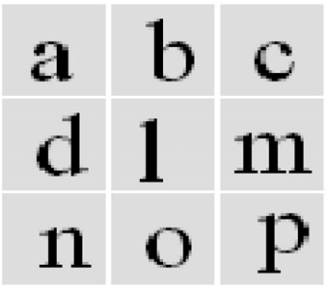

Fig. 4. The nine letters used to construct the dictionary. Each letter is extended by zero-padding to reach the size of the noisy image. It is then translated over the plane.

resembles more some machine learning approaches where sparsity is exploited, since it implies a decision/classification problem.

In order to illustrate the difficulty of this problem, we show in Fig. 5(c) the image obtained when denoising the noisy image with a soft thresholding in a wavelet dictionary. The letters of the original image cannot be seen in the denoised image. The purpose of this experiment is to illustrate that denoising the image with a "general purpose" denoising method before doing the detection is not likely to work.

In Fig. 5(d) we display the result obtained with the basis pursuit denoising (BPDN) (see (7) or [5,10,12,34]), with the dictionary described above. The parameter β (the weight for the l^1 term in (7)) for the BPDN has been carefully tuned to get better result. The negative coefficients are not displayed since they cannot represent letters. This makes some small difference as the white/negative letters (which are less reasonable in our case) are removed from the result image.

Some letters which we are looking for appear in Fig. 5(d). In the point view of component analysis [31,20], the noisy image v here should be regarded as

$$v=X+B+N,$$

where X is the clean images formed by letters, B is the background structure and N is the noise. Hence, the extremely strong noise level, the interference of the complex background structure, together with the coherence of the dictionary made the BPDN partially successful for this task.

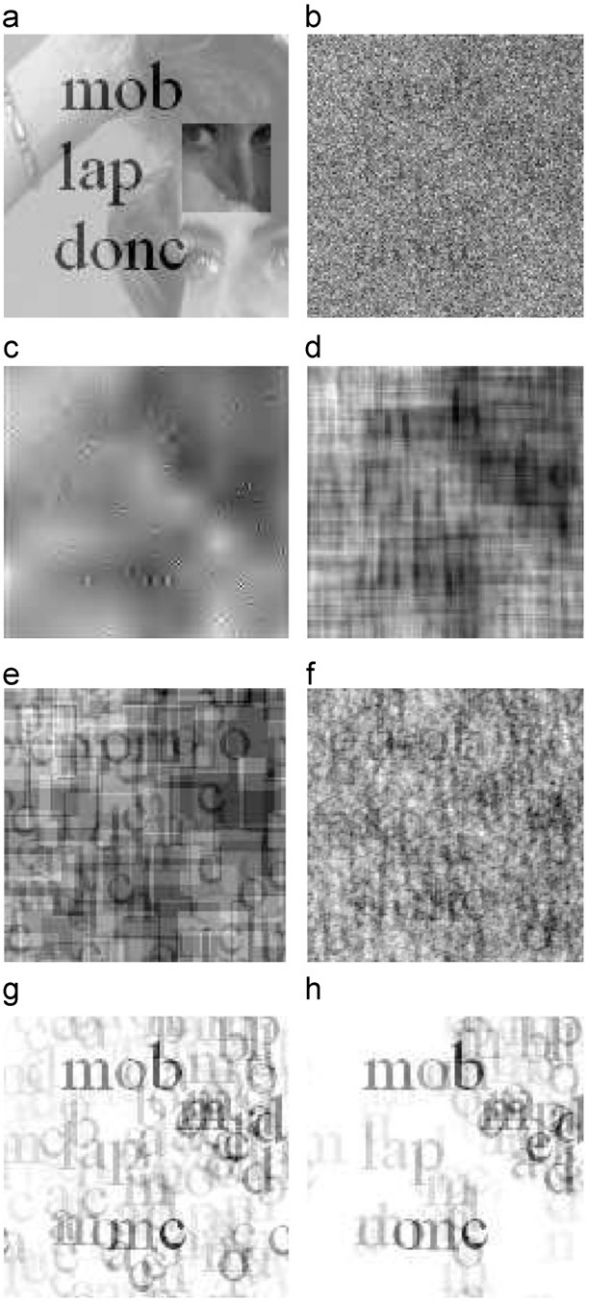

Fig. 5. Detecting letters with a coherent dictionary in the presence of a strong noise (a) clean image; (b) noisy image $\sigma = 150$; (c) wavelet soft shrinkage ($\tau = 400$); (d) BPDN ($\beta = 400$); (e) OMP; (f) MP; (g) denoised MP; (h) MPS with a soft-thresholding shrinkage function ($\tau = 400$).

