Sparse Regression Using Mixed Norms

Matthieu Kowalski

Sparse Regression using Mixed Norms

Matthieu Kowalski
LATP, CMI, Université de Provence,
39, rue F. Joliot-Curie,
13453 Marseille cedex

Abstract

Mixed norms are used to introduce in an easy way, both structures and sparsity in the framework of sparse regression problems, and then introduce implicitly couplings between regression coefficients. Corresponding algorithms are described and analyzed. Besides the classical sparse regression problem, at the same time the multi-layered expansion of signals are considered, using union of dictionaries. These sparse structured expansions are done subject to equality constraint, using a modified FOCUSS algorithm. When the mixed norms are used in the framework of regularized inverse problem, a thresholded Landweber iteration is used to minimize the corresponding variational problem.

Key words: Sparse regression, Structured regression, Mixed norms, FOCUSS, Thresholded Landweber iterations

1. Introduction

During the last few years, sparsity has emerged as a general principle for signal modeling. In a few words, whenever a signal may be represented sparsely, i.e. characterized by a small amount of data, many signal processing tasks turn out to become significantly easier. Among the success of sparse methods, one may mention application to signal coding and compression [7], denoising (Basis pursuit denoising and related techniques [4]), source separation [10] and many others.

Most sparsity based approaches start by expanding signals on a given waveform family (basis, frame, dictionary, ...), and process the coefficients of the expansion individually. Therefore, a coefficient independence assumption is implicitly done, although this latter assumption is generally an over simplification. A good example is provided by the (sparse) time-frequency representations displayed in figure 2, where significant coefficients are clearly organized in structured sets (vertical or horizontal lines). Such structures are clearly not accounted for when coefficients are treated individually.

The two main contributions of this paper are the following: we propose sparse expansion methods that explicitly introduce a notion of *structured sparsity*. We then combine this approach with multilayered signal expansion approaches, which aim at decomposing signals as sums of significantly different components (termed "layers") (see [3,8,7,22]).

Structured sparsity is modeled by introducing a coupling between coefficients in the same structured set. Some probabilistic approaches introduced this kind of structured with success (see [10,14]). In the framework

Email address: kowalski@cmi.univ-mrs.fr Phone: +33-(0)491054743 Fax: +33-(0)491054742 (Matthieu Kowalski).

of variational formulations, such a coupling may be introduced by suitable regularization terms, that combine sparsity and persistence.

Such regularization terms have been considered by Fornasier and Rauhut [11] and Teschke and Ramlau [25], under the name of joint sparsity: they have studied mixed norms $\ell_{p,1}$, focusing on multichannel signals. Our approach is based on mixed norms, which may be introduced whenever signal expansions on double labeled families are considered [1] :

$$s = \sum_{i,j} \alpha_{i,j} \phi_{i,j} \ .$$

where $\{\phi_{i,j}\}$ are the waveforms of a given basis or frame.

The mixed $\ell_{p,q}$ norm is then defined as

$$\|\alpha\|_{p,q} = \left(\sum_i \left(\sum_j |\alpha_{i,j}|^p \right)^{q/p} \right)^{1/q} \ ,$$

and we shall be mainly concerned with the regression problem

$$\min_{\alpha} \left[\|s - \sum_{i,j} \alpha_{i,j} \phi_{i,j}\|_2^2 + \lambda \|\alpha\|_{p,q}^q \right] \ ,$$

with $\lambda > 0$ a fixed parameter.

When $\{\phi\}_{i,j}$ is a basis, we give practical estimates for the regression coefficients $\alpha_{i,j}$, summarized in theorem 3, obtained by generalized soft thresholding. When $\{\phi\}_{i,j}$ is a frame, the latter estimates may be plugged in a Landweber iteration scheme to yield a minimizer of the corresponding functional.

In the case of the *exact reconstruction* regression problem

$$\min_{\alpha} \|\alpha\|_{p,q}^q \quad \text{such that} \quad s = \sum_{i,j} \alpha_{i,j} \phi_{i,j} \ ,$$

we also show that the FOCUSS algorithm [18] may be adapted to yield minimizers of that functional.

The extension to multilayered signal expansion exploits several doubly labeled families: in the case of two layers, one seeks expansions of the form

$$s = \sum_{i,j} \alpha_{i,j} \phi_{i,j} + \sum_{k,\ell} \beta_{k,\ell} \psi_{k,\ell} \ ,$$

with prescribed sparsity and persistence assumptions on the coefficients sets α and β. In a variational formulation, we consider regression problem

$$\min_{\alpha,\beta} \left[\|s - \sum_{i,j} \alpha_{i,j} \phi_{i,j} - \sum_{k,\ell} \beta_{k,\ell} \psi_{k,\ell}\|_2^2 + \lambda \|\alpha\|_{p,q}^q + \mu \|\beta\|_{p',q'}^{q'} \right] \ .$$

The thresholded Landweber iterations are studied with this more general formulation, and a modification of the FOCUSS algorithm is provided if an exact reconstruction estimate is required.

The paper is organized as follows. The mixed norms are defined in section 2, and we gives an overview of how some mixed norms are used in the literature. Section 3 used the FOCUSS algorithm to minimize these norms subject to an equality constraint, and extend the algorithm for multilayered expansion.

The noisy signal estimation problem is study in the section 4: after the recall of corresponding regularized FOCUSS algorithm, we discuss the thresholded Landweber iteration to minimize the corresponding functional. Section 5 gives a simple illustration of the possible uses of the algorithms, in order to give an idea of the mixed norms are relevant to easily model dependence between coefficients.

[1] The approach of [11,25] clearly applies directly to this more general situation.

The paper deals with the finite dimension case. This choice was made because of the attention carried on the practical algorithms which can be used.

2. Mixed norms

This section recalls the definition of the weighted mixed norms we shall be concerned with and some useful properties. To our knowledge, the corresponding mixed norm spaces were introduced and studied in [2], and some main properties are summarized in [20].

We use the following notations. Let the vector $\mathbf{x} \in \mathbb{C}^N$, $N = K \times F$ labeled using a double index, be such that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ and for all $k \in \{1, \ldots, K\}$, $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,\nu}, \ldots, x_{k,F})$.

This double index is purely conventional and is used to introduce some dependences between coefficients [2].

Definition 1 *Let $\mathbf{w} \in \mathbb{R}^N$, with $N = K \times F$, be such that for all $(k, \nu) \in \{1, \ldots, K\} \times \{1, \ldots, F\}$ $w_{k,\nu} > 0$. Let $p \geq 1$ and $q \geq 1$. We call (weighted) mixed norm of $\mathbf{x} \in \mathbb{C}^N$, the norm $\ell_{\mathbf{w};p,q}$ defined by*

$$\|\mathbf{x}\|_{\mathbf{w};p,q} = \left(\sum_{k=1}^{K} \left(\sum_{\nu=1}^{F} w_{k,\nu} |x_{k,\nu}|^p \right)^{q/p} \right)^{1/q} . \tag{1}$$

The cases $p = +\infty$ and $q = +\infty$ can be obtained by replacing the corresponding norm by the supremum.

The mixed norm $\ell_{\mathbf{w};p,q}$ can be seen as a "composition" of the norms $\ell_{\mathbf{w};p}$ and ℓ_q:

$$\begin{aligned} \|\mathbf{x}\|_{\mathbf{w};p,q} &= \left(\sum_{k=1}^{K} \|\mathbf{x}_k\|_{\mathbf{w};p}^q \right)^{1/q} \\ &= \left\| \left(\sum_{\nu=1}^{F} W_{.,\nu} |\mathbf{x}_{.,\nu}|^p \right)^{1/p} \right\|_q , \end{aligned} \tag{2}$$

where $|\mathbf{x}_{.,\nu}|^p$ is obtained from the vector $\mathbf{x}_{.,\nu}$ by raising each component modulus at power p, and where $W_{.,\nu} = \mathrm{diag}(w_{1,\nu}, \ldots, w_{K,\nu})$. When there is no ambiguity, this type of notation will be used in the following.

Remark 1 *The mixed norms generalize the usual p norms if $p = q$: $\|\mathbf{x}\|_{\mathbf{w};p,p} = \|\mathbf{x}\|_{\mathbf{w};p}$.*

Mixed norms explicitly introduce coupling between coefficients instead of the usual independence assumption behind the ℓ_p norms. This point will be more explicit in section 4, with the Bayesian formulation of the regression problem.

However, one can see that the coupling is strongly dependent of the choice of p and q. ℓ_p norms are usually used as measures of diversity for small values of p and sparsity for large values. Mixed norms allow one to mix those two concepts. Used as regularization terms in a regression context, they enforce some specific types of joint sparsity and diversity, as we shall see below.

In such a framework, to ensure the convergence to a global optimum, it is useful to guarantee the convexity of the problem. This convexity is given by the following proposition

Proposition 1 *If $p \geq 1$ and $q \geq 1$ then the norm $\ell_{\mathbf{w};p,q}$ is convex. The strict convexity is obtained for $p > 1$ and $q > 1$.*

PROOF. This property is a consequence of the homogeneity of the norm and the triangle inequality. ∎

[2] For example, $\mathbf{x}$ can denote the coefficients of a time-frequency or a time-scale expansion of a signal. In this particular context, we shall denote by (k, ν) the index, with k the time index and ν the frequency index. Then, $x_{k,\nu}$ represents the coefficient at time k and frequency ν, the vector $\mathbf{x}_k$ represents all the frequency coefficients at time k, and we denote by $\mathbf{x}_{.,\nu}$ the vector which contains all the time coefficients at frequency ν.

The mixed norms allow us to favor some type of structures we can find in signals. Classical properties of the norms (and, in particular, the convexity), allow us to use them in optimization problems, and then regression problems. The next subsection recalls some models already studied which use a mixed norm to group variables.

2.1. *Some mixed norms in the literature*

Some specific instances of mixed norms are already used in various situations. This section presents some functional spaces characterized by mixed norms, and some statistical regression problems which use particular mixed norms.

2.1.1. *Characterization of some functional spaces*

First, let us recall a few examples of functional spaces that can be defined in terms of mixed norms. In this paragraph, we leave an instant the finite dimensional case (the definition of mixed norm to infinite space is straightforward).

The Besov, Triebel-Lizorkin spaces and the modulation spaces are characterized with mixed norms. Besov and Triebel-Lizorkin spaces are described in [19], and [9] is a good overview for modulation spaces.

Let $s \in \mathbb{R}$ and $0 < p, q \leq \infty$. Let $\phi_0 \in \mathcal{S}$ following some specific properties (see [19]) and $\phi_j(t) = 2^j \phi_0(2^j t)$. The Besov space $B^s_{p,q}$ is defined by

$$B^s_{p,q} = \left\{ f \in \mathcal{S}' : \|f\|_{B^s_{p,q}} = \left(\sum_j 2^{jsq} \|\phi_j * f\|_p^q \right)^{1/q} < \infty \right\} .$$

The Besov norm $\|f\|_{B^s_{p,q}}$ is indeed a mixed norm with $\mathbf{w} = \{2^{jsq}\}$. For some particular ranges of values of p, q, s, Besov spaces are known to be spaces of sparse functions (or distributions), i.e. spaces within which nonlinear approximation converge faster than linear ones [16,5].

The Triebel-Lizorkin space $F^s_{p,q}$ is defined by

$$F^s_{p,q} = \left\{ f \in \mathcal{S}' : \|f\|_{F^s_{p,q}} = \left\| \left(\sum_j 2^{jsq} |\phi_j * f|^q \right)^{1/q} \right\|_p < \infty \right\} .$$

The Triebel-Lizorkin norm $\|f\|_{F^s_{p,q}}$ is also a mixed norm with $\mathbf{w} = \{2^{jsq}\}$. In comparison with Besov spaces, the roles of the two indices are interchanged [13].

The modulation spaces are characterized too with mixed norms. Let $\{g_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ be a Gabor frame. f belongs to the modulation space $\mathbf{M}^w_{p,q}$ if and only if [21]

$$\left(\sum_{m\in\mathbb{Z}} \left(\sum_{n\in\mathbb{Z}} w^p_{m,n} |\langle f, g_{m,n} \rangle|^p \right)^{q/p} \right)^{1/q} < \infty .$$

2.1.2. *Statistical regression and inverse problems*

Several sparse regression techniques were studied in a supervised learning context. The most classical one is the ℓ_1 regression, known as the lasso estimate [26], well known in the signal processing community as the Basis pursuit denoising [4].

The group-lasso [27] introduced by Yan and Lin, use the mixed norm $\ell_{2,1}$. This norm was introduced to preserve some entire groups of individuals. More recently Fornasier and Rauhut studied more generally some $\ell_{p,1}$ mixed norm for inverse problems in [11], and Teschke and Ramlau in [25].

Another example is provided by the hierarchical penalization [23] introduced by Szafranski and Grandvalet leading to a mixed norm $\ell_{\frac{4}{3},1}$ which is obtained after a hierarchical modeling of the variables.

In the statistical community, the mixed norm were used by Peng Zhao *et al* in [28], under the name of "Composite Absolute Penalties". That paper studies more particularly algorithms using the $\ell_{\infty,1}$ mixed norm.

