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 exploit in an easy way, both structure and sparsity in the framework of regression problems, and introduce implicitly couplings between regression coefficients. Regression is done through optimization problems, and corresponding algorithms are described and analyzed. Beside the classical sparse regression problem, multi-layered expansion on unions of dictionaries of signals are also considered. These sparse structured expansions are done subject to an exact reconstruction 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

Over 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 become significantly easier. Among the successes of sparse methods, one may mention applications to signal coding and compression [10], denoising (Basis Pursuit Denoising and related techniques [5]), source separation [13] 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, an assumption on the coefficients independence is implicitly done, although the latter is generally too coarse a modeling. A good example is provided by the (sparse) time-frequency representations displayed on Figure 2 (Section 5), 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.

On the one hand, we propose sparse expansion methods that explicitly introduce a notion of *structured sparsity*. On the other hand we combine this approach with multilayered signal expansion approaches, which aim at decomposing signals in sums of significantly different components (termed "layers") (see [3,11,10,28]).

In this paper, structured sparsity is modeled by introducing a coupling between coefficients in the same structured set. Some probabilistic approaches have implemented this kind of structured modeling with

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

success (see [13,17]). In the framework 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 [15] and Teschke and Ramlau [31], under the name of joint sparsity: these authors 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 doubly 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.

Considering the mixed norm $\ell_{p,q}$ norm defined as

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

we shall be mainly concerned with the regression problem

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

with $\lambda > 0$ a fixed parameter.

When $\{\phi_{i,j}\}$ is a basis, we give practical estimates for the regression coefficients $\alpha_{i,j}$, obtained by generalized soft thresholding. These estimates are summarized in Theorem 3. 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. This former case is well adapted when the observation of the signal is noisy. In the case of a noiseless observation, we shall be interested by an *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 FOCal Underdetermined System Solver (FOCUSS) algorithm [24,8] 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 the following regression problem

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

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. Mixed norms are defined in Section 2, and we give an overview of how some mixed norms have been used in the literature. Section 3 uses the FOCUSS algorithm to minimize these norms subject to an equality constraint, and extends the algorithm for multilayered expansion. The noisy signal estimation problem is studied in Section 4: after a reminder on the corresponding regularized FOCUSS

[1] The approach of [15,31] clearly applies directly to this more general situation.

algorithm, we discuss the thresholded Landweber iteration used to minimize the functional. Section 5 gives a simple illustration of possible uses of the algorithms, in order to illustrate the relevance of mixed norms for easily modeling dependences between coefficients.

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 [26].

We use the following notation. Let the vector $\mathbf{x} \in \ell_2(\mathbb{C})$, labeled using a double index, be such that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots)$ and for all k, $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,\ell}, \ldots)$. This double index is purely conventional and is used to introduce some dependences between coefficients

Example 1 *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, ℓ) *the index, with* k *the time index and* ℓ *the frequency index. Then,* $x_{k,\ell}$ *represents the coefficient at time* k *and frequency* ℓ*, the vector* $\mathbf{x}_k$ *represents all the frequency coefficients at time* k*, and we denote by* $\mathbf{x}_{.,\ell}$ *the vector which contains all the time coefficients at frequency* ℓ*.*

Definition 1 *Let* $\mathbf{w} \in \ell_2(\mathbb{R})$ *be such that for all* k, ℓ*,* $w_{k,\ell} > 0$*. Let* $p \geq 1$ *and* $q \geq 1$*. We call (weighted) mixed norm of* $\mathbf{x} \in \ell_2(\mathbb{C})$*, the norm* $\ell_{\mathbf{w};p,q}$ *defined by*

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

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

Let us stress that not all vectors in $\ell_2(\mathbb{C})$ have finite mixed-norm. However, applications of mixed norms will consider signal in some specific functional spaces for which the mixed-norms will be finite. A short recall of such spaces is made at the end of this section.

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

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

where $|\mathbf{x}_{.,\ell}|^p = (|x_{1,\ell}|^p, \ldots, |x_{k,\ell}, \ldots|^p)$, and where $W_{.,\ell} = \mathrm{diag}(w_{1,\ell}, \ldots, w_{k,\ell}, \ldots)$, i.e $W_{.,\ell}$ is the diagonal matrix with diagonal entries equal to the vector of weights $(w_{1,\ell}, \ldots, w_{k,\ell}, \ldots)$.

Remark 1 *The mixed norms generalize the usual* ℓ_p *norms. Indeed, 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 made more explicit in Section 4, with the Bayesian formulation of the regression problem. The coupling is strongly dependent of the choice of p and q. Indeed, if we consider a vector as sparse/concentrated if it contains a lot of values close to zero, minimising an ℓ_p norm encourages sparsity for small values of p ($p < 2$) and diversity for large values ($p \geq 2$). Thus, ℓ_p are usually used as measures of diversity for small values of p ($p < 2$) and sparsity for large values ($p \geq 2$) (see [24] and references therein for a more detailled discution about sparsity/diversity measures). 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 this framework, it is preferable to have to deal with a convex optimization problem in order to have guaranties of global optimability. 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. ∎

Mixed norms make it possible to favor certain kind of structures we can find in signals. Classical properties of norms (and, in particular, convexity) allow us to use them in regression problems. The following subsection recalls some models already studied which use a mixed norm to group variables.

2.1. *Mixed norms in the literature*

Some specific instances of mixed norms are already used in various situations. This section presents functional spaces characterized by mixed norms, and 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. The Besov, Triebel-Lizorkin spaces and the modulation spaces are characterized with mixed norms. Besov and Triebel-Lizorkin spaces are described in [25], and [12] gives a good overview for modulation spaces.

Let $s \in \mathbb{R}$ and $0 < p, q \leq \infty$. Let $\phi_0 \in \mathcal{S}$, $\mathcal{S}$ denoting the Schwartz space, following some specific properties (see [25]) 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 $\|f\|_{B^s_{p,q}}$ can be viewed as 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 [21,6].

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 $\|f\|_{F^s_{p,q}}$ can also be viewed as a mixed norm with $\mathbf{w} = \{2^{jsq}\}$. In comparison with Besov spaces, the roles of the two indices are interchanged [16].

The modulation spaces are characterized also 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 [27]

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

2.1.2. *Statistical regression and inverse problems*

Several sparse regression techniques have been studied in a supervised learning context. The most classical one is the ℓ_1 regression, known as the lasso estimate [32], well known in the signal processing community as the Basis pursuit denoising [5].

The group-lasso [33] introduced by Yan and Lin, uses the mixed norm $\ell_{2,1}$. This norm was introduced to preserve entire groups of individuals. More recently Fornasier and Rauhut studied more generally some $\ell_{p,1}$ mixed norm for inverse problems in [15], and Teschke and Ramlau used these norms in [31]. Cotter *et al.* [8] used this kind of norms in the same context and provided a comparison of two classes of algorithms: Matching Pursuit and FOCUSS. Another example is provided by the hierarchical penalization [29] 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, mixed norm were used by Peng Zhao *et al* in [34], under the name of "Composite Absolute Penalties". That paper more particularly studies algorithms using the $\ell_{\infty,1}$ mixed norm.

In the following sections, we focus on regression problems. We show how classical algorithms used with ℓ_p norms regularization can be extended to the use of mixed norms. Moreover, we give general algorithms able to handle the structure in layers for signal, like the morphological component analysis [28]. 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 also relevant.

3. Signal estimation under equality constraint

We first address the problem of function or signal estimation subject to an exact reconstruction, in the finite dimensional case. We prove that the FOCUSS algorithm can be adapted to tackle the case of mixed norms, and we generalize it for a decomposition into layers. We limit ourselves to the finite dimension setting to follow the original setting of FOCUSS. The generalisation to infinite dimension requires to deal with Lagrange multipliers in Banach spaces [20] and then is not straightforward.

Let $\mathbf{y} \in \mathbb{C}^M$. Let $\mathbf{x} \in \mathbb{C}^N$ with $N = K \times L$ be such that $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_K)$ and for all k, $\mathbf{x}_k = (x_{k,1}, \dots, x_{k,\ell}, \dots, x_{k,L})$. 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, while the problem remains convex for $1 \leq p, q \leq 2$.

3.1. *Minimization of a mixed norm*

We want to solve the problem:

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

Problem (3) can be solved with the FOCUSS algorithm [24], which has been designed to minimize a diversity measure (or equivalently, to maximise the sparsity of the solution), 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 $\bar{\mathbf{x}}$ denotes the conjugate of $\mathbf{x}$, $\alpha(\mathbf{x})$ is a scalar and $\Pi(\mathbf{x})$ is a matrix. 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 $x_{k,\ell}$ is

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

Then

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

FOCUSS is a simple iterative scheme to solve (3) given by the following algorithm [2] :

Algorithm 1

Let $\mathbf{x}^{(0)} \in \mathbb{C}^N$ *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 $\mathbf{x}^{(0)}$ of (3) is a vector which satisfies the constraint equality of the problem, and A^* denotes the Hermitian transpose of A.

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:* $\mathbf{x}^{(0)} = A^+\mathbf{y}$.