In Fig. 5(e) and (f), we display the results obtained with OMP and MP. These algorithms are stopped once the l^2-norm of the residual is less or equal to σ. Again, the negative coefficients are set to 0 for the display. Experimentally, the OMP stops after fewer iterations than MP. This leads to less coefficients and less noise. Those two results are however poor. Again, it seems that the noise level and the coherence of the dictionary are too important for these algorithms.

In order to denoise these images, we have tried to soft-threshold their coordinates. The result of this thresholding (at the threshold 400) on the result of the MP is shown in Fig. 5(g). Below we refer to this algorithm as denoised MP. Experimentally we observe that the same scheme works rather limitedly for OMP and BPDN. In these experiments, the denoised MP provides better results than the denoised OMP and denoised BPDN.

Finally, the result of the MP shrinkage is displayed in Fig. 5(h). Again it is applied with a soft-thresholding shrinkage function. The image corresponds to the threshold $\tau = 400$. Evidently, the choice of τ is critical, it is mandatorily tuned around 3σ to obtain good qualitative result. The algorithm is stopped once $|s_n| \leq 10^{-6}$ (this leads to about 500 iterations). Again, the negative coordinates are set to 0 for the display. The MPS leads to a better detection than the other methods. In particular, the word "donc" is recovered and this was not the case with the other algorithms. The difference between the denoised MP and MPS results is due to the impact of the false detection on the remaining iterations. This impact is better controlled with MPS than with denoised MP.

On this experiment, another point is that the false detections are located in the dark area of the original image. It seems therefore possible to detect these false detection by a post-processing. The false detection of the denoised MP are less structured and might be more difficult to detect.

To summarize, as the observed image is very noisy and the ideal image is sparse over the dictionary, the MP shrinkage with positive threshold performances much better than the MP and other related algorithms.

8. Conclusion and perspectives

In this paper, we investigate a modification of the MP algorithm. Its main characteristics are:

1. Similar to the MP, the results evolves along one element of the dictionary at a time.
2. Similar to the MP, it is greedy: once an element of the dictionary has been selected, this choice is no longer questioned.
3. Unlike MP and depending on the choice of the shrinkage function, the evolution in the direction of a selected atom can be slowed. The benefit of being cautious is to limit the consequences of bad selections of atoms.

Because it is a simple variant of the MP (and in particular satisfy the above items 1 and 2), a fast (real time for audio processing) version can be implemented (see [21]). However, because of item 3, MPS can present benefits similar to the algorithms that do not satisfy the above item 2 (such as CoSaMP [29]).

In the current paper, we propose a segmentation of the general shrinkage functions in three classes: general shrinkage functions, thresholding functions and gap functions. For each category we establish convergence guarantees for the MPS. Altogether, these results suggest that the convergence and the decay rate of the coordinate

(assessed in terms of l^2, l^1 and l^0 norms) mostly depend on the behavior of the shrinkage function in the vicinity of $\{t : \theta(t) = 0\}$.

Obviously, the behavior of $\theta(t)$ for large values of $|t|$ impacts the number of iterations needed to converge. If $\theta(t)$ is small for $|t|$ large, we will generally select an atom several times in raw before selecting another atom. This will of course slow the algorithm in proportion of the cost of an iteration (which is small). The benefit is to limit the negative impact of false selection as much as this strategy permits to limit it.

The main perspectives of this work are to

- Perform a more complete experimental study on the impact of the shrinkage function in several contexts such as denoising, source separation, compressed sensing, approximation, etc. The purpose of this study should be to better understand how to chose the shrinkage function.
- Perform a theoretical analysis evaluating the performance of MPS in a compressed sensing setting. In particular, the error made on wrongly selected atoms will be smaller than with the standard MP and OMP. The limitation of this error limit the risk of false detection in the remaining iterations.
- Perform a theoretical analysis evaluating the performance of MPS in the context of approximation. As mention above, being cautious leads to more iterations. However, it is unclear if these extra iterations increase the number of selected atoms or not. This might depend on the shrinkage function and also on the coherence of the dictionary. Moreover, the choice of the shrinkage functions, the theoretical guidance of the parameter will also be very interesting.