Mixed norms are now introduced. In the following sections, we focuss on regression problems. We show how classical algorithms used with ℓ_p norms can be adapted with mixed norms. Moreover, we give general algorithms able to handle the structure in layers for signal, like the morphological component analysis [22]. Although the $\ell_{p,1}$-like mixed norms were specifically used in the literature, we give results for general $\ell_{p,q}$ mixed norms, and in particular, we show that the $\ell_{1,q}$ mixed norms are relevant too.

3. Signal estimation under equality constraint

We first address the problem of function or signal estimation subject to an equality constraint. We prove that the FOCUSS algorithm can be adapted to tackle the case of mixed norms, and generalize it for a decomposition into layers.

Let $\mathbf{y} \in \mathbb{C}^M$. Let $\mathbf{x} \in \mathbb{C}^N$ with $N = K \times F$ be such that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ and for all k, $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,\nu}, \ldots, x_{k,F})$. Let $A \in \mathbb{C}^{M \times N}$ be a matrix whose columns are the vectors of a dictionary of $\mathbb{C}^M$, with $M \leq N$. We consider the cases $p \leq 2$ and $q \leq 2$ ($p, q \neq 0$) (with the straightforward definition of the "mixed norms" for $p < 1$ or $q < 1$). These cases allow us to promote sparsity in some directions of the index lattice, and the problem remains convex for $1 \leq p, q \leq 2$.

3.1. *Minimization of a mixe norm*

We want to solve the problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \operatorname{sgn}(pq)\|\mathbf{x}\|^q_{\mathbf{w};p,q} \\ \text{subject to} \quad & \mathbf{y} = A\mathbf{x} \ . \end{aligned} \tag{3}$$

The problem (3) can be solved with the FOCUSS algorithm [18], which has been designed to minimize a sparsity measure (or diversity measure, with $sparsity = \frac{1}{diversity}$), subject to an equality constraint.

In order to use FOCUSS, we have to write the gradient of the diversity measure under a factorized form. Denoting by $E(\mathbf{x})$ the diversity measure of $\mathbf{x}$, the factorized form of the gradient reads

$$\nabla_{\mathbf{x}} E(\mathbf{x}) = \alpha(\mathbf{x})\Pi(\mathbf{x})\bar{\mathbf{x}} \ .$$

where $\bar{\mathbf{x}}$ denote the conjugate of $\mathbf{x}$.

For example, with the ℓ_p diversity measure $E^{(p)}(\mathbf{x}) = \|\mathbf{x}\|_p^p$, $\alpha(\mathbf{x}) = |p|$ and $\Pi(\mathbf{x}) = \operatorname{diag}(|x_n|^{p-2})$. In our case, we choose the $\ell_{\mathbf{w};p,q}$ diversity measure, i.e. $E^{(p,q)}(\mathbf{x}) = \operatorname{sgn}(pq)\|\mathbf{x}\|^q_{\mathbf{w};p,q}$. The partial derivative with respect to $x_{k,\nu}$ is

$$\frac{\partial E^{(p,q)}(\mathbf{x})^q}{\partial x_{k,\nu}} = \operatorname{sgn}(p)|q| \ w_{k,\nu} \ \bar{x}_{k,\nu}|x_{k,\nu}|^{p-2}\|\mathbf{x}_k\|^{q-p}_{\mathbf{w}_k;p} \ ,$$

Then

$$\alpha(x) = \operatorname{sgn}(p)|q| \text{ and } \Pi(x) = \operatorname{diag}(w_{k,\nu}|x_{k,\nu}|^{p-2}\|\mathbf{x}_k\|^{q-p}_{\mathbf{w}_k;p}) \tag{4}$$

FOCUSS is a simple iterative scheme to solve (3) given by the following algorithm:

Algorithm 1
Let $x^{(0)} \in \mathbb{C}^N$ *be a bounded feasible solution of (3)*
Do
$\mathbf{x}^{(m+1)} = \Pi^{-1}(\mathbf{x}^{(m)})A^*(A\Pi^{-1}(\mathbf{x}^{(m)})A^*)^{-1}\mathbf{y}$
until *convergence*

where a feasible solution $x^{(0)}$ of (3) is a vector which satisfies the constraint equality of the problem, and A^* denotes the Hermitian transpose of A.

Theorem 1 *Starting from a bounded feasible solution $x^{(0)} \in \mathbb{R}^N$, the sequence of iterates generated by algorithm 1 is convergent, and minimizes the $\ell_{\mathbf{w};p,q}$ diversity measure.*

PROOF. We rewrite and adapt the original proof in [18]. The main point is to prove that $E^{(p,q)}(\mathbf{x}^{(m+1)}) < E^{(p,q)}(\mathbf{x}^{(m)})$.

To prove the convergence of the algorithm, we have to check the assumptions of the global convergence theorem [1], which we restate here for the sake of completeness.

Theorem 2 *Let $\mathcal{A}$ be an algorithm on a set X, and suppose that, given $x^{(0)}$, a sequence $\{x^{(m)}\}$ is generated, satisfying*

$$x^{(m+1)} = \mathcal{A}(x^{(m)}) \ .$$

Let a solution set $\Gamma \subset X$ be given, and suppose the following
(i) All points $x^{(m)}$ are contained in a compact set $S \subset X$.
(ii) There is a continuous function (the descent function) Z on X such that
(a) If $x \notin \Gamma$, then $Z(y) < Z(x)$, $\forall y \in \mathcal{A}(x)$;
(b) If $x \in \Gamma$, then $Z(y) \leq Z(x)$, $\forall y \in \mathcal{A}(x)$;
(iii) The mapping $\mathcal{A}$ is closed at point outside Γ.
Then, the limit of any convergent sub-sequence of $x^{(m)}$ is a solution, and $Z(x^{(m)}) \to Z(x^\star)$ for some $x^\star \in \Gamma$.

We define here

$$\Gamma = \left\{ x^\star : Ax^\star = y, \text{ and } x^\star = P^\star (AP^\star)^+ y \right\} \ ,$$

where $P^\star = (\Pi^{-1}(x^\star))^{1/2}$ and $+$ denote the Moore-Penrose pseudo-inverse.

Point *iii*) does not matter, because $\mathcal{A}$ is here a continuous function. Point *i*) is an immediate consequence, once the point *ii*) is proved.

To prove point *ii*), we use the same technique as the one used in the original proof, and make use of Hölder's inequality: if $x_i,\ y_i \geq 0$, $r > 1$, $\frac{1}{r} + \frac{1}{s} = 1$, then

$$\sum_i x_i y_i \leq \left(\sum_i x_i^r \right)^{\frac{1}{r}} \left(\sum_i y_i^s \right)^{\frac{1}{s}} \ .$$

The inequality is reversed for $r < 1$ ($r \neq 0$).

One can write $\mathbf{x}^{(m+1)}$ as a function of the solution $b^{(m+1)}$ with minimal ℓ_2 norm, of the problem $AP^{(m+1)}b = y$ where $P^{(m+1)} = \mathrm{diag}(w_{k,\nu}^{-1/2} |x_{k,\nu}^{(m)}|^{\frac{2-p}{2}} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{\frac{p-q}{2}})$. Then we have $\mathbf{x}^{(m+1)} = P^{(m+1)} b^{(m+1)}$

Let us introduce $\tilde{b}$ such that $\tilde{b}_{k,\nu} = w_{k,\nu}^{1/2} \mathrm{sgn}(x_{k,\nu}^{(m)}) |x_{k,\nu}^{(m)}|^{\frac{q}{2}} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{q-p}{2}}$ which is a suboptimal solution of $AP^{(m+1)}b = y$. If $\mathbf{x}^{(m+1)} \neq \mathbf{x}^{(m)}$ (i.e. the algorithm has not converged), then

$$\begin{aligned} \|b^{(m+1)}\|_2^2 < \|\tilde{b}\|_2^2 &= \sum_k \sum_\nu w_{k,\nu} |x_{k,\nu}^{(m)}|^p \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{q-p} \\ &= \sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{q-p} \sum_\nu w_{k,\nu} |x_{k,\nu}^{(m)}|^p \\ &= \sum_k \|\mathbf{x}_p^{(m)}\|_{\mathbf{w}_k;p}^{q-p} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^p = \|\mathbf{x}^{(m)}\|_{\mathbf{w};p,q}^q \ . \end{aligned} \tag{5}$$

We can stress that

$$x_{k,\nu}^{(m+1)} = |x_{k,\nu}^{(m)}|^{\frac{2-p}{2}} \|\mathbf{x}_k^{(m+1)}\|_{\mathbf{w}_k;p}^{\frac{p-q}{2}} \ b_{k,\nu}^{(m+1)} \ ,$$

and then, for $0 < p < 2$ and $0 < q < 2$

$$
\begin{aligned}
E(\mathbf{x}^{(m+1)}) &= \sum_k \left(\sum_\nu w_{k,\nu} |x_{k,\nu}^{(m+1)}|^p \right)^{q/p} \\
&= \sum_k \left(\sum_\nu w_{k,\nu} |x_{k,\nu}^{(m)}|^{\frac{p(2-p)}{2}} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{p(p-q)}{2}} |b_{k,\nu}^{(m+1)}|^p \right)^{q/p} \\
&= \sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{q(p-q)}{2}} \left(\sum_\nu w_{k,\nu} |x_{k,\nu}^{(m)}|^{\frac{p(p-2)}{2}} |b_{k,\nu}^{(m+1)}|^p \right)^{q/p} \\
&\leq \sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{q(p-q)}{2}} \left(\sum_\nu w_{k,\nu} |x_{k,\nu}^{(m)}|^p \right)^{\frac{q(2-p)}{2p}} \left(\sum_\nu |b_{k,\nu}^{(m+1)}|^2 \right)^{\frac{pq}{2p}} \\
&= \sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{q(p-q)}{2}} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{q(2-p)}{2}} \|\mathbf{b}_k^{(m+1)}\|_2^q \\
&= \sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{\frac{2q-q^2}{2}} \|\mathbf{b}_k^{(m+1)}\|_2^q \\
&\leq \left(\sum_k \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^q \right)^{\frac{2-q}{2}} \left(\sum_k \|\mathbf{b}_k^{(m+1)}\|_2^2 \right)^{\frac{q}{2}} \\
&< \left(\|\mathbf{x}^{(m)}\|_{\mathbf{w};p,q}^q \right)^{\frac{2-q}{2}} \left(\|\mathbf{x}^{(m)}\|_{\mathbf{w};p,q}^q \right)^{\frac{q}{2}} \\
&= E(\mathbf{x}^{(m)}) \ ,
\end{aligned}
$$

where we have used in the $4th$ line the Hölder inequality with $r = \frac{2}{2-p},\ s = \frac{2}{p}$, and in the $7th$ line the Hölder inequality with $r = \frac{2}{2-q},\ s = \frac{2}{q}$.

Point ii) is then proved. The cases $p = 2$ or $q = 2$ are simple enough to not be specifically written. The cases $p < 0$ or $q < 0$ are similar, but we used the reversed Hölder inequality.

In order to prove i), we just have to prove that the sequence $\|x^{(m)}\|$ is bounded. However, as the function E is decreasing, we have $|x_{k,\nu}^{(m)}| \leq (E(x^{(0)}))^{\frac{1}{q}}$, which allows us to conclude. ∎

Remark 2 *To initialize the algorithm, a simple bounded feasible solution is the solution with the minimum ℓ_2 norm obtained by the Moore-Penrose pseudo inverse of A:* $\mathbf{x}^{(0)} = A^+\mathbf{y}$.

3.2. *Extension to multilayered expansions*

In some situations, the signals under consideration contain significantly different features (termed *layers*), which are accurately encoded using different bases or frames. This leads to regression problems with dictionaries built as unions of these bases. Then, it makes sense to use different (mixed) norms on the corresponding coefficients. FOCUSS may be adapted to such a new situation, as we show below.

For simplicity, we limit the present discussion to the case of two layers only. The generalization to arbitrary numbers of layers is straightforward. The problem can be formulated as follows. Let $\mathbf{y} \in \mathbb{C}^M$ and $\mathbf{x} = \begin{pmatrix} \mathbf{x}^{[1]} \\ \mathbf{x}^{[2]} \end{pmatrix} \in \mathbb{C}^N$, with $M < N$ and for all $i \in \{1, 2\}$, $\mathbf{x}^{[i]} \in \mathbb{R}^{N_i}$, with $N_i = K_i \times F_i$. Suppose that $\mathbf{x}^{[i]} = (\mathbf{x}_1^{[i]}, \ldots, \mathbf{x}_{K_i}^{[i]})$ and for all $k \in \{1, \ldots, K_i\}$, $\mathbf{x}_k^{[i]} = (x_{k,1}^{[i]}, \ldots, x_{k,\nu}^{[i]}, \ldots, x_{K_i,F_i}^{[i]})$. Let $A \in \mathbb{C}^{M \times N}$ be such that $A = [A_1 A_2]$, with $A_i \in \mathbb{C}^{M \times N_i}$ for $i \in \{1, 2\}$. Now that all the notations are introduced, we want to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{x}_1,\mathbf{x}_2} \quad & \operatorname{sgn}(p_1q_1)\lambda_1 \ \|\mathbf{x}^{[1]}\|^{q_1}_{\mathbf{w}^{[1]},p_1,q_1} + \operatorname{sgn}(p_2q_2)\lambda_2\|\mathbf{x}^{[2]}\|^{q_2}_{\mathbf{w}^{[2]},p_2,q_2} \\ \text{subject to} \quad & \mathbf{y} = A\mathbf{x} = A_1\mathbf{x}^{[1]} + A_2\mathbf{x}^{[2]} \ , \end{aligned} \tag{6}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$ fixed.