Theorem 1 *Starting from a bounded feasible solution $\mathbf{x}^{(0)} \in \mathbb{C}^N$, the sequence of iterates generated by Algorithm 1 is convergent, and minimizes the $\ell_{\mathbf{w};p,q}$ diversity measure, for $p, q < 2$ $(p, q \neq 0)$.*

PROOF. We rewrite and adapt the original proof in [24]. 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

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

where $P^\star = \left(\Pi^{-1}(\mathbf{x}^\star)\right)^{1/2}$.

In the problem we consider, point *iii*) does not matter, because $\mathcal{A}$ is here a continuous function. Point *i*) can be proved exactly as in [24]. Then only the point *ii*) remains to be proved.

To prove point *ii*), we use the same technique as the one used in the original proof, and we 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 minimal ℓ_2 norm solution $\mathbf{b}^{(m+1)}$ of the problem $AP^{(m+1)}\mathbf{b} = \mathbf{y}$, where $P^{(m+1)} = \mathrm{diag}\left(w_{k,\ell}^{-1/2}|x_{k,\ell}^{(m)}|^{\frac{2-p}{2}}\|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{\frac{p-q}{2}}\right)$. Then we have $\mathbf{x}^{(m+1)} = P^{(m+1)}\mathbf{b}^{(m+1)}$.

[2] One can remark that the algorithm does not depend on $\alpha(\mathbf{x})$.

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

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

We can stress that

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

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

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

where we have used in the $4th$ line Hölder's 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's inequality. ∎

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 or frames. Then, it makes sense to use different (mixed) norms on the corresponding coefficients. FOCUSS may be adapted to this new situation, as we show below.

For simplicity, we limit the present discussion to the case of two layers only. The generalization to an arbitrary number 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 $N_i = K_i \times L_i$. Suppose that $\mathbf{x}^{[i]} = (\mathbf{x}_1^{[i]}, \ldots, \mathbf{x}_{K_i}^{[i]})$ and for all $k \in \{1, \ldots, K_i\}$, $\mathbf{x}_k^{[i]} = (x_{k,1}^{[i]}, \ldots, x_{k,\ell}^{[i]}, \ldots, x_{k,L_i}^{[i]})$. Let $A \in \mathbb{C}^{M \times N}$ be such that $A = A_1 \bigoplus A_2$, 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:

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

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

We denote the diversity measure by

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

$$= \operatorname{sgn}(p_1 q_1)\lambda_1 \sum_{k_1=1}^{K_1} \left(\sum_{\ell_1=1}^{L_1} w_{k_1,\ell_1}^{[1]} |x_{k_1,\ell_1}^{[1]}|^{p_1} \right)^{q_1/p_1} + \operatorname{sgn}(p_2 q_2)\lambda_2 \sum_{k_2=1}^{K_2} \left(\sum_{\ell_2=1}^{L_2} w_{k_2,\ell_2}^{[2]} |x_{k_2,\ell_2}^{[2]}|^{p_2} \right)^{q_2/p_2} \tag{8}$$

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

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

$$\frac{\partial E(\mathbf{x})}{\partial x_{k_1,\ell_1}^{[1]}} = \operatorname{sgn}(p_1)\lambda_1 |q_1| w_{k_1,\ell_1}^{[1]} x_{k_1,\ell_1}^{[1]} |x_{k_1,\ell_1}^{[1]}|^{p_1-2} \|\mathbf{x}_{k_1}^{[1]}\|_{\mathbf{w}^{[1]};p_1}^{q_1-p_1} \tag{10}$$

$$\frac{\partial E(\mathbf{x})}{\partial x_{k_2,\ell_2}^{[2]}} = \operatorname{sgn}(p_2)\lambda_2 |q_2| w_{k_2,\ell_2}^{[2]} x_{k_2,\ell_2}^{[2]} |x_{k_2,\ell_2}^{[2]}|^{p_2-2} \|\mathbf{x}_{k_2}^{[2]}\|_{\mathbf{w}^{[2]};p_2}^{q_2-p_2} \ . \tag{11}$$

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

$$\Pi_1(\mathbf{x}^{[1]}) = \operatorname{sgn}(p_1)\lambda_1 |q_1| \operatorname{diag}(w_{k_1,\ell_1}^{[1]} |x_{k_1,\ell_1}^{[1]}|^{p_1-2} \|\mathbf{x}_{k_1}^{[1]}\|_{\mathbf{w}^{[1]};p_1}^{q_1-p_1})$$

and

$$\Pi_2(\mathbf{x}^{[2]}) = \operatorname{sgn}(p_2)\lambda_2 |q_2| \operatorname{diag}(w_{k_2,\ell_2}^{[2]} |x_{k_2,\ell_2}^{[2]}|^{p_2-2} \|\mathbf{x}_{k_2}^{[2]}\|_{\mathbf{w}^{[2]};p_2}^{q_2-p_2}) \ .$$

A first idea would be to apply the FOCUSS Algorithm 1 without any change. Unfortunately, the 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 henceforth the convergence, let us take again the ideas of the previous proof, with the same notations. We rewrite inequality (5) as

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

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

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

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

Algorithm 2

Let $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *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 $E(\mathbf{x}^{(m+1)}) \geq E(\mathbf{x}^{(m)})$ **then**
if $E_1(\mathbf{x}^{[1]^{(m+1)}}) > E_1(\mathbf{x}^{[1]^{(m)}})$, **then** $\mathbf{x}^{[1]^{(m+1)}} = \mathbf{x}^{[1]^{(m)}}$
else $\mathbf{x}^{[2]^{(m+1)}} = \mathbf{x}^{[2]^{(m)}}$ *% (i.e* $E_2(\mathbf{x}^{[2]^{(m+1)}}) > E_2(\mathbf{x}^{[2]^{(m)}})$ *)%***endif**
endif
until *convergence*

This way, we ensure that E_1 or E_2 decrease stricly during the iterations, so that the energy E decreases strictly. This 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 [23] 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

In this section, we deal with the infinite dimensional case. Let $\mathbf{y} \in \mathcal{H}$, with $\mathcal{H} \subset L_2(\mathbb{R})$ a separable Hilbert space, let $\mathbf{x} \in \ell_2(\mathbb{C})$ and a linear operator $A : \ell_2(\mathbb{C}) \to \mathcal{H}$. We are interested here by the noisy case: i.e. the case of observations of the form $\mathbf{y} = A\mathbf{x} + b$, where $b \in \mathcal{H}$ is an unspecified noise. We follow the classical variational formulation of the problem: the regression is made by minimizing the ℓ_2 error between the observed signal and its estimate, regularized by a mixed norm. Hence, 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 usually justified by assuming a Gaussian i.i.d. distribution for the noise. The mixed norm corresponds to a prior on the coefficients of the form

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

$$= \prod_k \exp\{-\frac{\lambda}{q}\|\mathbf{x}_k\|_{\mathbf{w};p}^q\} \ , \tag{17}$$

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 minimize a functional of the form:

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

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

In the finite dimensional case, in order to minimize (18), the FOCUSS algorithm can be adapted as suggested in [23]. We provide here the modified algorithm for $I = 2$ only for the sake of simplicity

Algorithm 3
Let $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *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 $E(\mathbf{x}^{(m+1)}) \geq E(\mathbf{x}^{(m)})$ **then**
if $E_1(\mathbf{x}^{[1]^{(m+1)}}) > E_1(\mathbf{x}^{[1]^{(m)}})$, **then** $\mathbf{x}^{[1]^{(m+1)}} = \mathbf{x}^{[1]^{(m)}}$
else $\mathbf{x}^{[2]^{(m+1)}} = \mathbf{x}^{[2]^{(m)}}$ *%(i.e.* $E_2(\mathbf{x}^{[2]^{(m+1)}}) > E_2(\mathbf{x}^{[2]^{(m)}})$*) %***endif**
endif
until *convergence*

with

$$\Pi_1(x^{[1]}) = \mathrm{sgn}(p_1)\lambda_1|q_1| \ \mathrm{diag}(w_{k_1,\ell_1}^{[1]}|x_{k_1,\ell_1}^{[1]}|^{p_1-2}\|\mathbf{x}_{k_1}^{[1]}\|_{\mathbf{w}^{[1]};p_1}^{q_1-p_1}) \ , \tag{19}$$

and

$$\Pi_2(x^{[2]}) = \mathrm{sgn}(p_2)\lambda_2|q_2| \ \mathrm{diag}(w_{k_2,\ell_2}^{[2]}|x_{k_2,\ell_2}^{[2]}|^{p_2-2}\|\mathbf{x}_{k_2}^{[2]}\|_{\mathbf{w}^{[2]};p_2}^{q_2-p_2}) \ . \tag{20}$$

Unlike the algorithm given by Rao *et al.* in [23], one can notice 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 a matrix inversion involved at each iteration.

A valuable alternative is provided by thresholded Landweber iteration algorithm, like the algorithm introduced by Daubechies *et al.* in [9]. In the next subsection, we study the simple case where A is unitary (and corresponds to the operator of an orthogonal basis). In this case, the iterative thresholding algorithm is developed for $1 \leq p, q \leq 2$. Although this algorithm is more restrictive in terms of admissible values for p and q than FOCUSS, it is really faster.