Acknowledgment

This work was supported in part by RGC 203109, RGC 211710, the FRGs of Hong Kong Baptist University, and the PROCORE-France/Hong Kong Joint Research Scheme sponsored by the Research Grant Council of Hong Kong and the Consulate General of France in Hong Kong F-HK/05/08T. We would like to thank Prof. Alain Trouvé (ENS-Cachan), Prof. Michael Ng (HKBU) and Dr. Li Xiaolong (Peking Univ.) for the helpful discussions. We would also like to thank all the anonymous reviewers as their comments and suggestions contributed greatly to the quality of the paper.

Appendix A

A.1. Proof of Proposition 1

*Proof of item*1: Let $t_0 \in \mathbb{R}$ be such that $t_0 > \inf_{t:\theta(t)\neq 0}|t|$. We cannot simultaneously have $\theta(t_0) = 0$ and $\theta(-t_0) = 0$, since $\theta(\cdot)$ is non-decreasing. Let us denote

$$t = \begin{cases} t_0 & \text{if } \theta(t_0) \neq 0, \\ -t_0 & \text{if } \theta(t_0) = 0. \end{cases}$$

We have $\theta(t) \neq 0$ and given the definition of the gap, we know that

$$\mathrm{gap}(\theta)^2 \leq \theta(t)^2 + 2\theta(t)(t-\theta(t)), = t^2 - (t-\theta(t))^2 \leq t^2 = t_0^2.$$

As a conclusion, for any t_0 such that $t_0 > \inf_{t:\theta(t)\neq 0}|t|$, we have $\mathrm{gap}(\theta) \leq t_0$. So

$$\mathrm{gap}(\theta) \leq \inf_{t:\theta(t)\neq 0} |t|.$$

*Proof of item*2: Let us first deduce from the definition of general shrinkage functions and (11) that for all $t \in \mathbb{R}$

$$0 \leq \theta(t)^2 \leq |t||\theta(t)| = t\theta(t).$$

Therefore for all $t \in \mathbb{R}$:

$$t\theta(t) \leq 2t\theta(t) - \theta(t)^2 \leq 2t\theta(t).$$

We then immediately obtain that

$$\inf_{t:\theta(t)\neq 0} t\theta(t) \leq \mathrm{gap}(\theta)^2 \leq 2 \inf_{t:\theta(t)\neq 0} t\theta(t).$$

In order to prove the last inequality of item 2, let us first remark that, from the definition of a general shrinkage function and for t and $t' \in \mathbb{R}$

$$\begin{aligned} &\text{if } t \geq t' \geq 0 \text{ then } \theta(t) \geq \theta(t') \geq 0 \\ &\quad \text{and } t\theta(t) \geq t'\theta(t') \geq 0, \\ &\quad \text{if } t \leq t' \leq 0 \text{ then } \theta(t) \leq \theta(t') \\ &\qquad \leq 0 \text{ and } t\theta(t) \geq t'\theta(t') \geq 0. \end{aligned}$$

As a consequence, if we consider $\varepsilon > 0$ and $t \in \mathbb{R}$

$$\text{if } t \geq \tau^+ + \varepsilon \text{ then } t\theta(t) \geq (\tau^+ + \varepsilon)\theta(\tau^+ + \varepsilon) \geq \inf_{t:\ |t| \geq \tau^+ + \varepsilon,\ g(t)\neq 0} t\theta(t)$$

$$\text{if } t \leq -\tau^+ - \varepsilon \text{ then } t\theta(t) \geq (-\tau^+ - \varepsilon)\theta(-\tau^+ - \varepsilon) \geq \inf_{t:\ |t| \geq \tau^+ + \varepsilon,\ g(t)\neq 0} t\theta(t)$$

Therefore, for any $\varepsilon > 0$

$$\inf_{t:g(t)\neq 0} t\theta(t) = \inf_{t:\ |t| \geq \tau^+ + \varepsilon,\ g(t)\neq 0} t\theta(t).$$

It is then straightforward that, for $t \in \mathbb{R}$ such that $|t| \leq \tau^+ + \varepsilon$,

$$t\theta(t) = |t||\theta(t)| \leq (\tau^+ + \varepsilon)|\theta(t)|.$$

Therefore, for any $\varepsilon > 0$

$$\inf_{t:g(t)\neq 0} t\theta(t) \leq (\tau^+ + \varepsilon) \inf_{t:g(t)\neq 0} |\theta(t)|.$$

This immediately leads to the last unproved inequality.