We denote the diversity measure by

$$E(\mathbf{x}) = \operatorname{sgn}(p_1q_1)\lambda_1 \ \|\mathbf{x}^{[1]}\|^{q_1}_{\mathbf{w}^{[1]},p_1,q_1} + \operatorname{sgn}(p_2q_2)\lambda_2\|\mathbf{x}^{[2]}\|^{q_2}_{\mathbf{w}^{[2]},p_2,q_2} \tag{7}$$

$$= \operatorname{sgn}(p_1q_1)\lambda_1 \sum_{k_1=1}^{K_1}\left(\sum_{\nu_1=1}^{F_1} w^{[1]}_{k_1,\nu_1}|x^{[1]}_{k_1,\nu_1}|^{p_1}\right)^{q_1/p_1} + \operatorname{sgn}(p_2q_2)\lambda_2 \sum_{k_2=1}^{K_2}\left(\sum_{\nu_2=1}^{F_2} w^{[2]}_{k_2,\nu_2}|x^{[2]}_{k_2,\nu_2}|^{p_2}\right)^{q_2/p_2} \tag{8}$$

$$= E_1(\mathbf{x}^{[1]}) + E_2(\mathbf{x}^{[2]}) \ . \tag{9}$$

In order to write the gradient of E in factorized form, we calculate the partial derivatives

$$\frac{\partial E(\mathbf{x})}{\partial x^{[1]}_{k_1,\nu_1}} = \operatorname{sgn}(p_1)\lambda_1|q_1|w^{[1]}_{k_1,\nu_1}x^{[1]}_{k_1,\nu_1}|x^{[1]}_{k_1,\nu_1}|^{p_1-2}\|\mathbf{x}^{[1]}_{k_1}\|^{q_1-p_1}_{\mathbf{w}^{[1]},p_1} \tag{10}$$

$$\frac{\partial E(\mathbf{x})}{\partial x^{[2]}_{k_2,\nu_2}} = \operatorname{sgn}(p_2)\lambda_2|q_2|w^{[2]}_{k_2,\nu_2}x^{[2]}_{k_2,\nu_2}|x^{[2]}_{k_2,\nu_2}|^{p_2-2}\|\mathbf{x}^{[2]}_{k_2}\|^{q_2-p_2}_{\mathbf{w}^{[2]},p_2} \ . \tag{11}$$

The gradient of the energy can be written in factorized form with $\alpha(x) = 1$ and $\Pi(\mathbf{x}) = \begin{pmatrix} \Pi_1(x^{[1]}) & 0 \\ 0 & \Pi_2(\mathbf{x}^{[2]}) \end{pmatrix}$ where

$$\Pi_1(\mathbf{x}^{[1]}) = \operatorname{sgn}(p_1)\lambda_1|q_1| \ \operatorname{diag}(w^{[1]}_{k_1,\nu_1}|x^{[1]}_{k_1,\nu_1}|^{p_1-2}\|\mathbf{x}^{[1]}_{k_1}\|^{q_1-p_1}_{\mathbf{w}^{[1]},p_1})$$

and

$$\Pi_2(\mathbf{x}^{[2]}) = \operatorname{sgn}(p_2)\lambda_2|q_2| \ \operatorname{diag}(w^{[2]}_{k_2,\nu_2}|x^{[2]}_{k_2,\nu_2}|^{p_2-2}\|\mathbf{x}^{[2]}_{k_2}\|^{q_2-p_2}_{\mathbf{w}^{[2]},p_2}) \ .$$

A first idea would be to apply the FOCUSS algorithm 1 without any change. Unfortunately, this resulting algorithm does not converge in this case: the diversity measure E is not decreasing during the iterations. To modify the algorithm in order to ensure the decrease of the diversity measure and so the convergence, let us take again the ideas of the preceding proof, with the same notations. We rewrite the inequality (5):

$$\|b^{(m+1)}\|_2^2 = \|b^{[1]^{(m+1)}}\|_2^2 + \|b^{[2]^{(m+1)}}\|_2^2 \tag{12}$$

$$< \|\tilde{b}\|_2^2 = \|\mathbf{x}^{(m)}\|^q_{\mathbf{w};p,q} \tag{13}$$

$$= \|\mathbf{x}^{[1]^{(m)}}\|^q_{\mathbf{w};p,q} + \|\mathbf{x}^{[2]^{(m)}}\|^q_{\mathbf{w};p,q} \ . \tag{14}$$

To prove the strict decrease of E during the iterations, we would like to have $\|b^{[1]^{(m+1)}}\|_2^2 < \|\mathbf{x}^{[1]^{(m)}}\|^q_{\mathbf{w};p,q}$ (resp. $\|b^{[2]^{(m+1)}}\|_2^2 < \|\mathbf{x}^{[2]^{(m)}}\|^q_{\mathbf{w};p,q}$) and $\|b^{[2]^{(m+1)}}\|_2^2 \leq \|\mathbf{x}^{[2]^{(m)}}\|^q_{\mathbf{w};p,q}$ (resp. $\|b^{[1]^{(m+1)}}\|_2^2 \leq \|\mathbf{x}^{[1]^{(m)}}\|^q_{\mathbf{w};p,q}$). Then, we slightly modify the FOCUSS algorithm to guarantee the strict decrease of the energy

Algorithm 2

Let $x^{(0)} \in \mathbb{R}^N$ *be a bounded feasible solution.*

Do

$\mathbf{x}^{[1]^{(m+1)}} = \Pi_1^{-1}(\mathbf{x}^{[1]^{(m)}})A_1^*(A\Pi^{-1}(\mathbf{x}^{(m)})A^*)^{-1}\mathbf{y}$

$\mathbf{x}^{[2]^{(m+1)}} = \Pi_2^{-1}(\mathbf{x}^{[2]^{(m)}})A_2^*(A\Pi^{-1}(\mathbf{x}^{(m)})A^*)^{-1}\mathbf{y}$

if $E(\mathbf{x}^{(m+1)}) \geq E(\mathbf{x}^{(m+1)})$ **then**

if $E_1(\mathbf{x}^{[1]^{(m+1)}}) \geq E_1(\mathbf{x}^{[1]^{(m)}})$, **then** $\mathbf{x}^{[1]^{(m+1)}} = \mathbf{x}^{[1]^{(m)}}$ **endif**

if $E_2(\mathbf{x}^{[2]^{(m+1)}}) \geq E_2(\mathbf{x}^{[2]^{(m)}})$, **then** $\mathbf{x}^{[2]^{(m+1)}} = \mathbf{x}^{[2]^{(m)}}$ **endif**

endif

until *convergence*

In this way, we are sure to obtain a strict decrease of E_1 or E_2 during the iterations, so that the energy E is strictly decreasing. That ensures the convergence of the algorithm to the desired result.

The FOCUSS algorithms allow us to estimate the coefficients in a dictionary, under an exact signal reconstruction constraint. The observation of a signal is often noisy, so it can be useful to relax the strict equality constraint when estimating the signal and its layers. FOCUSS and the above generalization can be modified as in [17] to take the noise into account; however this does not seem to be the most efficient algorithm in this case, as we shall see in the next section, where an alternative approach is also proposed.

4. Signal estimation in the presence of noise

We are interested here by the noisy case, i.e. the case of observations of the form $\mathbf{y} = A\mathbf{x}+b$, where b is an unspecified noise. We follow the classical variational formulation of the problem: the regression is classically made by minimizing the ℓ_2 error between the observed signal and its estimate, regularized by a mixed norm. So that we consider the following functional to minimize

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\|\mathbf{x}\|_{\mathbf{w},p,q}^q \ , \tag{15}$$

with $\lambda \in \mathbb{R}_+$.

In a Bayesian setting, the choice of ℓ_2 norm for the data fidelity term is generally justified by assuming a Gaussian i.i.d. distribution for the noise. The mixed norm leads to a coefficient prior of the form

$$p(\mathbf{x}) = \exp\{-\frac{\lambda}{q}\|\mathbf{x}\|_{\mathbf{w},p,q}^q\} \tag{16}$$

$$= \prod_{k=1}^{K} \exp\{-\frac{\lambda}{q}\|\mathbf{x}_k\|_p^q\} \ , \tag{17}$$

which is a product of Gibbs distributions. This Bayesian formulation shows the coupling between coefficients as stressed in section 2.

More generally, in a multilayered signal decomposition setting (see subsection 3.2 above) our aim is to deal with a functional like:

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \sum_{i=1}^{I} \frac{\lambda_i}{q_i}\|\mathbf{x}^{[i]}\|_{\mathbf{w}_i,p_i,q_i}^{q_i} \ , \tag{18}$$

with $\lambda_i \in \mathbb{R}_+$.

In order to minimize (18), the FOCUSS algorithm can be adapted as suggested in [17]. We provide here the modified algorithm for $I = 2$ only for the sake of simplicity

Algorithm 3
Let $x^{(0)} \in \mathbb{R}^N$ *be a bounded feasible solution.*
Do
$\mathbf{x}^{[1]^{(m+1)}} = \Pi_1^{-1}(\mathbf{x}^{[1]^{(m)}})A_1^*(A\Pi^{-1}(\mathbf{x}^{(m)})A^* + Id)^{-1}\mathbf{y}$
$\mathbf{x}^{[2]^{(m+1)}} = \Pi_2^{-1}(\mathbf{x}^{[2]^{(m)}})A_2^*(A\Pi^{-1}(\mathbf{x}^{(m)})A^* + Id)^{-1}\mathbf{y}$
if $E(\mathbf{x}^{(m+1)}) \geq E(\mathbf{x}^{(m+1)})$ **then**
if $E_1(\mathbf{x}^{[1]^{(m+1)}}) \geq E_1(\mathbf{x}^{[1]^{(m)}})$, **then** $\mathbf{x}^{[1]^{(m+1)}} = \mathbf{x}^{[1]^{(m)}}$ **endif**
if $E_2(\mathbf{x}^{[2]^{(m+1)}}) \geq E_2(\mathbf{x}^{[2]^{(m)}})$, **then** $\mathbf{x}^{[2]^{(m+1)}} = \mathbf{x}^{[2]^{(m)}}$ **endif**
endif
until *convergence*

with

$$\Pi_1(x^{[1]}) = \mathrm{sgn}(p_1)\lambda_1|q_1| \ \mathrm{diag}(w_{k_1,\nu_1}^{[1]}|x_{k_1,\nu_1}^{[1]}|^{p_1-2}\|\mathbf{x}_{k_1}^{[1]}\|_{\mathbf{w}^{[1]},p_1}^{q_1-p_1}) \ , \tag{19}$$

and

$$\Pi_2(x^{[2]}) = \mathrm{sgn}(p_2)\lambda_2|q_2| \ \mathrm{diag}(w_{k_2,\nu_2}^{[2]}|x_{k_2,\nu_2}^{[2]}|^{p_2-2}\|\mathbf{x}_{k_2}^{[2]}\|_{\mathbf{w}^{[2]},p_2}^{q_2-p_2}) \ . \tag{20}$$

Unlike the algorithm given by Rao *et al* in [17], one can remark the presence of the term $(A\Pi^{-1}(\mathbf{x}^{(m)})A^* + Id)$ instead of $(A\Pi^{-1}(\mathbf{x}^{(m)})A^* + \lambda Id)$. This is because in our case, the λ_i must be integrated in the matrix Π_i to be able to write the algorithm.

This algorithm can perform the minimization for any $p, q < 2$ ($p, q \neq 0$). However, it is not very efficient and becomes very slow for high dimensional problems, due to the matrix inversion involved inside the iteration.

A valuable alternative is provided by thresholded Landweber iteration algorithm, like the algorithm introduced by Daubechies in [6]. We study in the next subsection the simple case where A is orthogonal (and corresponds to the operator of an orthogonal basis). Then, the iterative thresholding algorithm is developed for $1 \leq p, q \leq 2$. Although this algorithm is more restricted in terms of admissible values for p and q than FOCUSS, it is really faster.

4.1. *The orthogonal case*

In the whole subsection, A is assumed to be an orthogonal matrix. This allows us to introduce some useful operators corresponding to a given mixed norm. Moreover, the regression problem formulated in the orthogonal case give a good idea of the influence of the mixed norms when they are used as a penalty term.

The regression in the orthogonal case is equivalent to the optimization problem:

$$\min_{\mathbf{x}} \left[\frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\|\mathbf{x}\|_{\mathbf{w};p,q}^q \right] , \tag{21}$$

which can be written like the minimization with respect to $\mathbf{x}$ of

$$\begin{aligned} \Phi(\mathbf{x}) =& \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\left(\sum_{k=1}^{K} \left(\sum_{\nu=1}^{F} w_{k,\nu}|x_{k,\nu}|^p \right)^{q/p} \right) \\ =& \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\left(\sum_{k=1}^{N} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^q \right)^{1/q} . \end{aligned} \tag{22}$$

Several cases have to be taken into account

– $p > 1$ and $q > 1$ The functional Φ is differentiable at all points.
– $p > 1$ and $q = 1$ The functional Φ is not differentiable at points $\mathbf{x}_k = \mathbf{0}$.
– $p = 1$ and $q \geq 1$ The functional Φ is not differentiable at points $x_{k,\nu} = 0$.