4.1. *The unitary case*

In this subsection, A is assumed to be an unitary operator: denoting by A^* the adjoint of A, $A^*A = AA^* = Id$. This allows us to introduce some useful operators associated with a given mixed norm. Moreover, the regression problem formulated in the unitary case gives a good idea of the influence of the mixed norms when they are used as a penalty term.

Regression in the unitary 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

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

Several cases have to be dealt with:

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

As A is an unitary operator, we have

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

Let $\tilde{\mathbf{y}} = A^*\mathbf{y}$ and $\theta_{x_{k,\ell}}$ (resp. $\theta_{\tilde{y}_{k,\ell}}$) be the argument of $x_{k,\ell}$ (resp. $\tilde{y}_{k,\ell}$). We have, for all k, ℓ

$$|\tilde{y}_{k,\ell} - x_{k,\ell}|^2 = |\tilde{y}_{k,\ell}|^2 + |x_{k,\ell}|^2 - 2|x_{k,\ell}||\tilde{y}_{k,\ell}|\cos(\theta_{\tilde{y}_{k,\ell}} - \theta_{x_{k,\ell}}) \ .$$

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

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

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

$$2|x_{k,\ell}||\tilde{y}_{k,\ell}| \sin(\theta_{\tilde{y}_{k,\ell}} - \theta_{x_{k,\ell}}) = 0 \ . \tag{24}$$

From (23) and (24), one can deduce that $\theta_{x_{k,\ell}} = \theta_{\tilde{y}_{k,\ell}}$ and state that variational equations are equivalent to the following system:

$$\begin{cases} |x_{k,\ell}| & = |\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} |x_{k,\ell}|^{p-1} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p} \\ \arg(x_{k,\ell}) & = \arg(\tilde{y}_{k,\ell}) \end{cases} \tag{25}$$

The variational equations are coupled, so that their solution may be 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 shrinkage operation, as suggested by the variational equations. In all cases, the argument of $x_{k,\ell}$ is the same as $\tilde{y}_{k,\ell}$.

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

Let us introduce the function $\mathcal{F} : \ell_2(\mathbb{C}) \to \ell_2(\mathbb{R})$, $|\mathbf{v}| \mapsto \mathcal{F}(|\mathbf{v}|) = |\mathbf{v}| + \lambda W_k |\mathbf{v}|^{p-1} \|\mathbf{v}\|_{\mathbf{w};p}^{q-p}$. $\mathcal{F}$ 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 fixed point Algorithm 4 (see below).

Proposition 2 *Let A be a unitary operator. We suppose that, for all k, ℓ, $w_{k,\ell} = w_k$. The solution of Problem* (21)*, where $\ell_{p,q} = \ell_{2,1}$ is given by*[3]

$$x_{k,\ell} = \tilde{y}_{k,\ell} \left(1 - \frac{\lambda \sqrt{w_k}}{\|\tilde{\mathbf{y}}_k\|_2} \right)^+ \ .$$

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

Let us stress that the weighting term w_k in the solution above, depends only on the index k, and does not depend on the index ℓ. Hence, we can rewrite the solution in vector form:

$$\mathbf{x}_k = \tilde{\mathbf{y}}_k \left(1 - \frac{\lambda \sqrt{w_k}}{\|\tilde{\mathbf{y}}_k\|_2} \right)^+ \ . \tag{26}$$

Remark 3 *The result given in Proposition 2 shows a mixture of a weighting (remembering the ℓ_2 minimization), and a thresholding (remembering the ℓ_1 minimization) which acts on an entire group of variables. In this case, the groups with many "big"(or "significant") coefficients are kept (i.e. not set to zero) over the groups with small coefficients. The coupling appears to be between significant coefficients.*

Furthermore, note that this solution is identical to the group-lasso estimate, given in [33].

For the more general case $q = 1$ and $1 < p < 2$ we cannot give any analytical expression for the solution, and Equation (25) must be solved numerically. In the *finite dimensional case* a simple iterative thresholding algorithm can be constructed.

[3] For $x \in \mathbb{R}$, we shall set $x^+ = \max(0, x)$.

Algorithm 4
Let $\mathbf{x}^{(0)} = \tilde{\mathbf{y}}$
For *all* k, ℓ **do**
$x_{k,\ell}^{(m+1)} = \arg(\tilde{y}_{k,\ell}) \left(\tilde{y}_{k,\ell} - \lambda w_{k,\ell} |x_{k,\ell}|^{p-1} \|\mathbf{x}_k\|_{\mathbf{w}_k;p}^{q-p}\right)^+$
endfor

Proposition 3 *Let* $\tilde{\mathbf{y}} \in \mathbb{C}^{K \times L}$ *and* $L_{\mathbf{w}_k} = \sum_{\ell=1}^{L} w_{k,\ell}$. *Fixed point algorithm 4 converges for*

$$\lambda < \frac{L_{\mathbf{w}_k}^{\frac{2(p-1)}{p(2-p)}}}{2(p-1)} \min_k(\|\mathbf{y}_k\|_{\mathbf{w}_k;p-2}) \ .$$

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

Remark 4 *Fornasier and Rauhut studied more specifically this kind of mixed norms in [15]. They show that the solution is given by a shrinkage operator and give the analytical solutions for the norms* ℓ_1, $\ell_{2,1}$ *and* $\ell_{\infty,1}$. *The study of the shrinkage operator for the* $\ell_{p,1}$ *mixed norm was also done by Teschke and Ramlau in [31] for non linear inverse problems. Here, we gave Algorithm 4 to find the minimizer of 21 with* $q = 1$ *and* $1 < p < 2$, *in the unitary case, and Proposition 3 gave a sufficient condition for the convergence.*

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

We show here that the solution is obtained by a coordinatewise soft-thresholding. The following 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 a unitary operator. The solution of Problem* (21), *with* $\ell_{p,q} = \ell_{1,q}$ *is given by a soft thresholding operator.*

For each k, *we denote* $\check{y}_{k,\ell_k}$ *(resp.* $\check{w}_{k,\ell_k}$*) the coefficients* $|\tilde{y}_{k,\ell_k}|$ *(resp.* $\check{w}_{k,\ell_k}$*) sorted such that the* $r_{k,\ell_k} = \frac{|\tilde{y}_{k,\ell_k}|}{w_{k,\ell_k}}$ *are ordered by descending order (for each* k*).*

The threshold is given by $w_{k,\ell}\xi_k$, *with* ξ_k *the (unique) solution in* $\mathbb{R}_+$ *of the following equation*

$$\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} L_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} |||\tilde{\mathbf{y}}_k|||_{\mathbf{w}_k} \ ,$$

with $|||\tilde{\mathbf{y}}_k|||_{\mathbf{w}_k} = \sum_{\ell=1}^{L_k} \check{w}_{k,\ell} \check{y}_{k,\ell}$ *and* $L_{\mathbf{w}_k} = \sum_{\ell_k=1}^{L_k} \check{w}_{k,\ell_k}^2$. *The quantity* L_k *is the number such that*

$$r_{k,L_k+1} \leq \lambda \left(\sum_{\ell_k=1}^{L_k+1} w_{k,\ell_k}^2 (r_{k,\ell_k} - r_{k,L_k+1}) \right)^{q-1} \quad \text{and} \quad r_{k,L_k} > \lambda \left(\sum_{\ell_k=1}^{L_k} w_{k,\ell_k}^2 (r_{k,\ell_k} - r_{k,L_k}) \right)^{q-1} \ .$$

In particular, for $q = 2$, *the threshold* $w_{k,\ell}\xi_k$ *is equal to*

$$\frac{\lambda w_{k,\ell}}{1 + L_{\mathbf{w}_k}\lambda} |||\tilde{\mathbf{y}}_k|||_{\mathbf{w}_k} \ .$$

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

Remark 5 *This proposition generalises the result given in the wavelet and Besov spaces framework in [4].*

In the case $q = 2$, the $\ell_{1,2}$ norm is associated with the problem called *elitist-lasso* [18] (as opposite of group-lasso). Let us point out some properties of this estimator in the following two remarks. These properties illustrate the expected consequence of using this mixed norm in the context of regression.

Remark 6 *After a suitable rewriting, the solution is obtained by a mixture of ℓ_2-like weighting and ℓ_1-like soft thresholding:*

$$\begin{aligned}|x_{k,\ell}| &= |\tilde{y}_{k,\ell}| - \frac{\lambda}{1+\lambda L_{\mathbf{w}_k}} \|\tilde{\mathbf{y}}_k\|_{\mathbf{w}_k} \\ &= |\tilde{y}_{k,\ell}| \left(1 - \frac{\lambda w_{k,\ell}}{1+\lambda L_{\mathbf{w}_k}}\right) - \frac{\lambda}{1+\lambda L_{\mathbf{w}_k}} \sum_{l=1,\check{y}_l \neq y_\ell}^{L_k} \check{w}_{k,l} \check{y}_{k,l} \ .\end{aligned}$$

This mixture of weighting and thresholding is here very different from the one we obtained for the case of the mixed norm $\ell_{2,1}$ (see Proposition 2). Here, we weight each coefficient, before comparing it to a threshold which depends on the norm of the group k. 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. In other words, only the largest coefficients – compared to the others – are kept, for each index k.