A.2. Lemma used in the proof of Theorem 1

This lemma is a variation on the lemma used for the proof of Theorem 1 in [27].

Lemma 2. *Let $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ be two non-negative sequences of reals such that*

$$\sum_{k=0}^{+\infty} x_k y_k < +\infty.$$

One of the following alternatives holds:

- *either*

$$\sum_{k=0}^{+\infty} x_k < +\infty$$

- *or*

$$\liminf_{j\to+\infty} y_j \sum_{k=0}^{j} x_k = 0.$$

Proof. First, since $(y_k)_{k\in\mathbb{N}}$ is a sequence of nonnegative real numbers, its inferior limit always exists. We

- either have $\liminf_{k\to+\infty} y_k > 0$
- or $\liminf_{k\to+\infty} y_k = 0$.

Let us first assume that

$$\liminf_{k\to+\infty} y_k > 0.$$

There exists $\varepsilon > 0$ and $n > 0$ such that for any $k \geq n$, one has $y_k \geq \varepsilon$. Therefore, we have

$$\varepsilon \sum_{k=n}^{+\infty} x_k \leq \sum_{k=n}^{+\infty} x_k y_k < +\infty$$

and finally

$$\sum_{k=0}^{+\infty} x_k < +\infty.$$

The first alternative holds.

Let us now assume that

$$\liminf_{k\to+\infty} y_k = 0$$

and consider $\varepsilon > 0$ and $m \geq 0$. Since $\sum_{k=0}^{+\infty} x_k y_k < +\infty$, there is $n \geq m$ such that

$$\sum_{k=n}^{+\infty} x_k y_k < \frac{\varepsilon}{2}. \tag{36}$$

Since $\liminf_{k\to+\infty} y_k = 0$, there is $p \geq 0$ such that

$$y_{n+p} < \frac{1}{2\sum_{k=0}^{n-1} x_k}\varepsilon. \tag{37}$$

Let $j \in \{n, \ldots, n+p\}$ be such that

$$y_j \leq y_k, \quad \forall k \in \{n, \ldots n+p\}. \tag{38}$$

We have

$$\begin{aligned} y_j \sum_{k=0}^{j} x_k &= y_j \sum_{k=0}^{n-1} x_k + y_j \sum_{k=n}^{j} x_k \leq y_{n+p} \sum_{k=0}^{n-1} x_k + y_j \sum_{k=n}^{j} x_k \quad \text{from (38)} \\ &< \frac{\varepsilon}{2} + \sum_{k=n}^{j} x_k y_k \quad \text{from (37) and (38)} \\ &< \varepsilon \quad \text{from (36)}. \end{aligned}$$

As a conclusion, for any $\varepsilon > 0$ and any $m \geq 0$, there is $j \geq m$ such that

$$y_j \sum_{k=0}^{j} x_k < \varepsilon.$$

This means that the second alternative holds. □

References

[1] T. Blumensath, M.E. Davies, Iterative hard thresholding for compressed sensing, Appl. Comput. Harmon. Anal. 27 (3) (2009) 265–274.

[2] T. Blumensath, M.E. Davies, Gradient pursuits, IEEE Trans. Signal Process. 56 (6) (2008) 2370–2382.

[3] E. Candès, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Ann. Statist. 35 (6) (2007) 2313–2351.

[4] C. Chaux, L. Duval, A. Benazza-Benyahia, J.-C. Pesquet, A nonlinear Stein-based estimator for multichannel image denoising, IEEE Trans. Signal Process. 56 (8) (2008) 3855–3870.

[5] S.S. Chen, D. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.

[6] P.L. Combettes, J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, SIAM J. Optim. 18 (4) (2007) 1351–1376.

[7] S. Cotter, J. Adler, R. Rao, and K. Kreutz-Delgado, Forward sequential algorithms for best basis selection, in: IEE, Proceedings in Vision, Image and Signal Processing, vol. 146, no. 5, October 1999, pp. 235–244.