As A is an orthogonal matrix, we have

$$\|\mathbf{y} - A\mathbf{x}\|_2^2 = \|A^*\mathbf{y} - \mathbf{x}\|_2^2 = \sum_{k,\nu} ([A\mathbf{y}]_{k,\nu} - x_{k,\nu})^2$$

Denoting by A^* the adjoint of A and by $\theta_{x_{k,\nu}}$ (resp. $\theta_{[A^*y]_{k,\nu}}$) the argument of $x_{k,\nu}$ (resp. $[A^*y]_{k,\nu}$), we have for all k, ν

$$|[A^*y]_{k,\nu} - x_{k,\nu}|^2 = |[A^*y]_{k,\nu}|^2 + |x_{k,\nu}|^2 - 2|x_{k,\nu}||[A^*y]_{k,\nu}| \cos(\theta_{[A^*y]_{k,\nu}} - \theta_{x_{k,\nu}}) .$$

Then, one can differentiate the functional Φ with respect to the modulus of $x_{k,\nu}$, for a fixed couple k, ν, and obtains

$$|x_{k,\nu}| = |[A^*y]_{k,\nu}| \cos(\theta_{[A^*y]_{k,\nu}} - \theta_{x_{k,\nu}}) - \lambda w_{k,\nu}|x_{k,\nu}|^{p-1}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} . \tag{23}$$

The differentiation of Φ with respect to $\theta_{x_{k,\nu}}$ gives

$$2|x_{k,\nu}||[A^*y]_{k,\nu}| \sin(\theta_{[A^*y]_{k,\nu}} - \theta_{x_{k,\nu}}) = 0 . \tag{24}$$

From (23) and (24), one can deduce that $\theta_{x_{k,\nu}} = \theta_{[A^*y]_{k,\nu}}$ and state that variational equations are equivalent to the following system:

$$|x_{k,\nu}| = |[A^*\mathbf{y}]_{k,\nu}| - \lambda w_{k,\nu}|x_{k,\nu}|^{p-1}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} \tag{25}$$

$$\arg(x_{k,\nu}) = \arg([A^*\mathbf{y}]_{k,\nu}) \tag{26}$$

The variational equations are coupled, so that the solution seems difficult to obtain. The following discussion shows that an analytical solution can be obtained in most cases, otherwise an iterative algorithm is given to obtain the solution. It will appear that the solution is obtained by a soft-thresholding operation, as suggested by the variational equations. In all cases, the argument of $x_{k,\nu}$ is the same as $[A^*y]_{k,\nu}$.

4.1.1. *$p > 1$ and $q > 1$*

Let us introduce the function $\mathcal{F} : \mathbb{R}^F \to \mathbb{R}^F$, $|\mathbf{v}| \mapsto \mathcal{F}(|\mathbf{v}|) = |\mathbf{v}| + \lambda W_k|\mathbf{v}|^{p-1}\|\mathbf{v}\|_{\mathbf{w};p}^{q-p}$. $\mathcal{F}$ is bijective, and the system has a unique solution which can be obtained numerically.

4.1.2. *$p > 1$ and $q = 1$*

For the particular case $p = 2$ and $q = 1$, proposition 2 below gives an analytical expression of the solution. The more general cases $1 < p < 2$ and $q = 1$ are solved using the fixed point algorithm 4 (see below).

Proposition 2 *Let A be an orthogonal matrix. We suppose that, for all $k \in \{1, \dots, K\}$, and for all $\nu \in \{1, \dots, F\}$, $w_{k,\nu} = w_k$. The solution of the problem (21), where $\ell_{p,q} = \ell_{2,1}$ is given by*

$$|x_{k,\nu}| = |[A^*\mathbf{y}]_{k,\nu}| \left(1 - \frac{\lambda\sqrt{w_k}}{\|[A^*\mathbf{y}]_k\|_2}\right)^+ .$$

PROOF. The proof is postponed to appendix A.1. ∎

Let us stress that the weighting term which allows us to obtain the solution from $[A^*\mathbf{y}]_{k,\nu}$, depends only of the index k, and does not depend of the index ν. So that, we can rewrite the solution vectorially:

$$|\mathbf{x}_k| = |[A^*\mathbf{y}]_k| \left(1 - \frac{\lambda\sqrt{w_k}}{\|[A^*\mathbf{y}]_k\|_2}\right)^+ . \tag{27}$$

Remark 3 *The result given in proposition 2 shows a mixture of a ℓ_2-like weighting, and a thresholding (remembering the ℓ_1 minimization) which work on a entire group of variables. In this case, the groups with a lot of "big"(or "significant") coefficients are kept rather than the groups with small coefficients. The coupling appears to be between significant coefficients.*

Furthermore, let us stress that this solution is identical to the group-lasso estimate, given in [27].

For the more general case $q = 1$ and $1 < p < 2$ we cannot give any analytical expression for the solution. But an iterative thresholding algorithm can be constructed.

Algorithm 4

Let $\mathbf{x}^{(0)} = [A^*\mathbf{y}]$

For *all* k, ν **do**

$$|x_{k,\nu}^{(m+1)}| = \begin{cases} |[A^*\mathbf{y}]_{k,\nu}| - \lambda w_{k,\nu}|x_{k,\nu}|^{p-1}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} & \text{if } |[A^*\mathbf{y}]_{k,\nu}| > \lambda w_{k,\nu}|x_{k,\nu}|^{p-1}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} \\ 0 & \text{if } |[A^*\mathbf{y}]_{k,\nu}| \leq \lambda w_{k,\nu}|x_{k,\nu}|^{p-1}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} \end{cases}$$

endfor

Proposition 3 *The fixed point algorithm 4 converges for $\lambda < \frac{F^{\frac{2(p-1)}{p(2-p)}}}{2(p-1)} \min_k(\|\mathbf{y}_k\|_{\mathbf{w}_k;p-2})$.*

PROOF. The proof is postponed to appendix A.2 ∎

Remark 4 *Despite the fact that the solution is not given by an analytical expression, the result is obtained by successive soft-thresholding. So that, the solution corresponds to a soft-thresholding.*

Fornasier and Rauhut studied more specifically this kind of mixed norms in [11]. They show that the solution is given by a soft-thresholding operator and gives the analytical solutions for the norms ℓ_1, $\ell_{2,1}$ and $\ell_{\infty,1}$. The study of the threshold operator for the $\ell_{p,1}$ mixed norm is also done by Teschke and Ramlau in [25] for non linear inverse problems.

4.1.3. *$p = 1$ and $q > 1$*

We show here that the solution is obtained again by a soft-threholding. Next proposition gives the threshold analytically for the case $q = 2$, and shows how to obtain a numerical estimation for the other cases.

Proposition 4 *Let A be an orthogonal matrix. The solution of the problem (21), where $\ell_{p,q} = \ell_{1,q}$ is given by a soft thresholding operator. The threshold ξ_k is the (unique) solution on $\mathbb{R}_+$ of the following equation*

$$\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} F_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} \|[A^*\mathbf{y}]_k\|_{\mathbf{w}_k;1}$$

with $F_{\mathbf{w}_k} = \sum_{l=1}^{F} w_{k,l}$.

In particular, for $q = 2$, the threshold ξ_k is equal to

$$\frac{\lambda}{1 + F_{\mathbf{w}_k}\lambda} \|[A\mathbf{y}]_k\|_{\mathbf{w}_k;1} \ .$$

PROOF. The proof is postponed to appendix A.3. ∎

In the case $q = 2$, let us stress some properties of the $\ell_{1,2}$ norm in the following two remarks. These properties illustrate the expected behavior of this mixed norm.

Remark 5 *After suitable rewriting of the solution, the solution is obtained by a mixture of ℓ_2-like weighting and ℓ_1-like soft thresholding:*

$$\begin{aligned} |x_{k,\nu}| &= |[A^*y]_{k,\nu}| - \frac{\lambda}{1 + \lambda F_{\mathbf{w}_k}} \|[A^*y]_k\|_{\mathbf{w}_k;1} \\ &= |[A^*y]_{k,\nu}| \left(1 - \frac{\lambda w_{k,\nu}}{1 + \lambda F_{\mathbf{w}_k}}\right) - \frac{\lambda}{1 + \lambda F_{\mathbf{w}_k}} \sum_{l=1, l\neq\nu}^{K} w_{k,l} |[A^*y]_{k,l}| \ . \end{aligned}$$

This mixture of weighting and thresholding is here very different from the one we obtained for the case of the mixed norm $\ell_{2,1}$. Here, we weight each coefficient, before comparing it to a threshold which depends of the norm of the group k: hence each coefficient is soft thresholded. The sparsity is then enforced while preserving the structure: the threshold depends of the sparsity according to ν, at index k. The sparser the signal at index k, the smaller the threshold.

Contrary to the $\ell_{2,1}$ mixed norm, the coupling is not between the significant coefficients: a coefficient appears significant if the others are unsignificant.

Remark 6 *Let us take a look to the particular case $\lambda = \infty$ and $w_{k,\nu} = 1$ for all k, ν. Then, for a fixed k, the threshold is equal to $\frac{\|[A^*y]_k\|_1}{F}$. So that, one kept only the coefficients $x_{k,\nu}$ which are bigger than the mean (respect to ν, k being fixed) of $|A^*y|_{k,\nu}$.*

Consequently, the $\ell_{1,2}$ mixed norms cannot give estimates as sparse as the ℓ_1 norm, because of this upper bound of the threshold. This "price" is the consequence of the structures and illustrate well the coupling between significant and unsignificant coefficients as said in remark 5 before.

4.2. *Summary of the main results*

In order to re-use the results proved above and make the reading easier, we rewrite them in a single theorem.

Theorem 3 *Let $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{y} \in \mathbb{C}^N$. Let $A \in \mathcal{M}_N(\mathbb{C})$ be an orthogonal matrix. Let $1 \leq p, q \leq 2$ and $\mathbf{w} \in {\mathbb{R}_+^*}^N$ such that $\|\mathbf{x}\|_{\mathbf{w};p,q}^q = \sum_{k=1}^K \left(\sum_{\nu=1}^F w_{k,\nu}|x_{k,\nu}|^p\right)^{q/p}$. Then the solution of the following optimization problem*

$$\min_{\mathbf{x}\in\mathbb{C}^N} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathbf{w};p,q}^q \ ,$$

is given by $\mathbf{S}_{\mathbf{w},p,q}^{\lambda}(A^*\mathbf{y})$, with $\mathbf{S}_{\mathbf{w},p,q}^{\lambda}$ a soft-thresholding operator defined coordinatewise by, for all k, ν:

$$v_{k,\nu} \mapsto (|v_{k,\nu}| - \xi_{k,\nu})^+$$

where the $\xi_{k,\nu}$ are given here after.

- *If $p > 1$ and $q > 1$. Then the thresholds are given a posteriori, the solution being given by the inverse of $\mathcal{F} : \mathbb{R}^F \to \mathbb{R}^F$, $|\mathbf{v}| \mapsto \mathcal{F}(|\mathbf{v}|) = |\mathbf{v}| + \lambda W_k |\mathbf{v}|^{p-1} \|\mathbf{v}\|_{\mathbf{w};p}^{q-p}$;*
- *If $p = 1$ and $q = 2$. Then, for all ν, $\xi_{k,\nu} = \xi_k$, and*

$$\xi_k = \frac{\lambda}{1 + F_{\mathbf{w}_k}\lambda} \|[A^*\mathbf{y}]_k\|_{\mathbf{w}_k;1} \ ,$$

  with $F_{\mathbf{w}_k} = \sum_{\nu=1}^F w_{k,\nu}$;
- *If $p = 1$ and $1 \leq q < \infty$. Then $\xi_{k,\nu}$ is solution on $\mathbb{R}_+$ of*

$$\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} F_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} \|[A^*\mathbf{y}]_k\|_{\mathbf{w}_k;1} \ ,$$

  with $F_{\mathbf{w}_k} = \sum_{\nu=1}^F w_{k,\nu}$;
- *If $p = 2$ and $q = 1$, and for a fixed k, $w_{k,\nu} = w_k \ \forall \nu$. Then*

$$\xi_{k,\nu} = \frac{\lambda\sqrt{w_k}}{[A^*y]_{k,\nu}\|[A^*\mathbf{y}]_k\|_2} \ ;$$

- *If $1 < p < 2$ and $q = 1$. Then the solution is given by the algorithm 4.*

The preceding theorem shows that the minimizer is obtained by a soft-thresholding operator. This remark gives us the following corollary

Corollary 1 *Let $\mathbf{x} \in \mathbb{C}^N$. Let $A \in \mathcal{M}_N(\mathbb{C})$ be an orthogonal matrix. Let $1 \leq p, q \leq 2$ and $\mathbf{w} \in {\mathbb{R}_+^*}^N$ such that $\|\mathbf{x}\|_{\mathbf{w},p,q}^q = \sum_{k=1}^K \left(\sum_{\nu=1}^F w_{k,\nu}|x_{k,\nu}|^p\right)^{q/p}$.*

Then, for all $\mathbf{y} \in \mathbb{C}^N$, there exist $\xi \in {\mathbb{R}_+^*}^N$ (which depend of $\mathbf{y}$) such that the minimum of the functional

$$\Phi(\mathbf{x}) = \|\mathbf{y} - A\mathbf{x}\|_2^2 + \|\mathbf{x}\|_{\mathbf{w};p,q}^q \ ,$$

coincides with the minimum of the functional

$$\Psi(\mathbf{x}) = \|\mathbf{y} - A\mathbf{x}\|_2^2 + \|\mathbf{x}\|_{\xi;1} \ .$$

We saw how to obtain an estimation of the signal in the case where A is an orthogonal matrix. So that, we defined operators allowing us to obtain the solution. Next section sets the problem in a more general case, and shows that the corresponding iterative algorithm inspired by Daubechies *et al* in [6] and by Teschke in [24] may be extended to this new setting, with the same convergence properties.