Remark 7 *Let us take a look at the particular case $\lambda >> 1$ and $w_{k,\ell} = 1$ for all k, ℓ, for the $\ell_{1,2}$ mixed norm. Then, for a fixed k, at least one coefficient is not set to zero. Indeed, if all the coefficients are set to zero, then $\|\tilde{\mathbf{y}}_k\| = 0$ and no coefficient is thresholded, which is a contradiction. Consequently, the $\ell_{1,2}$ mixed norm cannot give estimates as sparse as the ℓ_1 norm. This property is the consequence of the structures and illustrate well the coupling between significant and unsignificant coefficients as explained before in Remark 6.*

4.2. *Summary of the main results*

Here, we summarize the preceeding results in a single theorem:

Theorem 3 *Let $\mathbf{x} \in \ell_2(\mathbb{C})$ and $\mathbf{z} \in \ell_2(\mathbb{C})$. Let $1 \leq p, q \leq 2$ and $\mathbf{w} = (w_{k,\ell})$ a sequence of stricly positive weights such that $\|\mathbf{x}\|^q_{\mathbf{w};p,q} = \sum_k \left(\sum_\ell w_{k,\ell} |x_{k,\ell}|^p\right)^{q/p}$. Then, the solution of the following optimization problem*

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|^q_{\mathbf{w};p,q} \ ,$$

is given by $\mathbf{S}^\lambda_{\mathbf{w};p,q}(\mathbf{z})$, with $\mathbf{S}^\lambda_{\mathbf{w};p,q}$ a "generalized" soft-thresholding (or shrinkage) operator defined coordinatewise by, for all k, ℓ:

$$v_{k,\ell} \mapsto \arg(v_{k,\ell}) \left(|v_{k,\ell}| - \xi_{k,\ell}(\lambda)\right)^+ \ ,$$

where the $\xi_{k,\ell}(\lambda)$, simply denoted by $\xi_{k,\ell}$, 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 $\mathcal{F} : \ell_2(\mathbb{R}) \to \ell_2(\mathbb{R})$, $|\mathbf{v}| \mapsto \mathcal{F}(|\mathbf{v}|) = |\mathbf{v}| + \lambda W_k |\mathbf{v}|^{p-1} \|\mathbf{v}\|^{q-p}_{\mathbf{w};p}$;*
- *If $p = q = 1$ then $\xi_{k,\ell} = \lambda w_{k,\ell}$.*
- *If $p = q = 2$ then $\xi_{k,\ell} = \frac{\lambda w_{k,\ell}}{1+\lambda w_{k,\ell}}$.*
- *If $p = 1$ and $1 < q < \infty$. For each k, we denote $\check{z}_{k,\ell_k}$ (resp. $\check{w}_{k,\ell_k}$) the coefficients $|z_{k,\ell_k}|$ (resp. $\check{w}_{k,\ell_k}$) sorted such that the $r_{k,\ell_k} = \frac{|z_{k,\ell_k}|}{w_{k,\ell_k}}$ are ordered by descending order. The threshold is $\xi_{k,\ell} = w_{k,\ell}\xi_k$, where ξ_k is the solution on $\mathbb{R}_+$ of*

$$\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} L_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} \|\mathbf{z}_k\|_{\mathbf{w}_k} \ ,$$

with $\|\mathbf{z}_k\|_{\mathbf{w}_k} = \sum_{\ell_k=1}^{L_k} \check{w}_{k,\ell_k} \check{z}_{k,\ell_k}$ and L_k is the quantity verifying:

$$r_{k,L_k+1} \leq \lambda \left(\sum_{\ell_k=1}^{L_k+1} w^2_{k,\ell}(r_{k,\ell_k} - r_{k,L_k+1})\right)^{q-1} \quad \text{and} \quad r_{k,L_k} > \lambda \left(\sum_{\ell_k=1}^{L_k} w^2_{k,\ell}(r_{k,\ell_k} - r_{k,L_k})\right)^{q-1} \ .$$

and $L_{\mathbf{w}_k} = \sum_{\ell_k=1}^{L_k} \check{w}^2_{k,\ell_k}$.

In particular, if $p = 1$ *and* $q = 2$,

$$\xi_{k,\ell} = \frac{\lambda w_{k,\ell}}{1 + L_{\mathbf{w}_k}\lambda} \|\mathbf{z}_k\|_{\mathbf{w}_k} \ .$$

– *If* $p = 2$ *and* $q = 1$, *and for a fixed* k, $w_{k,\ell} = w_k \ \forall \ell$. *Then*

$$\xi_{k,\ell} = \frac{\lambda\sqrt{w_k}}{|z_{k,\ell}|\|\mathbf{z}_k\|_2} \ ;$$

– *If* $1 < p < 2$ *and* $q = 1$. *Then the solution is given by Algorithm 4 in finite dimension.*

Remark 8 *Theorem 3 gives the so-called proximity operator associated with the mixed norm* $\ell_{p,q}$. *Proximity operators were introduced by Moreau [22] in the 60's and nicely used more recently by Combettes* et al. *in particular in [7] to minimize some nondifferentiable convex functionals.*

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

Corollary 1 *Let* $\mathbf{x} \in \ell_2(\mathbb{C})$. *Let* $A : \ell_2(\mathbb{C}) \to \mathcal{H}$ *be an unitary operator. Let* $1 \leq p, q \leq 2$ *and* $\mathbf{w} = (w_{k,\ell})$ *a strictly positive sequence such that* $\|\mathbf{x}\|^q_{\mathbf{w};p,q} = \sum_k \left(\sum_\ell w_{k,\ell}|x_{k,\ell}|^p\right)^{q/p}$.

Then, for all $\mathbf{y} \in \mathcal{H}$, *there exists a strictly positive sequence* $\boldsymbol{\xi} = (\xi_{k,\ell})$ *(which depends on* $\mathbf{y}$*) such that the minimum of the functional*

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

coincides with the minimum of the functional

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

We saw how to obtain an estimation of the minimizer when A is a unitary operator, and we defined operators allowing us to obtain the solution. Next section generalizes this problem, and shows that the corresponding iterative algorithm inspired by Daubechies *et al.* in [9] and by Teschke in [30] 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 Problem (21). A is now 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 a union of orthogonal bases, one can apply the Block Coordinate Relaxation algorithm, as it was shown in [18]. The algorithm studied here is more general, and can be applied to any linear operator A, e.g. a matrix corresponding to frame (or union of frames), convolution operator, etc.

Let us introduce the functional

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \sum_{i=1}^{I} A_i\mathbf{x}^{[i]}\|_2^2 + \sum_{i=1}^{I} \frac{\lambda_i}{q_i}\Psi_i(\mathbf{x}^{[i]}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \boldsymbol{\lambda}\Psi(\mathbf{x}) \ . \tag{27}$$

with $\mathbf{y} \in \mathcal{H}$, for all i, A_i is a linear operator and $A = \bigoplus A_i$. Let $\mathbf{x} = (\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[I]}) \in \ell_2(\mathbb{C})^I$ 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 $\Psi(\mathbf{x}) = (\Psi_1(\mathbf{x}^{[1]}), \ldots, \Psi_I(\mathbf{x}^{[I]}))^T$.

Here, the penalty term is 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 Problem (27), following [9,30], we introduce a surrogate functional

$$\Phi^{sur}(\mathbf{x},\mathbf{a}) = \frac{1}{2}\|\mathbf{y}-A\mathbf{x}\|_2^2 + \frac{C}{2}\|\mathbf{x}-\mathbf{a}\|_2^2 - \frac{1}{2}\|A\mathbf{x}-A\mathbf{a}\|_2^2 + \boldsymbol{\lambda}\Psi(\mathbf{x}) \tag{28}$$

$$= \frac{1}{2}\sum_{i=1}^{I}\left(\sum_k\sum_\ell\left(C(x_{k,\ell}^{[i]})^2 - 2x_{k,\ell}^{[i]}[A_i^*\mathbf{y}+C\mathbf{a}^{[i]}-A_i^*A\mathbf{a}]_{k,\ell}\right) + \frac{\lambda_i}{q_i}\sum_{k=1}^{K}\left(\sum_\ell w_{k,\ell}^{[i]}|x_{k,\ell}^{[i]}|^{p_i}\right)^{q_i/p_i}\right) \tag{29}$$

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

with C such that $C > C_1 + \ldots + C_I$, where each C_i bounds the operator norm of $A_i^*A_i$.