[8] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inform. Theory 55 (5) (2009) 2230–2249.

[9] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, Constr. Approx. 13 (1) (1997) 57–98.

[10] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 57 (11) (2004) 1413–1457.

[11] R. DeVore, V. Temlyakov, Some remarks on greedy algorithms, Adv. Comput. Math. 5 (1996) 173–187.

[12] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Statist. 1 (2) (2007) 302–332.

[13] J. Friedman, W. Stuetzle, Projection pursuit regression, J. Amer. Statist. Assoc. 76 (376) (1981) 817–823.

[14] P. Huber, Projection pursuit, Ann. Statist. 15 (2) (1985) 435–525.

[15] H.-Y. Gao, Wavelet shrinkage denoising using the non-negative garrote, J. Comput. Graph. Statist. 7 (4) (1998) 469–488.

[16] H.-Y. Gao, A.G. Bruce, WaveShrink with firm shrinkage, Statist. Sinica 7 (4) (1997) 855–874.

[17] R. Gribonval, E. Bacry, Harmonic decomposition of audio signals with matching pursuit, IEEE Trans. Signal Process. 51 (1) (2003) 101–111.

[18] R. Gribonval, M. Nielsen, Approximate Weak Greedy Algorithms, Advances in Computational Mathematics 14 (4) (2001) 361–378.

[19] R. Gribonval, P. Vandergheynst, On the exponential convergence of matching pursuits in quasi-incoherent dictionaries, IEEE Trans. Inform. Theory 52 (1) (2006) 255–261.

[20] M. Kowalski, B. Torrésani, Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients, Signal Image Video Process. 3 (3) (2009) 251–264.

[21] S. Krstulovic, R. Gribonval, MPTK: Matching Pursuit made tractable, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), vol. 3, Toulouse, France, May 2006, pp. 496–499.

[22] S. Lintner, F. Malgouyres, Solving a variational image restoration model which involves L^∞ constraints, Inverse Problem 20 (3) (2004) 815–831.

[23] B. Mailhé, R. Gribonval, F. Bimbot, P. Vandergheynst, Low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries, in: IEEE ICASSP 2009, Taiwan.

[24] B. Mailhé, Boris, R. Gribonval, P. Vandergheynst, F. Bimbot, Fast orthogonal sparse approximation algorithms over local dictionaries, Signal Process., doi:10.1016/j.sigpro.2011.01.004.

[25] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, San Diego, 1998.

[26] G. Davis, S. Mallat, Z. Zhang, Adaptive time–frequency decompositions, SPIE J. Opt. Eng. 33 (7) (1994) 2183–2191.

[27] S. Mallat, Z. Zhang, Matching pursuits with time–frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397–3415.

[28] Y. Meyer, Oscillating Patterns in Image Processing and Nonlinear Evolution Equations, The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures, American Mathematical Society, Boston, USA, 2001.

[29] D. Needell, J. Tropp, CoSaMP: iterative recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. 26 (3) (2009) 301–321.

[30] Y. Pati, R. Rezaiifar, P. Krishnaprasad, Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition, in: Proceedings of 27th Asimolar Conference on Signals, Systems and Computers, Los Alamitos, 1993.

[31] J.-L. Starck, M. Elad, D. Donoho, Image decomposition via the combination of sparse representation and a variational approach, IEEE Trans. Image Process. 14 (10) (2005) 1570–1582.
[32] B.L. Sturm, J.J. Shynk, Sparse approximation and the pursuit of meaningful signal models with interference adaptation, IEEE Trans. Audio Speech Language Process. 18 (3) (2010) 461–472.
[33] V. Temlyakov, Greedy algorithms and *m*-term approximation with regard to redundant dictionaries, J. Approx. Theory 98 (1) (1999) 117–145.
[34] T. Wu, K. Lange, Coordinate descent algorithms for lasso penalized regression, Ann. Appl. Statist. 2 (1) (2008) 224–244.
[35] T. Zeng, Études de modèles variationnels et apprentissage de dictionnaires, Ph.D. Thèsis, Université Paris Nord, October 2007.
[36] Z. Zhao, Wavelet shrinkage denoising by generalized threshold function, in: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18–21 August 2005.