4.3. *A thresholded Landweber iteration*

We study here a more general case than the problem (21). A is a general linear operator, and we want to exploit the structure in layers which can appear in a signal. In the particular case where A is the matrix of a dictionary constructed as an union of orthogonal bases, one can apply the Block Coordinate Relaxation

algorithm, as it was shown in [15]. The algorithm studied here is more general, and can be applied to any linear operator A, like matrix corresponding to frame (or union of frame), convolution operator, etc.

Introduce the functional

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \sum_{i=1}^{I} A_i\mathbf{x}^{[i]}\|_2^2 + \sum_{i=1}^{I}\frac{\lambda_i}{q_i}\Psi_i(\mathbf{x}^{[i]}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \boldsymbol{\lambda}\Psi(\mathbf{x}) \ . \tag{28}$$

with $\mathbf{y} \in \mathbb{C}^M$, for all i, $A_i \in \mathbb{C}^{M\times N_i}$ is a linear operator and $A = \bigoplus A_i \in \mathbb{C}^{M\times N}$. Let $\mathbf{x} = (\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[I]}) \in \mathbb{C}^{N\times I}$ and $\boldsymbol{\lambda} = (\frac{\lambda_1}{q_1}, \ldots, \frac{\lambda_I}{q_I}) \in \mathbb{R}^I$. We have $A\mathbf{x} = \sum_i A_i\mathbf{x}^{[i]}$ and $\Psi(\mathbf{x}) = (\Psi_1(\mathbf{x}^{[1]}), \ldots, \Psi_I(\mathbf{x}^{[I]}))$.

We take here for penalty the mixed norm introduced before. So we have $\Psi_i(\mathbf{x}^{[i]}) = \|\mathbf{x}^{[i]}\|_{\mathbf{w}^{[i]};p_i,q_i}$.

To solve the problem (28), following [6,24], we introduce a surrogate functional

$$\Phi^{sur}(\mathbf{x},\mathbf{a}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{C}{2}\|\mathbf{x} - \mathbf{a}\|_2^2 - \frac{1}{2}\|A\mathbf{x} - A\mathbf{a}\| + \boldsymbol{\lambda}\Psi(\mathbf{x}) \tag{29}$$

$$= \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{\nu=1}^{F}\left(C(x_{k,\nu}^{[i]})^2 - 2x_{k,\nu}^{[i]}[A_i^*\mathbf{y} + C\mathbf{a}^{[i]} - A_i^*A\mathbf{a}]_{k,\nu} + \frac{\lambda_i}{q_i}\Psi_i(\mathbf{x}^{[i]})\right)$$

$$+ \frac{1}{2}\left(\|\mathbf{y}\|_2^2 + C\|\mathbf{a}\|_2^2 - \|A\mathbf{a}\|_2^2\right) \tag{30}$$

$$= \frac{1}{2}\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\sum_{\nu=1}^{F}\left(C(x_{k,\nu}^{[i]})^2 - 2x_{k,\nu}^{[i]}[A_i^*\mathbf{y} + C\mathbf{a}^{[i]} - A_i^*A\mathbf{a}]_{k,\nu}\right) + \frac{\lambda_i}{q_i}\sum_{k=1}^{K}\left(\sum_{\nu=1}^{F} w_{k,\nu}^{[i]}|x_{k,\nu}^{[i]}|^{p_i}\right)^{q_i/p_i}\right) \tag{31}$$

$$+ \frac{1}{2}\left(\|\mathbf{y}\|_2^2 + C\|\mathbf{a}\|_2^2 - \|A\mathbf{a}\|_2^2\right) \ ,$$

with C such that $C > C_1 + \ldots + C_I$, where each C_i is the square of the bound norm of the operator A_i.

Then, the solution of the associated variational problem verifies, for all i, k, ν:

$$x_{k,\nu}^{[i]} = C^{-1}[A_i^*\mathbf{y} + C\mathbf{a}^{[i]} - A_i^*A\mathbf{a}]_{k,\nu} - \frac{\lambda_i}{C}\mathrm{sgn}(x_{k,\nu}^{[i]})w_{k,\nu}^{[i]}|x_{k,\nu}^{[i]}|^{p_i-1}\|\mathbf{x}_k^{[i]}\|_{\mathbf{w}_k^{[i]},p_i}^{q_i-p_i} \ . \tag{32}$$

In the usual situation, the introduction of the surrogate functional decouples the problem into scalar problems. In our case, it also performs some decoupling, yielding vector subproblems of smaller dimension, which can be solved as described in section 4.1. Consequently, for all i, the operators $\mathbf{S}_{\mathbf{w}^{[i]},p_i,q_i}^{\lambda_i/C}$ given by theorem 3 in section 4.2, allow us to obtain the solution of (32) from $C^{-1}[A_i^*\mathbf{y} + C\mathbf{a}^{[i]} - A\mathbf{a}]$.

Proposition 5 *Let* $\mathbf{a}$ *be fixed. The surrogate functional* $\Phi^{sur}(\mathbf{x},\mathbf{a})$ *has a minimum given by*

$$\mathrm{argmin}(\Phi^{sur}(\mathbf{x},\mathbf{a})) = \mathbf{S}(C^{-1}[A^*\mathbf{y} + C\mathbf{a} - A^*A\mathbf{a}]) \ ,$$

where $\mathbf{S} = \left(\mathbf{S}_{\mathbf{w}^{[1]},p_1,q_1}^{\lambda_1/C}, \ldots, \mathbf{S}_{\mathbf{w}^{[I]},p_I,q_I}^{\lambda_I/C}\right)$.

PROOF. The surrogate functional allow us to decouple the variational equations, so we just have to apply the work done in section 4.1. ∎

Then, we can deduce the following iterative algorithm

Algorithm 5
Let $\mathbf{x}^{(0)} \in \mathbb{C}^N$
Do
For $i = 1 : I$

$$|\mathbf{x}^{[i]^{(m+1)}}| = \mathbf{S}^{\lambda_i/C}_{\mathbf{w}_i,p_i,q_i}(C^{-1}[A_i^*\mathbf{y} + C\mathbf{x}^{[i]^{(m)}} - A_i^*A\mathbf{x}^{(m)}])$$
$$= \left(C^{-1}[A_i^*\mathbf{y} + C\mathbf{x}^{[i]^{(m)}} - A_i^*A\mathbf{x}^{(m)}] - \xi^{[I]^{(m)}}\right)^+$$

EndFor

until *convergence*

where the $\xi^{[i]^{(m)}}$ are the vectors which contain the thresholds $\xi^{[i]^{(m)}}_{k,\nu}$. These thresholds are given by the operators $\mathbf{S}^{\lambda_i/C}_{\mathbf{w}^{[i]},p_i,q_i}$, associated to the adequate ℓ_{p_i,q_i} mixed norm, applied to $C^{-1}[A_i^*\mathbf{y}+C\mathbf{x}^{[i]^{(m)}}-A_i^*A\mathbf{x}^{(m)}]$. The theorem 3 gives the practical way to obtain the thresholds.

As in [6,24], we show that the preceding algorithm is convergent, then the convergence point is the desired minimum (i.e., a solution for the problem (28)). As we work here in the simpler case of finite dimension, we just have to prove the weak convergence of the algorithm.

Theorem 4 *Suppose the operator A maps a Hilbert space $\mathcal{H}$ to another Hilbert space $\mathcal{H}'$, with $\|A\| < \sqrt{C}$, and with $dim(\mathcal{H}) < \infty$ and $dim(\mathcal{H}') < \infty$. Suppose $\mathbf{y}$ is an element of $\mathcal{H}'$. Then, the sequence of iterates generated by algorithm 5 with $\mathbf{x}^{(0)}$ arbitrarily chosen in $\mathbb{C}^N$, converges to a fixed point which is a minimizer of the functional (29).*

Moreover, the obtained fixed point is a minimizer of the functional (28)

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \sum_i A_i\mathbf{x}^{[i]}\|_2^2 + \sum_i \lambda_i\Psi_i(\mathbf{x}^{[i]}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \boldsymbol{\lambda}\Psi(\mathbf{x}) \ .$$

PROOF. The proof follows the lines originally given by Daubechies *et al* in [6], and resumed in various papers as [24] to prove the (weak) convergence of a thresholded Landweber iteration. We choose then to postpone the proof to appendix A.4. Although we work in finite dimension, the given proof does not use this hypothesis. So that, the weak convergence in infinite dimension can be proven too. ∎

Remark 7 *We limited ourselves to the finite dimensional case, consequently the weak convergence is enough. If one wan to study the convergence in norm, only the lemma 3.18 of [6] must be proven.*

We have now several algorithms to solve some specific regression problems. Next section gives some illustrations of these algorithms and the influence of the mixed norms.

5. Two illustrations

We choose to limit ourselves to two illustrations for our algorithms, in the audio signal processing domain. The first one is an application to signal declicking and illustrates the thresholded Landweber algorithm. The FOCUSS algorithm is illustrated by a decomposition of an audio signal in "transients + tonal" layers, following [8,7]. Applications to similar image processing problems may also be done in a straightforward way.

5.1. *Illustration of the thresholded Landweber algorithm*

Our declicking example is a "toy example" which allows us to show the different behaviors and the influence of the mixed norms compared to the classical ℓ_1 norm. In this example we limit ourselves to $\ell_{1,2}$ norm. This choice is justified in Remark 8 below.

We choose a 3 sec. long (2^{17} samples) trumpet signal. We added to this signal some random clicks simulated by Dirac pulses with amplitudes ± 1 The resulting signal is the $\mathbf{y}$ signal of the theory and its Signal to Noise

Ratio (SNR) is equal to 20.33 dB. The time representation of the samples of the original signal and its clicked version is provided in figure 1.

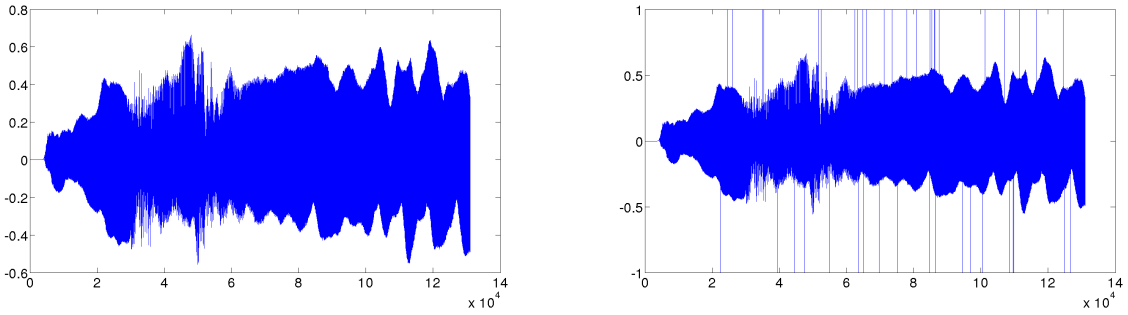

Fig. 1. Trumpet signal (left) and its clicked version (right).

The signal is then decomposed in a Gabor frame with a 2048 samples length Gaussian window, with a time shifting of 128 samples, and 2 samples in frequency. The corresponding matrix of the Gabor frame operator is the matrix A. As can be seen on the time-frequency representation in figure 2, the clicks appears clearly like vertical lines that are sparse in time, but cover all the frequencies.

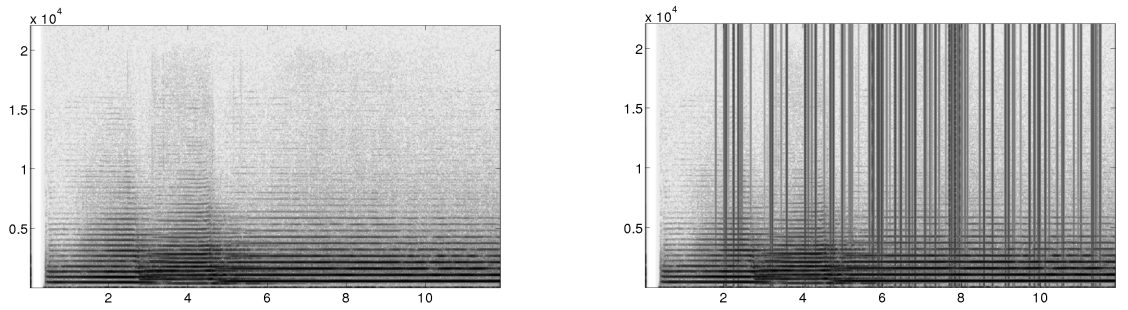

Fig. 2. Time-frequency representation of the Gabor coefficients for the original trumpet signal (left) and its clicked version (right).

Several strategies can be imagined to declick the signal. First, we used the thresholded Landweber algorithm with a ℓ_1 norms penalty and compared it with the same algorithm using a mixed norms penalty.