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

$$\left|x_{k,\ell}^{[i]}\right| = C^{-1}\left|[A_i^*\mathbf{y}+C\mathbf{a}^{[i]}-A_i^*A\mathbf{a}]_{k,\ell}\right| - \frac{\lambda_i}{C}w_{k,\ell}^{[i]}|x_{k,\ell}^{[i]}|^{p_i-1}\|\mathbf{x}_k^{[i]}\|_{\mathbf{w}_k^{[i]};p_i}^{q_i-p_i} \tag{30}$$

$$\arg\left(x_{k,\ell}^{[i]}\right) = \arg\left([A_i^*\mathbf{y}+C\mathbf{a}^{[i]}-A_i^*A\mathbf{a}]_{k,\ell}\right) \tag{31}$$

In the usual situation of the ℓ_1 norm [9,30], 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, the solution of (30) is obtained from $z = C^{-1}[A_i^*\mathbf{y}+C\mathbf{a}^{[i]}-A_i^*A\mathbf{a}]$, using the operators $\mathbf{S}^{\lambda_i/C}_{\mathbf{w}^{[i]};p_i,q_i}$ given by Theorem 3 in Section 4.2, for all i.

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

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

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

PROOF. The surrogate functional decouples the variational equations, so we just have to exploit the work done in Section 4.1. ∎

Then, we can deduce the following iterative algorithm

Algorithm 5
Let $\mathbf{x}^{(0)} \in \ell_2(\mathbb{C})$
Do
For $i = 1 : I$

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

EndFor
until *convergence*

where the $\boldsymbol{\xi}^{[i]^{(m)}}$ are the vectors containing the thresholds $\xi_{k,\ell}^{[i]^{(m)}}$. These thresholds are given by the operators $\mathbf{S}^{\lambda_i/C}_{\mathbf{w}^{[i]};p_i,q_i}$, associated with the adequate ℓ_{p_i,q_i} mixed norm, applied to $\mathbf{z} = C^{-1}[A_i^*\mathbf{y}+C\mathbf{x}^{[i]^{(m)}}-A_i^*A\mathbf{x}^{(m)}]$. Theorem 3 explains how to obtain these thresholds in practice.

As in [9,30], we show that the preceeding algorithm converges, and then that limit is the desired minimizer (i.e. a solution of Problem (27)).

Theorem 4 *Let $I \in \mathbb{N}$. For all $i \in \{1, \dots I\}$, let A_i be a linear operator, $A_i : \ell_2(\mathbb{C}) \to \mathcal{H}$. Let A be the linear operator such that $A = \bigoplus A_i$ and C such that $C > \|A^*A\|$. Suppose $\mathbf{y}$ is an element of $\mathcal{H}$, and the sequence of weights $\mathbf{w}$ is uniformly bounded from below by a stricly positive number. Then, the sequence of iterates generated by Algorithm 5 with $\mathbf{x}^{(0)}$ arbitrarily chosen in $\ell_2(\mathbb{C})$, converges weakly to a fixed point which is a minimizer of functional* (27)*:*

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

PROOF. The proof follows the lines originally given by Daubechies *et al.* in [9], and summarized in various papers (see e.g. [30]) to prove the weak convergence of a thresholded Landweber iteration. We just have to check [9, Lemma 3.6] (or [30, Lemma 14]), and [9, Proposition 3.10] ([30, Proposition 17]). We choose to postpone these proofs to Appendix A.4.

The same algorithm and the weak convergence can be obtained with the approach of proximal algorithm introduced by Combette *et al.* in [7] ∎

The weak convergence of the thresholded Landweber algorithm allows us to state the following theorem

Theorem 5 *Let $I \in \mathbb{N}$. For all $i \in \{1, \dots I\}$, let A_i be a linear operator, $A_i : \ell_2(\mathbb{C}) \to \mathcal{H}$. Let A be the linear operator such that $A = \bigoplus A_i$ and C such that $C > \|A^*A\|$*

There exists a (non unique) sequence of $\mathbb{R}_+$ $\boldsymbol{\xi} = (\xi_{k,\ell}^{[i]})$ such that

$$\Phi(\mathbf{x}) = \|\mathbf{y} - A\mathbf{x}\|_2^2 + \boldsymbol{\lambda}\Psi(\mathbf{x})$$

and

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

have the same minimum reached at the same point $\mathbf{x}^\star$. Moreover, a choice of $\boldsymbol{\xi}$ is given coordinatewise by

$$\xi_{k,\ell}^{(i)} \begin{cases} = \lambda_i w_{k,\ell}^{[i]} |x_{k,\ell}^{[i]^\star}|^{p_i - 1} \|\mathbf{x}_k^{[i]^\star}\|_{\mathbf{w}_k^{[i]};p_i,q_i}^{q_i} & \text{if} \quad x_{k,\ell}^{[i]^\star} \neq 0 \\ \geq B & \text{otherwise.} \end{cases}$$

where

$$B = \prod_{i=1}^{I} \min\{\xi_{k,\ell}^{(i)} \mid x_{k,\ell}^{[i]^\star} \neq 0\}^{-4/p_i} \|\mathbf{y}\|_2^{(2p_i(6+q_i-p_i)-4q_i)/(p_i q_i)} \ .$$

PROOF. Eq. (30) tells us that the minimum of Φ is attained at $\mathbf{x}^\star$ such that, for any $C > \|A^*A\|$

$$|x_{k,\ell}^{[i]^\star}| = C^{-1}[A^*\mathbf{y} + C\mathbf{x}^{[i]^\star} - A_i^* A\mathbf{x}^\star]_{k,\ell} - \frac{\lambda_i}{C} w_{k,\ell}^{[i]} |x_{k,\ell}^{[i]^\star}|^{p_i - 1} \|\mathbf{x}_k^{[i]^\star}\|_{\mathbf{w}_k^{[i]};p_i}^{q_i - p_i} \ .$$

We denote by $\tilde{\mathbf{x}}^\star$ the minimizer of $\tilde{\Phi}(\mathbf{x})$, which is the limit of the sequence $\tilde{\mathbf{x}}^{(m)}$ generated by Algorithm 5 applied to $\tilde{\Phi}$. Lemma 1 ensure that there exist a uniform bound $B(\mathbf{x}^{(0)})$ such that $\|\tilde{\mathbf{x}}^{(m)}\| \leq B(\tilde{\mathbf{x}}^{(0)})$. Then, $\xi_{k,\ell} \geq B(\tilde{\mathbf{x}}^{(0)})$ for $x_{k,\ell}^\star = 0$ ensure that $\tilde{x}_{k,\ell}^\star = 0$. As the convergence does not depend of the choice of $\tilde{\mathbf{x}}^{(0)}$, we can choose $\xi_{k,\ell} \geq B(\mathbf{0})$ where B is given by Lemma 1.

Hence, with $\xi_{k,\ell}^{[i]} = \lambda_i w_{k,\ell}^{[i]} |x_{k,\ell}^{[i]^\star}|^{p_i - 1} \|\mathbf{x}_k^{[i]^\star}\|_{\mathbf{w}_k^{[i]},p_i}^{q_i - p_i}$ when $x_{k,\ell}^{[i]^\star} \neq 0$, we have

$$\mathbf{x}^\star = \operatorname*{argmin}_{\mathbf{x}} \Phi(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x}} \tilde{\Phi}(\mathbf{x}) = \tilde{\mathbf{x}}^\star \ .$$

Moreover, for all $i \in \{1, \dots I\}$

$$\begin{aligned}\|\mathbf{x}^{[i]^\star}\|_{\boldsymbol{\xi}^{(i)};1} &= \lambda_i \sum_{k,\ell \mid x_{k,\ell}^{[i]^\star}\neq 0} w_{k,\ell}^{[i]} |x_{k,\ell}^{[i]^\star}|^{p_i-1} \|\mathbf{x}_k^{[i]^\star}\|_{\mathbf{w}_k^{[i]};p_i}^{q_i-p_i} |x_{k,\ell}^{[i]^\star}| \\ &= \lambda_i \sum_k \|\mathbf{x}_k^{[i]^\star}\|_{\mathbf{w}_k^{[i]};p_i}^{q_i-p_i} \sum_\ell w_{k,\ell}^{[i]} |x_{k,\ell}^{[i]^\star}|^{p_i} \\ &= \lambda_i \|\mathbf{x}^{[i]^\star}\|_{\mathbf{w}^{[i]};p_i,q_i}^{q_i}\end{aligned}$$

then $\boldsymbol{\lambda}\Psi(\mathbf{x}^\star) = \|\mathbf{x}^\star\|_{\boldsymbol{\xi};1}$ and $\Phi(\mathbf{x}^*) = \tilde{\Phi}(\mathbf{x}^\star)$. ∎

Using the previous theorem, we can state

Theorem 6 *The sequence $\mathbf{x}^{(n)}$ of iterates generated by Algorithm 5 with $\mathbf{x}^{(0)}$ arbitrarily chosen in $\ell_2(\mathbb{C})$, converges strongly to a fixed point which is a minimizer of functional* (27).

PROOF. Let $\mathbf{x}^\star$ be the weak limit of $\mathbf{x}^{(n)}$. Let $\tilde{\Phi}$ and $\boldsymbol{\xi}$ be defined by applying Theorem 5 to Φ. Let $\tilde{\mathbf{x}}^{(m)}$ the sequence of iterates generated by Algorithm 5 applied to $\tilde{\Phi}$. By Theorem 5 this sequence converges – strongly (see [9]) – to $\mathbf{x}^\star$ which is a minimiser of Φ and $\tilde{\Phi}$.