The functional Φ that one wants to minimize is the following:

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\left(\sum_{k=1}^{K}\left(\sum_{\nu=1}^{F}|x_{k,\nu}|^p\right)^{q/p}\right) \quad , \tag{33}$$

where $\mathbf{y}$ is the clicked signal and A the matrix corresponding to the operator of the Gabor frame. The p, q are chosen as follows

- $p = q = 1$. This correspond to the classical ℓ_1 norm.
- $p = 1$ and $q = 2$. The k index correspond to the time, and the ν index correspond to the frequency. This choice is made to promote sparsity in frequency.

Figure 3 provides the SNR as a function of the number of retained coefficients. The mixed norm obviously outperforms the classical ℓ_1 norm. To clearly illustrate the behavior of the mixed norm, figure 4 shows the time-frequency representation of the Gabor coefficients, for a comparable number of retained coefficients of ℓ_1 and $\ell_{1,2}$ norms. It clearly appears on the time-frequency representation that the clicks are better eliminated with the $\ell_{1,2}$ mixed norm than with the classical ℓ_1 norm: the ℓ_1 norm keeps more vertical lines which correspond to clicks. Moreover, the partials are better preserved by the mixed norm.

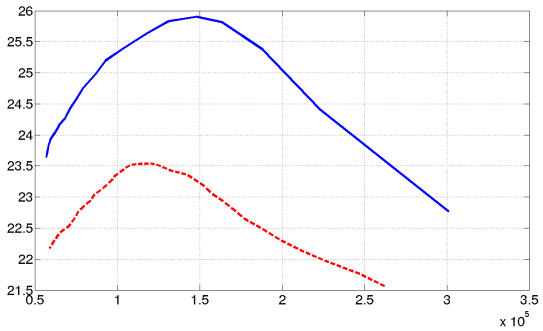

Fig. 3. Evolution of the SNR in function of the number of coefficients. solid line: ℓ_{12} mixed norm, dashed line: ℓ_1 norm.

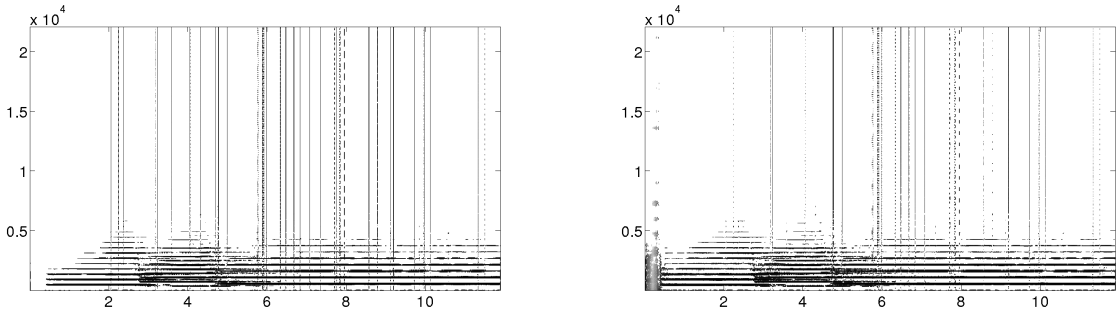

Fig. 4. Time-frequency representation of the Gabor coefficients for the denoised version. Left: ℓ_1 estimate. Right: $\ell_{1,2}$ estimate.

.

Remark 8 *The $\ell_{1,2}$ mixed norm appeared specially adapted for the problem we choose if we look back to the estimate given in the orthogonal case at proposition 4 in subsection 4.1.3. For a fixed time index k, the threshold is equal to $\frac{\lambda}{1+\lambda F}\|[A^*\mathbf{y}]_k\|_1$. Thus, when a click appears at the time index k, one expects that the threshold is higher than at a time index without a click.*

We did not use the $\ell_{2,1}$ mixed norm because this norm keeps entire groups (the sparsity are on the groups, not on the coefficients). This structure in "lines" does not seem to be very adapted to estimate the trumpet signal: the partials can evolve slowly in time, and their number may jump from a time frame to another.

However, this structure in lines could be adapted to estimate only the clicks, and then obtain the clean signal in the residual of the functional. This strategy corresponds to an another functional instead of (33) and we did not try this strategy here.

5.2. *Illustration of the FOCUSS algorithm*

To illustrate the modified FOCUSS algorithm, we choose a xylophone signal of 0.7 sec long (2^{15} samples) represented in Figure 5. The goal is to provide a decomposition in two layers "transient + tonal" subject to an exact reconstruction. For this, we choose to expand the signal in a dictionary constructed as the union of two MDCT bases. The first MDCT basis is chosen with a 4096 samples long window (about 90 msec) and is adapted for the tonal layer. The second one is chosen with a 128 samples long window and is adapted for the transient layer. The MDCT coefficients of transient layer are represented in Figure 5.

The classical strategy is to minimize the ℓ_1 norm of all the coefficients. Each layer is then obtained by the inverse transform of the corresponding MDCT coefficients. This minimization is done by the classical FOCUSS algorithm.

Our adaptation of FOCUSS is used with two mixed norms. For the tonal layer the ℓ_{p_1,q_1} mixed norm are choose to promote sparsity in frequency with $p_1 = 1.2$ and $q_1 = 2$ (and then, the k index correspond to

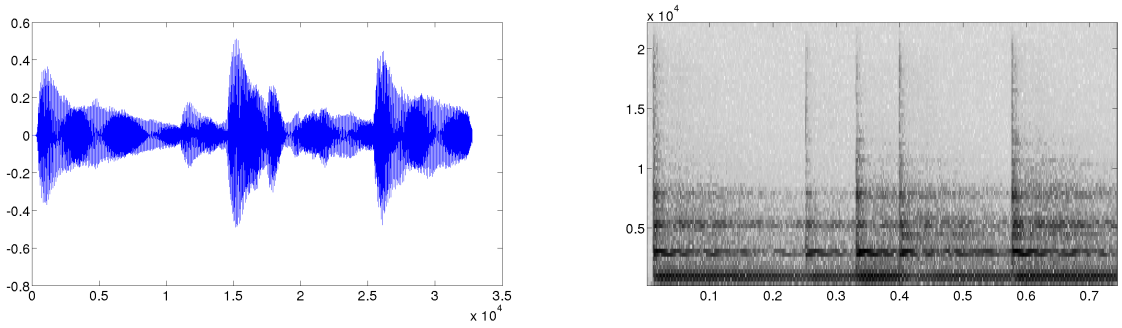

Fig. 5. Left: the xylophone signal. Right: MDCT coefficients of the xylophone, with a window of 128 samples length.

the time and the ν index to the frequency). For the transient layer, with the k index corresponding to the frequency and the ν index to the time, we choose a ℓ_{p_2,q_2} mixed norm with $p_2 = 1$ and $q_2 = 1/2$. This last choice was made to obtain a very sparse layer, but with a "structured sparsity". In order to balance between the penalty between the two mixed norms, we choose $\lambda_1 = 1$ and $\lambda_2 = 5$.

We provide in Figure 6 the MDCT coefficients of the transient layer estimate by the ℓ_1 norm and the mixed norms. One can see that the estimate obtained by the mixed norm is sparser than the ℓ_1 estimate, and one can observe that the chosen mixed norm promote some structures compared to the ℓ_1 norm.

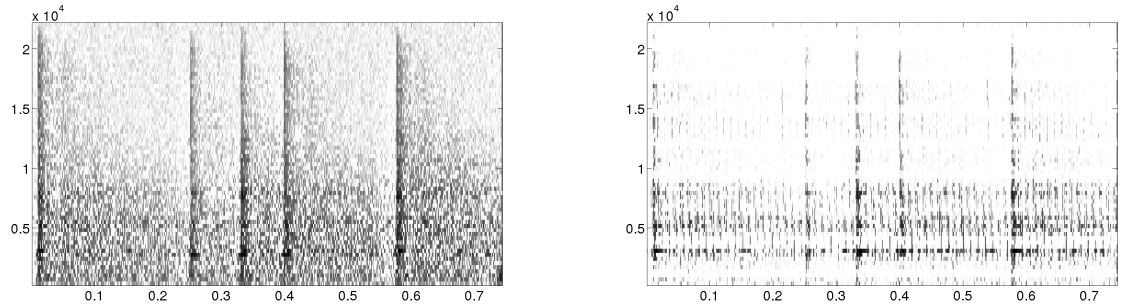

Fig. 6. MDCT coefficients of two estimates of the transient layer for the xylophone.

6. Conclusion and perspectives

This paper showed that when the data can be labelled by a double index, the mixed norms can help to introduce easily structures. The labels can indeed be used to introduced a hierarchy throught the coefficients. This hierarchy is then explicitely used in the model and modelized by mixed norms $\ell_{p,q}$. Then, one can play both on p and q to promote a structured sparsity.

These norms are adapted to two differents situations of regression:

– Signal estimation subject to an equality constraint ;

– Noisy signal estimation.

Two algorithms were presented, corresponding to these different problems. For the sake of clarity, we summarize their strengths and weaknesses in the table 1.

Let us notice that in the particular case of regression in an union of orthogonal bases, the Block Coordinate Relaxation (BCR) algorithm with mixed norms [15] provide a valuable alternative to the thresholded Landweber iteration presented here. Our numerical experiments (not provided here) seems to show that both algorithms perform quite similar.

The behavior of the iterative thresholding algorithms was illustrated on a specific example in order to stress the influence of mixed norms compared to the classical ℓ_1 norm. Audio signal was chosen for the intuitive structures provided by the time-frequency representations. But let us stress that mixed norm are

	FOCUSS	Iterative thresholding
Range value for p and q	$p, q \leq 2$ and $p, q \neq 0$	$1 \leq p, q \leq 2$, and for any $p \leq 1$ if $q = 2$. The cases $q = 1$ and $1 \leq p < 2$ use an iterative algorithm which works for suitably chosen λ.
Speed	–	+
Ease of implementation	+	+
Optimization subject to equality constraint	+	Not designed for
Optimization subject to inequality constraint	+	+

Table 1
Compared advantages and shortcomings of the algorithms.

certainly not specific of audio signal and can be used on any applications with structures given by a suitable double indexing of the coefficients of the signal's expansion.

Already, the $\ell_{p,1}$-like norms already enjoyed significant succes in the statistical community for variable selection [27,23,28], and were more specifically studied and applied for color image restoration in [11] and [25].

Our work studied the mixed norm in a general manner, and we want to stress the utility of the $\ell_{1,q}$-like norms, to encourage some structures, without imposing sparsity only on grouped variables (see remark 5 in section 4.1.3, and remark 8 in section 5). The simple example provided here encourages us to use mixed norm in signal restoration. The preceding paper [15] gave encouraging results in multichannel denoising and multilayered "tonal + transients + noise" decomposition. The author is currently studying source separation of under-determined anechoic mixture with the help of mixed norms. Promising results will be presented in a forthcoming paper.

Some natural extensions can also be studied, as mixed norms with more than a two levels index, or using the sum of mixed norms (as in elastic-net [29] which use a sum of ℓ_1 and ℓ_2 penalty, or the regularization penalty proposed in [11,12]) for the regularization term in a functional. Furthermore, it could be interesting to use the mixed norm for the data term in the transformed domain. Indeed, the ℓ_2 norm is used in a bayesian context to model Gaussian noise. The use of such norms could be adapted to penalyse noise which are known to be not Gaussian.

Acknowledgement

The author wishes to gratefully thanks Bruno Torrésani for his advices and help during this work.

Appendix A. Proofs

A.1. *Proof of proposition 2*

The variational equations for problem (21) are as follows. For all k, ν, s.t. $x_{k,\nu} \neq 0$ we have:

$$|x_{k,\nu}| = \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda w_k |x_{k,\nu}| \|\mathbf{x}_k\|_{\mathbf{w}_k,2}^{-1}\right)^+ .$$

So we can give an expression for λ:

$$\lambda = \frac{\left(|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|\right)\right)^+}{w_k |x_{k,\nu}| \|\mathbf{x}_k\|_{\mathbf{w}_k,2}^{-1}} .$$

We can now eliminate λ from the expression of $|x_{k,\nu}|$:

$$
\begin{aligned}
|x_{k,\nu}| &= \left(|[A^*\mathbf{y}]_{k,\nu}| - \frac{|[A^*\mathbf{y}]_{k,l}| - |x_{k,l}|}{|x_{k,l}|} |x_{k,\nu}| \right)^+ \\
|x_{k,\nu}| \left(1 + \frac{(|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,l}|)^+}{|x_{k,l}|} \right) &= |[A^*\mathbf{y}]_{k,\nu}| \\
|x_{k,\nu}| &= \frac{|[A^*\mathbf{y}]_{k,\nu}|}{1 + \frac{(|[A^*\mathbf{y}]_{k,l}| - |x_{k,l}|)^+}{|x_{k,l}|}} = \frac{|[A^*\mathbf{y}]_{k,\nu}|}{|[A^*\mathbf{y}]_{k,l}|} |x_{k,l}| \ \forall l \text{ and } |[A^*y]_{k,l}| \neq 0 \ .
\end{aligned}
$$

So that, we obtain

$$
\begin{aligned}
|x_{k,\nu}| &= \left(|[A^*\mathbf{y}]_{k,\nu}| - \frac{\lambda |x_{k,\nu}| w_k}{\sqrt{\sum_l w_k |x_{k,l}|^2}} \right)^+ \\
|x_{k,\nu}| &= \left(|[A^*\mathbf{y}]_{k,\nu}| - \frac{\lambda |x_{k,\nu}| w_k}{\sqrt{\sum_l \left(w_k |[A^*\mathbf{y}]_{k,l}|^2 \frac{|x_{k,\nu}|^2}{|[A^*\mathbf{y}]_{k,\nu}|)^2} \right)}} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \frac{\lambda |x_{k,\nu}| w_k}{\frac{|x_{k,\nu}|}{|[A^*\mathbf{y}]_{k,\nu}|} \sqrt{w_k} \|[A^*\mathbf{y}]_k\|_2} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \frac{\lambda \sqrt{w_k} (|[A^*\mathbf{y}]_{k,\nu}|)}{\|\mathbf{y}_k\|_2} \right)^+ \\
&= |[A^*\mathbf{y}]_{k,\nu}| \left(1 - \frac{\sqrt{w_k}\lambda}{\|[A^*\mathbf{y}]_k\|_2} \right)^+ ,
\end{aligned}
$$

which is the desired result.