With L the operator defined by $L = \sqrt{CId - A^*A}$, we have

$$\Phi^{sur}(\mathbf{x}^{(n)}, \tilde{\mathbf{x}}^{(m)}) = \Phi(\mathbf{x}^{(n)}) + \|L(\mathbf{x}^{(n)} - \tilde{\mathbf{x}}^{(m)})\|_2^2 \ .$$

Moreover,

$$\lim_{n\to\infty, m\to\infty} \Phi^{sur}(\mathbf{x}^{(n)}, \tilde{\mathbf{x}}^{(m)}) = \Phi^{sur}(\mathbf{x}^\star, \mathbf{x}^\star) = \Phi(\mathbf{x}^\star) \ .$$

Then, for any $\varepsilon > 0$, there exist N, M, such that for all $n > N$ and $m > M$,

$$\|L(\mathbf{x}^{(n)} - \tilde{\mathbf{x}}^{(m)})\|_2^2 = |\Phi^{sur}(\mathbf{x}^{(n)}, \tilde{\mathbf{x}}^{(m)}) - \Phi(\mathbf{x}^{(n)})| < \varepsilon \ .$$

Finally, for all $n > N$ and $m > M$, with μ a stricly positive lower bound for the spectrum of L^*L we have

$$\|\mathbf{x}^{(n)} - \mathbf{x}^\star\|_2^2 \leq \|\mathbf{x}^{(n)} - \tilde{\mathbf{x}}^{(m)}\|_2^2 + \|\tilde{\mathbf{x}}^{(m)} - \mathbf{x}^\star\|_2^2 < \frac{1}{\mu}\|L(\mathbf{x}^{(n)} - \tilde{\mathbf{x}}^{(m)})\|_2^2 + \epsilon < (\frac{1}{\mu} + 1)\varepsilon \ .$$

Therefore, we can conclude that the sequence of $\mathbf{x}^{(n)}$ converges strongly to the fixed point $\mathbf{x}^\star$. ∎

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

5. Two illustrations

We choose to limit ourselves to two illustrations for our algorithms, in the field of audio signal processing. 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 [11,10]. These technique may be similarly applied to image processing problems in a straightforward manner.

5.1. *Illustration of the thresholded Landweber algorithm*

Our declicking example is a "toy example" which allows us to show the consequences of using the mixed norms compared to the classical ℓ_1 norm. In this example we limit ourselves to the $\ell_{1,2}$ norm. This choice is justified in Remark 9 below.

We choose a 44.1 KHz sampled, 3 secondes long (2^{17} samples) trumpet signal. We add to this signal random clicks simulated by Dirac pulses with amplitudes ± 1 to obtain a Signal to Noise Ratio (SNR) equal

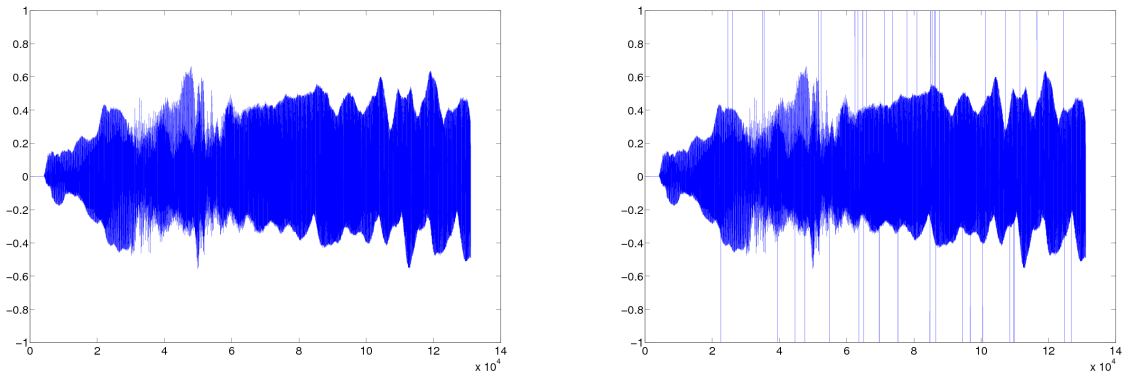

Fig. 1. Trumpet signal (left) and its clicked version (right).

to 20.33 dB. The time representation of the samples of the original signal and its clicked version are provided in Figure 1.

The signal is then decomposed in a Gabor frame with a 2048 sample long Gaussian window, with a time shift of 128 samples, and 2 samples in frequency. As it can be seen on the time-frequency representation in Figure 2, the clicks appear clearly as vertical lines that are sparse in time, but cover all the frequencies.

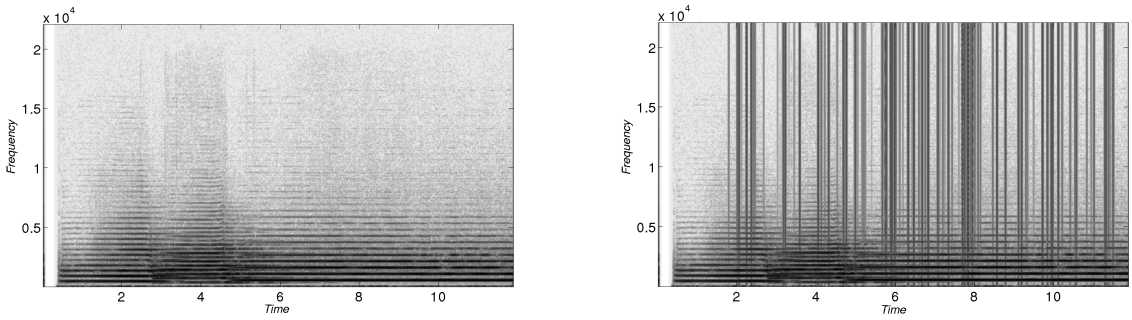

Fig. 2. Time-frequency representation of the Gabor coefficients for the original trumpet signal (left) and its clicked version (right).

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

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

$$\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\left(\sum_k \left(\sum_\ell |x_{k,\ell}|^p\right)^{q/p}\right) \quad , \tag{32}$$

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

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

The minimization of Φ was performed for various values of λ: the bigger the λ, the smaller the number of nonzero coefficients. 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 consequence of using 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. The time-frequency representations clearly show 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 partial harmonics are better preserved by the mixed norm.

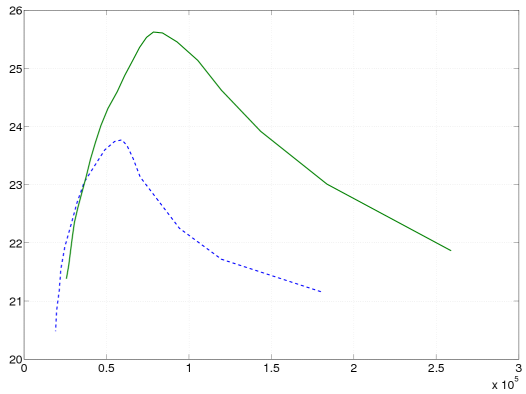

Fig. 3. Evolution of the SNR as a function of the number of coefficients. solid line: ℓ_{12} mixed norm, dashed line: ℓ_1 norm.

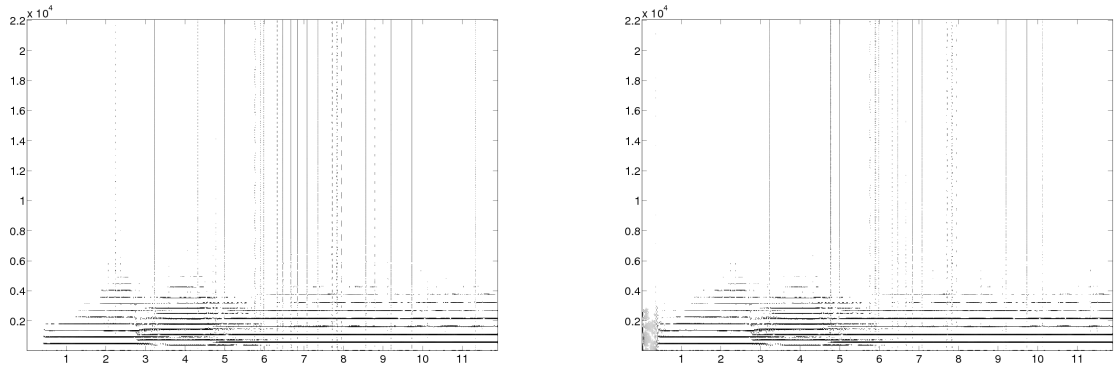

Fig. 4. Time-frequency representation of the Gabor coefficients for the denoised version. Left: ℓ_1 estimate. Right: $\ell_{1,2}$ estimate.

.

Remark 9 *The $\ell_{1,2}$ mixed norm appeared well adapted for the problem we chose if we look back to the estimate given in the orthogonal case in Proposition 4 in Subsection 4.1.3. For a fixed time index k, the threshold is equal to $\frac{\lambda}{1+\lambda L}|||[A^*\mathbf{y}]_k|||$. Thus, when a click appears at 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 than (32) *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 constraint. To this end, we choose to expand the signal in a dictionary constructed as the union of two MDCT (Modified Discrete Cosine Transform) bases. The first MDCT basis is chosen with a 4096 sample long window (about 90 msec) and is adapted for the tonal layer. The second one is chosen with a 128 sample long window and is adapted for the transient layer. The MDCT coefficients of transient layer are represented in Figure 5.

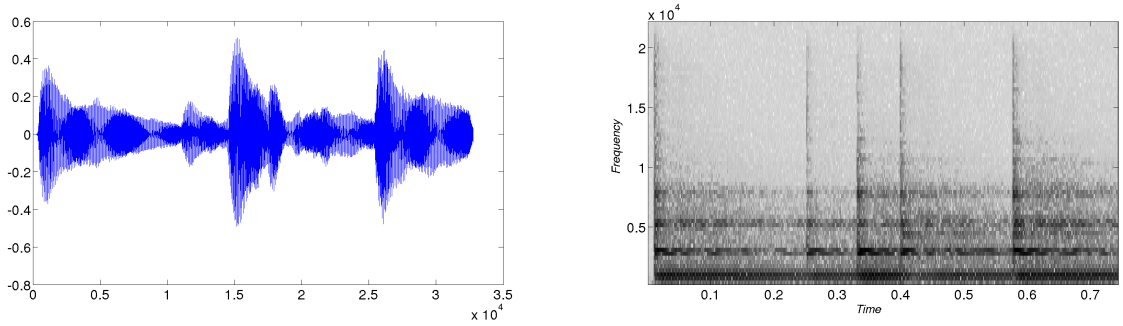

Fig. 5. Left: the xylophone signal. Right: MDCT coefficients of the signal, with a window of length 128 samples.

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 original FOCUSS algorithm.

Our adaptation of FOCUSS is used with two mixed norms. For the tonal layer the ℓ_{p_1,q_1} mixed norm is chosen to promote sparsity in frequency with $p_1 = 1.2$ and $q_1 = 2$ (and k is the time index and ℓ the frequency index). For the transient layer, with k the frequency index and ℓ the frequency one, 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 the penalty with the two mixed norms terms, we choose $\lambda_1 = 1$ and $\lambda_2 = 5$.

in Figure 6, we provide the MDCT coefficients of the transient layer estimated 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.

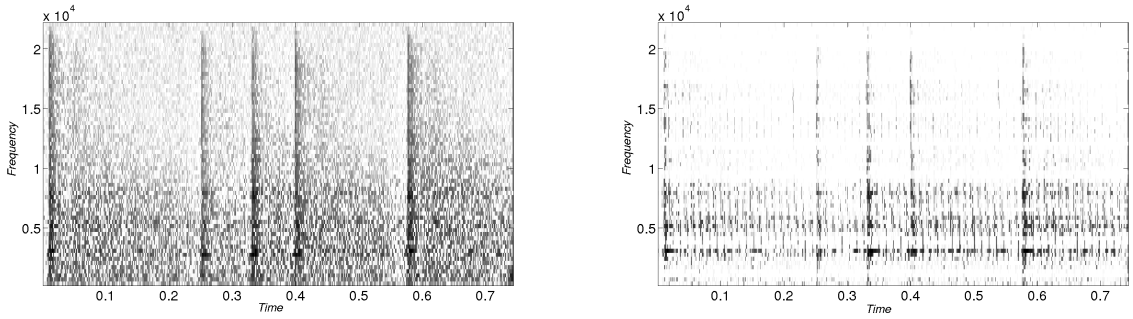

Fig. 6. MDCT coefficients of two estimates of the transient layer for the xylophone. Left: ℓ_1 estimate. Right: $\ell_{1,1/2}$ estimate.

6. Conclusion and outlooks

In this paper we showed that when the data can be labelled by a double index, mixed norms can help easily introduce sparsity structure. This indexing can indeed be used to introduced a hierarchy throught the coefficients. This hierarchy is then explicitly used in the model through mixed norms $\ell_{p,q}$. Then, one can play on both p and q to promote different structured sparsity.

Mixed norms $\ell_{p,q}$ are well adapted to two differents situations in regression:

– Signal estimation subject to an exact reconstruction;
– Noisy signal estimation.
Two algorithms were presented, corresponding to these different problems. For the sake of clarity, we summarize their strengths and weaknesses in Table 1.

	FOCUSS	Iterative thresholding
Range of 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.

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 [18] provides a valuable alternative to the thresholded Landweber iteration presented here. Our numerical experiments (not provided here) seem to show that both algorithms perform quite similarly.

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. The audio signals were chosen for the intuitive structures provided by their time-frequency representations. But let us stress that mixed norm are 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.

The $\ell_{p,1}$-like norms have already enjoyed significant succes in the statistical community for variable selection [33,29,34], and were more specifically studied and applied for color image restoration in [15] and [31]. Our work studied mixed norms in a general manner, and we want to stress the utility of the $\ell_{1,q}$-like norms to promote structures without imposing sparsity only on grouped variables (see remark 6 in section 4.1.3, and remark 9 in section 5). The simple example provided here encourages us to use mixed norms in signal restoration. The preceding paper [18] gave promising results in multichannel denoising and multilayered "tonal + transients + noise" decomposition. In [19], the thresholded Landweber iteration with the $\ell_{1,2}$ norms was applied with success to source separation of underdetermined convolutive mixtures.

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 [35] which uses a sum of ℓ_1 and ℓ_2 penalty, or the regularization penalty proposed in [15,14]) for the regularization term in the context of regression. Furthermore, it could be interesting to use mixed norms 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 penalise noise which is known not to be Gaussian.

Acknowledgement

The author wishes to gratefully thank Bruno Torrésani, Sandrine Anthoine and Liva Ralaivola for their precious advices and help during this work.

Appendix A. Proofs

A.1. *Proof of Proposition 2*

From (25), we have, for all k, ℓ such that $\mathbf{x}_k \neq \mathbf{0}$

$$|x_{k,\ell}| = \left(|\tilde{y}_{k,\ell}| - \lambda w_k |x_{k,\ell}| \|\mathbf{x}_k\|_{\mathbf{w}_k;2}^{-1}\right)^+ .$$

Then for all ν, $\lambda w_k \|\mathbf{x}_k\|_{\mathbf{w}_k;2}^{-1} = \frac{|\tilde{y}_{k,\nu}| - |x_{k,\nu}|}{|x_{k,\nu}|}$, which gives

$$|x_{k,\ell}| = \frac{|\tilde{y}_{k,\ell}|}{1 + \frac{|\tilde{y}_{k,\nu}| - |x_{k,\nu}|}{|x_{k,\nu}|}} = \frac{|\tilde{y}_{k,\ell}|}{|\tilde{y}_{k,\nu}|} |x_{k,\nu}| \ \forall l \ .$$

Then, we obtain

$$\begin{aligned}
|x_{k,\ell}| &= \left(|\tilde{y}_{k,\ell}| - \frac{\lambda |x_{k,\ell}| w_k}{\sqrt{\sum_\nu w_k |x_{k,\nu}|^2}}\right)^+ \\
|x_{k,\ell}| &= \left(|\tilde{y}_{k,\ell}| - \frac{\lambda |x_{k,\ell}| w_k}{\sqrt{\sum_\nu \left(w_k |\tilde{y}_{k,\nu}|^2 \frac{|x_{k,\ell}|^2}{|\tilde{y}_{k,\ell}|)^2}\right)}}\right)^+ \\
&= \left(|\tilde{y}_{k,\ell}| - \frac{\lambda |x_{k,\ell}| w_k}{\frac{|x_{k,\ell}|}{|\tilde{y}_{k,\ell}|} \sqrt{w_k} \|\tilde{\mathbf{y}}_k\|_2}\right)^+ \\
&= |\tilde{y}_{k,\ell}| \left(1 - \frac{\sqrt{w_k}\lambda}{\|\tilde{\mathbf{y}}_k\|_2}\right)^+ ,
\end{aligned}$$

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

$$|\mathbf{x}_k^{(m+1)}| = s(|\mathbf{x}_k^{(m)}|) = \left(|\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} |x_{k,\ell}^{(m)}|^{p-1} \|\mathbf{x}_k^{(m)}\|_{\mathbf{w}_k;p}^{1-p}\right)^+ . \tag{A.1}$$

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

We have

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

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

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 $L_{\mathbf{w}_k} = \sum_{\ell=1}^L w_{k,\ell}$, and as we have $1 < p < 2$, we can give an upper bound for the ℓ_1 norm:

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

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

A.3. *Proof of Proposition 4*

For all k, ℓ, we have from (25)

$$|x_{k,\ell}| = \left(|\tilde{y}_{k,\ell}| - \lambda\ w_{k,\nu}\|\mathbf{x}_k\|_{\mathbf{w}_k;1}^{q-1} \right)^+ ,$$