A.2. *Proof of proposition 3 (convergence of the fixed point algorithm)*

We prove that for all k, the sequence of vectors $\mathbf{x}_k$ converges to an unique fixed vector. We denote by s the soft-thresholding operator which maps $|\mathbf{x}_k^{(m)}|$ to

$$
\begin{aligned}
|\mathbf{x}_k^{(m+1)}| &= s(|\mathbf{x}_k^{(m)}|) && \text{(A.1)} \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda w_{k,\nu} |x_{k,\nu}^{(m)}|^{p-1} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{1-p} \right)^+ . && \text{(A.2)}
\end{aligned}
$$

To prove the proposition, we simply apply the Picard's fixed point theorem to s.

We have

$$\frac{\partial s(|\mathbf{x}_k|)}{\partial |x_{k,\nu}|} = \begin{pmatrix} -\lambda w_{k,1}|x_{k,1}|^{p-1} w_{k,\nu}|x_{k,\nu}|^{p-1}(1-p)\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} \\ \vdots \\ -\lambda(1-p)w_{k,\nu}^2|x_{k,\nu}|^{2p-2}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} - \lambda(p-1)w_{k,\nu}|x_{k,1}|^{p-2}\|x_{k,\nu}\|_{\mathbf{w}_k;p}^{1-p} \\ \vdots \end{pmatrix}$$

$$= -\lambda w_{k,\nu}|x_{k,\nu}|^{p-1}(1-p)\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} \begin{pmatrix} w_{k,1}|x_{k,1}|^{p-1} \\ \vdots \\ w_{k,\nu}|x_{k,\nu}|^{p-1} \\ \vdots \end{pmatrix} - \lambda(p-1)w_{k,\nu}|x_{k,\nu}|^{p-2}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-p} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{pmatrix} .$$

Since we want to give an upper bound for the ℓ_1 norm of s, we use the general mean inequality: let the following quantities

$$M_p = \left(\frac{1}{\sum_{n=1}^N w_n} \sum_{n=1}^N w_n |x_n|^p \right)^{\frac{1}{p}} .$$

If $\alpha < \beta$, then $M_\alpha < M_\beta$ for all α and β in $\mathbb{R}^*$.

Denoting by $F_{\mathbf{w}_k} = \sum_{\nu=1}^F w_{k,\nu}$, and as we have $1 < p < 2$, we can give an upper bound for the ℓ_1 norm:

$$\begin{aligned}
\left\| \frac{\partial s(|\mathbf{x}_k|)}{\partial |x_{k,\nu}|} \right\|_1 \leq & \lambda(p-1) \left(\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} w_{k,\nu}|x_{k,\nu}|^{p-1} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-1}^{p-1} + w_{k,\nu}|x_{k,\nu}|^{p-2}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-p} \right) \\
\leq & \lambda(p-1) \left(\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-1}^{p-1} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-1}^{p-1} + \|\mathbf{x}_k\|_{\mathbf{w}_k;p-2}^{p-2} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-p} \right) \\
\leq & \lambda(p-1) \left(F_{\mathbf{w}_k}^{\frac{2}{p}} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{1-2p} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{p-1} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{p-1} + F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-2}^{p-2} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-2}^{1-p} \right) \text{ (means inequality)} \\
\leq & \lambda(p-1) \left(F_{\mathbf{w}_k}^{\frac{2}{p}} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{-1} + F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \|\mathbf{x}_k\|_{\mathbf{w}_k;p-2}^{-1} \right) \leq \lambda(p-1)\|\mathbf{x}_k\|_{\mathbf{w}_k;p-2}^{-1} \left(F_{\mathbf{w}_k}^{\frac{2}{p}+\frac{2}{p(p-2)}} + F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \right) \\
\leq & \lambda(p-1)\|\mathbf{y}_k\|_{\mathbf{w}_k;p-2}^{-1} \left(2F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \right) \\
\leq & \lambda(p-1) \max_k(\|\mathbf{y}_k\|_{\mathbf{w}_k;p-2}^{-1}) \left(2F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \right) \\
\leq & \lambda(p-1) \frac{1}{\min_k(\|\mathbf{y}_k\|_{\mathbf{w}_k;p-2})} \left(2F_{\mathbf{w}_k}^{\frac{2(p-1)}{p(p-2)}} \right) .
\end{aligned}$$

If one chooses λ small enough, one can make this quantity strictly smaller than 1. So that, the application s is contractive and assure the convergence of the algorithm.

A.3. *Proof of proposition 4*

For all k, ν, we have

$$|x_{k,\nu}| = \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \, \|W_k\mathbf{x}_k\|_1^{q-1} \right)^+ ,$$

so we can write λ:

$$\lambda = \frac{(|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|)^+}{\|W_k\mathbf{x}_k\|_1^{q-1}} ,$$

and we can give the expression of any $|x_{k,\nu}|$ in terms of any $|x_{k,l}|$:

$$|x_{k,\nu}| = (|[A^*\mathbf{y}]_{k,\nu}| - (|[A^*\mathbf{y}]_{k,l}| - |x_{k,l}|))^+ \ .$$

So that, we have

$$\begin{aligned}
|x_{k,\nu}| &= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ \left(\sum_{l=1}^{F} w_{k,l}|x_{k,l}| \right)^{q-1} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ \left(\sum_{l} w_{k,l}(|[A^*\mathbf{y}]_{k,l}| - (|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|) \right)^{q-1} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ \left(\sum_{l=1}^{F} w_{k,l} \left(|[A^*\mathbf{y}]_{k,l}| + (|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|)\right) \right)^{q-1} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ \left(\sum_{l=1}^{F} w_{k,l}|[A^*\mathbf{y}]_{k,l}| + F_{\mathbf{w}_k}(|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|) \right)^{q-1} \right)^+ \\
&= \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ (\|W_k[A^*\mathbf{y}]_k\|_1 + F_{\mathbf{w}_k}(|[A^*\mathbf{y}]_{k,\nu}| - |x_{k,\nu}|))^{q-1} \right)^+ \ , \qquad \text{(A.3)}
\end{aligned}$$

with $F_{\mathbf{w}_k} = \sum_{l=1}^{F} w_{k,l}$.

If $q = 2$, we have the analytical expression

$$|x_{k,\nu}| = \left(|[A^*\mathbf{y}]_{k,\nu}| - \lambda \ \frac{\lambda}{1 + \lambda F_{\mathbf{w}_k}} \|W_k[A^*\mathbf{y}]_k\|_1 \right)^+ \ .$$

If $1 \leq q < 2$, equation (A.3) can be solved numerically. We denote $\xi_k = [A^*\mathbf{y}]_{k,\nu} - x_{k,\nu}$. We just have to solve in $\mathbb{R}_+$:

$$\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} F_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} \|[A^*\mathbf{y}]_k\|_1 \ .$$

One can easily verify that this equation has a unique solution on $\mathbb{R}_+$.

Then we have

$$|x_{k,\nu}| = (|[A^*\mathbf{y}]_{k,\nu}| - \xi_k)^+ \ .$$

In the case $q = 2$, $\xi_k = \frac{\lambda}{1 + F_{\mathbf{w}_k}\lambda} \|[A^*\mathbf{y}]_k\|_1$.

A.4. *Proof of theorem 4*

In order to prove the convergence of the sequence of $\mathbf{x}^{(m)}$ obtained with the algorithm 5, we apply Opial's theorem, which we recall here for convenience:

Theorem 5 *Let the mapping A from $\mathcal{H}$ to $\mathcal{H}$, satisfy the following conditions:*

i A is non-expansive: $\forall u, v \in \mathcal{H}$, $\|Au - Av\| \leq \|u - v\|$,

ii A is asymptotically regular: $\forall v \in \mathcal{H}$, $\|A^{n+1}v - A^n v\| \to 0$,

iii The set $\mathcal{E}$ of fixed points of A in $\mathcal{H}$ is not empty.

Then, $\forall v \in \mathcal{H}$, the sequence $(A^n v)_{n\in\mathbb{N}}$ converges weakly to a fixed point in $\mathcal{E}$.

Let the mapping $\mathbf{T}$ be such that:

$$\mathbf{Tx} = \mathbf{S}(C^{-1}[A^*\mathbf{y} + C\mathbf{x} - A^*A\mathbf{x}]) \ .$$

In order to apply Opial's theorem to T, we just have to check the three hypotheses. For that, we prove a series of lemma.

Lemma 1 *The operator* $\mathbf{S}$ *is non-expansive, i.e. for all* $u, v \in \mathcal{H}$,

$$\|\mathbf{S}u - \mathbf{S}v\| \leq \|u - v\| \ .$$

PROOF. The non-expansiveness comes from the fact that $\mathbf{S}$ is a soft-thresholding operator. ∎

Lemma 2 *The mapping* $\mathbf{T}$ *is non-expansive.*

PROOF. This follows from the fact that the operator $\mathbf{S}$ is non-expansive. We have then

$$\begin{aligned}\|\mathbf{T}u - \mathbf{T}v\| &\leq \|(I - C^{-1}A^*A)(u - v)\| \\ &\leq \|I - C^{-1}A^*A\|\|u - v\| \\ &\leq \|u - v\| \text{ because } \|A\|^2 < C \ .\end{aligned}$$

∎

The first hypothesis in Opial's theorem is now verified. Let us now verify that the second hypothesis is satisfied.

Lemma 3 *Both* $\left(\Phi(\mathbf{x}^{(m)})\right)_{m\in\mathbb{N}}$ *and* $\left(\Phi^{sur}(\mathbf{x}^{(m+1)}, \mathbf{x}^{(m)})\right)_{m\in\mathbb{N}}$ *are non-increasing sequences.*

PROOF. Let us introduce the operator $L = \sqrt{CI - A^*A}$, so that $C\|\mathbf{h}\|^2 - \|A\mathbf{h}\|^2 = \|L\mathbf{h}\|^2$. As $\mathbf{x}^{(m+1)}$ is the minimizer of the functional $\Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)})$,

$$\Phi(\mathbf{x}^{(m+1)}) + \|L(\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)})\|^2 = \Phi^{sur}(\mathbf{x}^{(m+1)}, \mathbf{x}^{(m)}) \leq \Phi^{sur}(\mathbf{x}^{(m)}, \mathbf{x}^{(m)}) = \Phi(\mathbf{x}^{(m)}) \ .$$

Moreover,

$$\Phi^{sur}(\mathbf{x}^{(m+2)}, \mathbf{x}^{(m+1)}) \leq \Phi(\mathbf{x}^{(m+1)}) \leq \Phi(\mathbf{x}^{(m+1)}) + \|L(\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)})\|^2 = \Phi^{sur}(\mathbf{x}^{(m+1)}, \mathbf{x}^{(m)}) \ .$$

∎

Lemma 4 *The series* $\sum_{m=0}^{\infty} \|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\|^2$ *is convergent.*

PROOF. Since L is a strictly positive operator, we have

$$\sum_{m=0}^{M} \|\mathbf{x}^{m+1} - \mathbf{x}^{m}\|^2 \leq \frac{1}{\mu} \sum_{m=0}^{M} \|L(\mathbf{x}^{m+1} - \mathbf{x}^{m})\|^2 \ ,$$

where μ is a strictly positive lower bound for the spectrum of L^*L. By lemma 3,

$$\sum_{m=0}^{M} \|L(\mathbf{x}^{m+1} - \mathbf{x}^{m})\|^2 \leq \sum_{m=0}^{M} [\Phi(\mathbf{x}^{(m)}) - \Phi(\mathbf{x}^{(m+1)})] = \Phi(\mathbf{x}^{(0)}) - \Phi(\mathbf{x}^{(M+1)}) \leq \Phi(\mathbf{x}^{(0)} \ .$$

It follows that the series $\sum_{m=0}^{\infty} \|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\|^2$ is convergent. ∎

Consequently, we have

Lemma 5 *The operator* T *is asymptotically regular, i.e.*

$$\|T^{(m+1)}\mathbf{x}^{(0)} - T^{(m)}\mathbf{x}^{(0)}\| = \|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\| \to 0 \ .$$

The hypothesis (ii) of Opial's theorem is verified. We still have to check the last hypothesis.

Lemma 6 *We can uniformly bound below the sequence formed by* $\mathbf{w}$ *by a strictly positive real number. Then the* $\|\mathbf{x}^{(m)}\|$ *are uniformly bounded in* m.