Then, $\forall k, \ell$, $\lambda\|\mathbf{x}_k\|_{\mathbf{w}_k;1}^{q-1} = \frac{|\tilde{y}_{k,\ell}| - |x_{k,\ell}|}{w_{k,\ell}}$, which gives, for all k, ν, ℓ such that $x_{k,\nu} \neq 0$ and $x_{k,\ell} \neq 0$

$$|x_{k,\nu}| = |\tilde{y}_{k,\nu}| - \frac{w_{k,\nu}(|\tilde{y}_{k,\ell}| - |x_{k,\ell}|)}{w_{k,\ell}} .$$

With $L_{\mathbf{w}_k} = \sum_{\nu:|x_{k,\nu}|\neq 0} w_{k,\nu}^2$, and $\|\tilde{\mathbf{y}}_k\|_{\mathbf{w}_k} = \sum_{\nu:|x_{k,\nu}|\neq 0} w_{k,\nu}|y_{k,\nu}|$, we have

$$
\begin{aligned}
|x_{k,\ell}| &= \left(|\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} \left(\sum_{\nu} w_{k,\nu} |x_{k,\nu}| \right) \right)^{q-1}_{+} \\
&= \left(|\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} \left(\sum_{\nu:|x_{k,\nu}|\neq 0} w_{k,\nu} \left[|\tilde{y}_{k,\nu}| - \frac{w_{k,\nu}(|\tilde{y}_{k,\ell}| - |x_{k,\ell}|)}{w_{k,\ell}} \right] \right)^{q-1}_{+} \right)^{+} \\
&= \left(|\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} \left(\left[\sum_{\nu:|x_{k,\nu}|\neq 0} w_{k,\nu} |\tilde{y}_{k,\nu}| \right] - L_{\mathbf{w}_k} \frac{|\tilde{y}_{k,\ell}| - |x_{k,\ell}|}{w_{k,\ell}} \right)^{q-1}_{+} \right)^{+} \\
&= \left(|\tilde{y}_{k,\ell}| - \lambda w_{k,\ell} \left(\|\tilde{y}_k\|_{\mathbf{w}_k} - L_{\mathbf{w}_k} \frac{|\tilde{y}_{k,\ell}| - |x_{k,\ell}|}{w_{k,\ell}} \right)^{q-1}_{+} \right)^{+} , \qquad \text{(A.2)}
\end{aligned}
$$

Denoting by $\xi_k = \lambda \|\mathbf{x}_k\|_{\mathbf{w}_k;1}$, we must now determine the set $\{\nu : |x_{k,\nu}| \neq 0\} = \{\nu : |y_{k,\nu}| > w_{k,\ell}\xi_k\}$. To do so, for each k, we sort the $|y_{k,\ell_k}|$ (resp. w_{k,ℓ_k}) such that the $r_{k,\ell_k} = \frac{y_{k,\ell_k}}{w_{k,\ell_k}}$ are ordered by decreasing order. We denote the ordered coefficients by $\check{y}_{k,\ell_k}$ (resp. $\check{w}_{k,\ell_k}$).

We must have $\frac{\check{y}_{k,L_k+1}}{w_{k,L_k+1}} \leq \xi_k < \frac{\check{y}_{k,L_k}}{w_{k,L_k}}$, then, from (25) L_k is such that

$$
r_{k,L_k+1} \leq \lambda \left(\sum_{\ell_k=1}^{L_k+1} w_{k,\ell_k}^2 (r_{k,\ell_k} - \xi_k) \right)^{q-1} \quad \text{and} \quad r_{k,L_k} > \lambda \left(\sum_{\ell_k=1}^{L_k} w_{k,\ell_k}^q (r_{k,\ell_k} - \xi_k) \right)^{q-1} .
$$

And finally, L_k is such that

$$
r_{k,L_k+1} \leq \lambda \left(\sum_{\ell_k=1}^{L_k+1} w_{k,\ell_k}^2 (r_{k,\ell_k} - r_{k,L_k+1}) \right)^{q-1} \quad \text{and} \quad r_{k,L_k} > \lambda \left(\sum_{\ell_k=1}^{L_k} w_{k,\ell_k}^2 (r_{k,\ell_k} - r_{k,L_k}) \right)^{q-1} .
$$

Then, $\|\tilde{\mathbf{y}}_k\|_{\mathbf{w}_k} = \sum_{\ell_k=1}^{L_k} \check{w}_{k,\ell_k} \check{y}_{k,\ell_k}$, and the thresholds ξ_k can be found by solving in $\mathbb{R}_+$ the equation

$$
\xi_k^{\frac{1}{q-1}} + \lambda^{\frac{1}{q-1}} L_{\mathbf{w}_k} \xi_k = \lambda^{\frac{1}{q-1}} \|\tilde{\mathbf{y}}_k\|_{\mathbf{w}_k} .
$$

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

$$
|x_{k,\ell}| = (|\tilde{y}_{k,\ell}| - w_{k,\ell}\xi_k)^+ .
$$

In the case $q = 2$, $w_{k,\ell}\xi_k = \frac{\lambda w_{k,\ell}}{1+L_{\mathbf{w}_k}\lambda} \|\tilde{\mathbf{y}}_k\|_{\mathbf{w}_k}$.

A.4. *Proof of Theorem 4*

We need to prove the following Lemma in order to be able to prove that the sequence $\mathbf{x}^{(m)}$ converges weakly to a fixed point, following [9].

Lemma 1 *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. The sequence formed by $\mathbf{w}$ is bounded from below by a strictly positive number, so we have $w_{k,\ell} \geq c$, uniformly in (k,ℓ), with $c > 0$.

We can write

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

because $\Phi(\mathbf{x}^{(m)})$ is a non-increasing sequence ([9, Lemma 3.5]), the $\Psi(\mathbf{x}^{(m)})$ are then uniformly bounded.

So, for all i

$$\|\mathbf{x}^{[i]^{(m)}}\|_{\mathbf{w},p_i,q_i}^{q_i} \leq \Phi(\mathbf{x}^{(0)}) \ , \tag{A.3}$$

and then, for all k,

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

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

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

Furthermore, we can show that

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

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

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

which, with Equations (A.3) and (A.4), allow us to give an uniform upper bound for $\|\mathbf{x}^{(m)}\|^2$. ∎

We also need to prove that the obtained fixed point is a minimizer of Functional (27). To do so, we denote by $\mathbf{x}^\star$ a fixed point of 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 $\partial\Phi^{sur}(\mathbf{x}, \mathbf{a})$:

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

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 concludes 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] A. Chambolle, D. R. A., N.-Y. Lee, B. J. Lucier, Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage, IEEE Transaction on Image Processing 7 (1998) 320–353.

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

[6] 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.

[7] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Modeling and Simulation 4 (4) (2005) 1168–1200.

[8] S. Cotter, B. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, IEEE Transactions on Signal Processing 53 (7) (2005) 2477–2488.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] M. Fornasier, H. Rauhut, Iterative thresholding algorithms, Applied and Computational Harmonic Analysis 25 (2) (2008) 187–208.

[15] M. Fornasier, H. Rauhut, Recovery algorithm for vector-valued data with joint sparsity constraints, SIAM Journal on Numerical Analysis 46 (2) (2008) 577–613.

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

[17] M. Kowalski, B. Torrésani, Random models for sparse signals expansion on unions of basis with application to audio signals, IEEE Transaction On Signal Processing 56 (8) (2008) 3468–3481.

[18] M. Kowalski, B. Torrésani, Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients, Signal, Image and Video Processing doi:10.1007/s11760-008-0076-1.

[19] M. Kowalski, E. Vicent, R. Gribonval, Under-determined source separation via mixed-norm regularized minimization, Proceedings of the European Signal Processing Conference.

[20] S. Kurcyusz, On the existence and nonexistence of lagrange multipliers in banach spaces, Journal of Optimization Theory and Applications 20 (1) (1976) 81–110.

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

[22] J.-J. Moreau, Proximit et dualit dans un espace hilbertien, Bull. Soc. Math. France 93 (1965) 273–299.

[23] 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.

[24] B. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection, IEEE Transaction On Signal Processing 47 (1) (1999) 187–200.

[25] 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.

[26] S. Samarah, S. Obeidat, R. Salman, A Shur test for weighted mixed-norm spaces, Analysis Mathematica 31 (2005) 277–289.

[27] S. Samarah, R. Salman, Local Fourier bases and modulation spaces, Turkish Journal of Mathematics 30 (4) (2006) 447–462.

[28] 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.

[29] 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.

[30] G. Teschke, Multi-frames representations in linear inverse problems with mixed multi-constraints, Applied and Computational Harmonic Analysis 22 (1) (2006) 43–60.

[31] 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, Inverse Problems 23 (2007) 1851–1870.
[32] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Serie B 58 (1) (1996) 267–288.
[33] 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.
[34] P. Zhao, G. Rocha, B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection, The Annals of Statistics (To appear).
[35] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of Royal Statistical Society Serie B 67 (2) (2005) 301–320.