PROOF. Thanks to finite dimension we can uniformly bound bellow the sequence formed by $\mathbf{w}$ by a strictly positive number (in infinite dimension, this hypothesis must be made). So we have $w_{k,\nu} \geq c$, uniformly in (k,ν), with $c > 0$.

We can write

$$\Psi(\mathbf{x}^{(m)}) \leq \Phi(\mathbf{x}^{(m)}) \leq \Phi(\mathbf{x}^{(0)}) \ ,$$

thanks to lemma 3. The $\Psi(\mathbf{x}^{(m)})$ are then uniformly bounded.

So, for all i

$$\|\mathbf{x}^{[i]^{(m)}}\|^{q_i}_{\mathbf{w},p_i,q_i} \leq \Phi(\mathbf{x}^{(0)}) \ , \tag{A.4}$$

and then, for all k,

$$\|\mathbf{x}^{[i]^{(m)}}_k\|^{q_i}_{\mathbf{w},p_i} \leq \Phi(\mathbf{x}^{(0)}) \ . \tag{A.5}$$

For all i, we can bound $\|\mathbf{x}^{[i]^{(m)}}\|_2^2$

$$\begin{aligned}
\|\mathbf{x}^{[i]^{(m)}}\|_2^2 \leq & \sum_k \left(\sum_\nu |x^{[i]^{(m)}}_{k,\nu}|^2 \right)^{2-q_i/p_i} \left(\sum_\nu |x^{[i]^{(m)}}_{k,\nu}|^2 \right)^{q_i/p_i} \\
\leq & \max_k \left(\left(\|\mathbf{x}^{[i]^{(m)}}_k\|_2^2 \right)^{2-q_i/p_i} \right) \sum_k \left(\sum_\nu |x^{[i]^{(m)}}_{k,\nu}|^2 \right)^{q_i/p_i} \\
\leq & c^{-2q_i/p_i^2} \max_k \left(\left(\|\mathbf{x}^{[i]^{(m)}}_k\|_2^2 \right)^{2-q_i/p_i} \right) \sum_k \left(\sum_\nu w^{(2-p_i)/p_i}_{k,\nu} |x^{[i]^{(m)}}_{k,\nu}|^{2-p_i} w_{k,\nu} |x^{[i]^{(m)}}_{k,\nu}|^{p_i} \right)^{q_i/p_i} \\
\leq & c^{-2q_i/p_i^2} \max_k \left(\left(\|\mathbf{x}^{[i]^{(m)}}_k\|_2^2 \right)^{2-q_i/p_i} \right) \max_k (\|\mathbf{x}^{[i]^{(m)}}_k\|^{2-p_i}_{\mathbf{w},p_i}) \|\mathbf{x}^{[i]^{(m)}}\|^{q_i}_{\mathbf{w},p_i,q_i} \ .
\end{aligned}$$

Furthermore, we can show that

$$\|\mathbf{x}^{[i]^{(m)}}_k\|_2^2 \leq c^{-2/p_i} \max(w^{(2-p_i)/p_i} |x^{[i]^{(m)}}_{k,\nu}|^{2-p_i}) \|\mathbf{x}^{[i]^{(m)}}_k\|^{p_i}_{\mathbf{w},p_i} \leq c^{-2/p_i} \|\mathbf{x}^{[i]^{(m)}}_k\|^{2-p_i}_{\mathbf{w},p_i} \|\mathbf{x}^{[i]^{(m)}}_k\|^{p_i}_{\mathbf{w},p_i} = c^{-2/p_i} \|\mathbf{x}^{[i]^{(m)}}_k\|^2_{\mathbf{w},p_i} \ ,$$

and then, as we have $2 - q_i/p_i \geq 0$

$$\|\mathbf{x}^{[i]^{(m)}}\|^2 \leq c^{-4/p_i} \max_k (\|\mathbf{x}^{[i]^{(m)}}_k\|^{2(2-q_i/p_i)}_{\mathbf{w},p_i}) \max_k (\|\mathbf{x}^{[i]^{(m)}}_k\|^{2-p_i}_{\mathbf{w},p_i}) \|\mathbf{x}^{[i]^{(m)}}\|^{q_i}_{\mathbf{w},p_i,q_i} \ ,$$

which, with equations (A.4) and (A.5), allow us to give an uniform upper bound for $\|\mathbf{x}^{(m)}\|^2$.

Obviously, the finite dimension allows one to conclude so must faster, thanks to the equivalence of the norms. But we choose to provide a proof which can be applied in infinite dimension. ∎

As pointed out in [6], we have

Lemma 7 *Suppose the mapping T satisfies the conditions (i) and (ii) in Opial's theorem. Then if a subsequence of $(T^n v)_{n\in\mathbb{N}}$ converges weakly, then its limit is a fixed point of T.*

Lemma 8 *The sequence $\mathbf{x}^{(m)} = T^m \mathbf{x}^{(0)}, m = 1, 2, \ldots$ converges weakly, and its limit is a fixed point of T.*

PROOF. Lemma 6 and Banach-Alaoglu's theorem prove that T has a weak accumulation point. Lemma 7 proves that this accumulation point is a fixed point of T. So that, the set of fixed point of T is non empty. ∎

So that, all hypothesis in Opial's theorem are verified, and we can conclude that the sequence of $\mathbf{x}^{(m)}$ is converging to a minimizer of (29).

We want now to prove that the obtained fixed point is a minimizer of the functional (28).

We denote by $\mathbf{x}^\star$ a fixed point of the algorithm 5. We have $\Phi^{sur}(\mathbf{x}^\star + \mathbf{h}; \mathbf{x}^\star) = \Phi(\mathbf{x}^\star + \mathbf{h}) + C\|\mathbf{h}\|_2^2 - \|A\mathbf{h}\|_2^2$.

We first prove that, if $\mathbf{x}$ is a critical point of $\Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)})$, then

$$\Phi^{sur}(\mathbf{x}+\mathbf{h}, \mathbf{x}^{(m)}) - \Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)}) \geq C\|\mathbf{h}\|_2^2 \ .$$

For this, we calculate $\partial\Phi^{sur}(\mathbf{x}, \mathbf{a})$:

$$\partial\Phi^{sur}(\mathbf{x}, \mathbf{a}) = -A^*(y - A\mathbf{a}) + 2C(\mathbf{x}-\mathbf{a}) - 2A^*(A\mathbf{x}-\mathbf{a}) + \lambda\partial\Psi(\mathbf{x}) \ ,$$

so that, one can check that

$$\Phi^{sur}(\mathbf{x}+\mathbf{h}, \mathbf{x}^{(m)}) - \Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)}) = \partial\Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)})\mathbf{h} + C\|\mathbf{h}\|_2^2 + \lambda\{\Psi(\mathbf{x}+\mathbf{h}) - \Psi(\mathbf{x}) - \partial\Psi(\mathbf{x})\mathbf{h}\} \ .$$

As $\mathbf{x}$ is a critical point, i.e. for all $\mathbf{v}$ in $\partial\Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)})$ and for all $\mathbf{h}$, we have $\partial\Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)})\mathbf{h} = 0$, then

$$\Phi^{sur}(\mathbf{x}+\mathbf{h}, \mathbf{x}^{(m)}) - \Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)}) = C\|\mathbf{h}\|_2^2 + 2\lambda\{\Psi(\mathbf{x}+\mathbf{h}) - \Psi(\mathbf{x}) - \partial\Psi(\mathbf{x})\mathbf{h}\} \ .$$

By definition of the sub-gradient, an element $\mathbf{v}$ belong to $\partial\Psi(\mathbf{x})$if and only if for all $\mathbf{y}$ $\Psi(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y}-\mathbf{x}\rangle \leq \Psi(\mathbf{y})$. In particular, for $\mathbf{y} = \mathbf{x}+\mathbf{h}$, this give for all $\mathbf{h}$ and for all $\mathbf{v} \in \partial\Psi(\mathbf{x})$

$$\Psi(\mathbf{x}) + \langle \mathbf{v}, \mathbf{h}\rangle \leq \Psi(\mathbf{y}) \text{ i.e. } 0 \leq \Psi(\mathbf{x}+\mathbf{h}) - \Psi(\mathbf{x}) - \partial\Psi(\mathbf{x})\mathbf{h} \ .$$

Finally

$$\Phi^{sur}(\mathbf{x}+\mathbf{h}, \mathbf{x}^{(m)}) - \Phi^{sur}(\mathbf{x}, \mathbf{x}^{(m)}) \geq C\|\mathbf{h}\|_2^2 \ .$$

As $\Phi^{sur}(\mathbf{x}^\star, \mathbf{x}^\star) = \Phi(\mathbf{x}^\star)$ and $\Phi^{sur}(\mathbf{x}^\star+\mathbf{h}, \mathbf{x}^\star) = \Phi(\mathbf{x}^\star+\mathbf{h}) + C\|h\|_2^2 - \|A\mathbf{h}\|_2^2$, we can conclude that for all $\mathbf{h}$:

$$\Phi(\mathbf{x}^\star+\mathbf{h}) \geq \Phi(\mathbf{x}^\star) + \|A\mathbf{h}\|_2^2 \ ,$$

which conclude the proof.

References

[1] M. Bazaraa, C. Shetty, Nonlinear Programming: Theory and Algorithms, wiley ed., New York, 1979.

[2] A. Benedek, R. Panzone, The space l^p with mixed norm, Duke Mathematical Journal 28 (1961) 301–324.

[3] J. Berger, R. Coifman, M. Goldberg, Removing noise from music using local trigonometric bases and wavelet packets, J. Audio Eng. Soc. 42 (10) (1994) 808–818.

[4] S. S. Chen, D. L. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1) (1998) 33–61.

[5] A. Cohen, I. Daubechies, O. G. Guleryuz, M. T. Orchard, On the importance of combining wavelet-based nonlinear approximation with coding strategies, IEEE Trans. Inform. Theory 48 (7) (2002) 1895–1921.

[6] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on Pure and Applied Mathematics 57 (11) (2004) 1413 – 1457.

[7] L. Daudet, S. Molla, B. Torrésani, Towards a hybrid audio coder, in: J. P. Li (ed.), International Conference Wavelet analysis and Applications, Chongqing, China, 2004, pp. 13–24.

[8] L. Daudet, B. Torrésani, Hybrid representations for audiophonic signal encoding, Signal Processing 82 (11) (2002) 1595–1617, special issue on Image and Video Coding Beyond Standards.
URL `http/::www.cmi.univ-mrs.fr/~torresan/papers/SigPro.ps.gz`

[9] H. G. Feichtinger, Modulation spaces: Looking back and ahead, Sampling Theory in Signal and Image Processing 5 (3) (2006) 109–140.

[10] C. Févotte, L. Daudet, S. J. Godsill, B. Torrésani, Sparse regression with structured priors: Application to audio denoising, in: IEEE International Conference on Acoustics, Speech, and Audio Signal, Toulouse, France, 2006.

[11] M. Fornasier, H. Rauhut, Recovery algorithm for vector-valued data with joint sparsity constraints, SIAM (to appear).

[12] M. Fornasier, H. Rauhut, Iterative thresholding algorithms, Applied and Computational Harmonic AnalysisDoi:10.1016/j.acha.2007.10.005.

[13] K. Gröchenig, S. Samarah, Nonlinear approximation with local Fourier bases, Constr. Approx. 16 (3) (2000) 317–331.

[14] M. Kowalski, B. Torrésani, Random models for sparse signals expansion on unions of basis with application to audio signals, Preprint (2007).

[15] M. Kowalski, B. Torrésani, Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients, Preprint (2007).

[16] S. Mallat, A wavelet tour of signal processing, Academic Press, 1998.

[17] B. Rao, K. Engan, S. Cotter, J. Palmerand, K. Kreutz-Delgado, Subset selection in noise based on diversity measure minimization, IEEE Transaction On Signal Processing 51 (3) (2003) 760–770.
[18] B. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection, IEEE Transaction On Signal Processing 47 (1) (1999) 187–200.
[19] V. S. Rychkov, On restrictions and extensions of the besov and triebel–lizorkin spaces with respect to lipschitz domains, Journal of London Mathematical Society 60 (1) (1999) 237–257.
[20] S. Samarah, S. Obeidat, R. Salman, A shur test for weighted mixed-norm spaces, Analysis Mathematica 31 (2005) 277–289.
[21] S. Samarah, R. Salman, Local fourier bases and modulation spaces, Turkish Journal of Mathematics 30 (4) (2006) 447–462.
[22] J.-L. Starck, M. Elad, D. L. Donoho, Image decomposition via the combination of sparse representation and a variational approach, IEEE Transaction on Image Processing 14 (10) (2005) 1570–1582.
[23] M. Szafranski, Y. Grandvalet, P. Morizet-Mahoudeaux, Hierarchical penalization, in: J. Platt, D. Koller, Y. Singer, S. Roweis (eds.), Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA, 2008, in press.
[24] G. Teschke, Multi-frames representations in linear inverse problems with mixed multi-constraints, Applied and Computational Harmonic Analysis 22 (1) (2006) 43–60.
[25] G. Teschke, R. Ramlau, An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector valued regimes and an application to color image inpainting, Preprint (2007).
[26] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Serie B 58 (1) (1996) 267–288.
[27] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society Serie B 68 (1) (2006) 49–67.
[28] P. Zhao, G. Rocha, B. Yu, Grouped and hierarchical model selection through composite absolute penalties, Preprint (2006).
[29] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of Royal Statistical Society Serie B 67 (2) (2005) 301–320.