Improved iterative shrinkage-thresholding for sparse signal recovery via Laplace mixtures models

Ravazzi and Magli *EURASIP Journal on Advances in Signal Processing* (2018) 2018:46
https://doi.org/10.1186/s13634-018-0565-5

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Improved iterative shrinkage-thresholding for sparse signal recovery via Laplace mixtures models

Chiara Ravazzi[1*] and Enrico Magli[1,2]

Abstract

In this paper, we propose a new method for support detection and estimation of sparse and approximately sparse signals from compressed measurements. Using a double Laplace mixture model as the parametric representation of the signal coefficients, the problem is formulated as a weighted ℓ_1 minimization. Then, we introduce a new family of iterative shrinkage-thresholding algorithms based on double Laplace mixture models. They preserve the computational simplicity of classical ones and improve iterative estimation by incorporating soft support detection. In particular, at each iteration, by learning the components that are likely to be nonzero from the current MAP signal estimate, the shrinkage-thresholding step is adaptively tuned and optimized. Unlike other adaptive methods, we are able to prove, under suitable conditions, the convergence of the proposed methods to a local minimum of the weighted ℓ_1 minimization. Moreover, we also provide an upper bound on the reconstruction error. Finally, we show through numerical experiments that the proposed methods outperform classical shrinkage-thresholding in terms of rate of convergence, accuracy, and of sparsity-undersampling trade-off.

Keywords: Compressed sensing, Sparse recovery, Gaussian mixture models, MAP estimation, Mixture models, Reweighted ℓ_1 minimization

1 Introduction

In this paper, we consider the standard compressed sensing (CS) setting [1], where we are interested in recovering high-dimensional signals $x^\star \in \mathbb{R}^n$ from few linear measurements $y = Ax^\star + \eta$, where $A \in \mathbb{R}^{m\times n}$, $m \ll n$, and η is a Gaussian i.i.d. noise. The problem is underdetermined and has infinitely many solutions. However, much interest has been focused on finding the sparsest solution, i.e., the one with the smallest number of nonzero components [2]. This involves the minimization of ℓ_0 pseudonorm [3], which is NP-hard.

A practical alternative is to use the ℓ_1 regularization, leading to the basis pursuit (BP, [4]) problem in the absence of noise, or the least absolute shrinkage and selection operator (Lasso, [5]) in the presence of noise. They can be efficiently solved by iterative shrinkage-thresholding algorithms (ISTA, [6–8]) that are generally first-order methods followed by a shrinkage-thresholding step. Due to its implementation simplicity and suitability for high-dimensional problems, a large effort has been spent to improve their speed of convergence [9–12], asymptotic performance in the large system limit [13, 14], and ease of use [15].

In a Bayesian framework, ℓ_1 minimization is equivalent to a maximum a posteriori (MAP) estimate [16] modeling the signal coefficients using a Laplace prior, in the sense that we need to solve the same optimization problem. Although the Laplace probability density function does not provide a relevant generative model for sparse or compressible signals [17], the non-differentiability at zero of the cost function leads to select a sparse solution, providing empirical success of ℓ_1 regularization.

However, ℓ_1 minimization alone does not fully exploit signal sparsity. In fact, in some cases, a support estimate [18] could be employed to reduce the number of measurements needed for good reconstruction via BP or Lasso, e.g., by combining support detection with weighted or

*Correspondence: chiara.ravazzi@ieiit.cnr.it
[1]National Research Council of Italy, IEIIT-CNR, c/o Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
Full list of author information is available at the end of the article

truncated ℓ_1 minimization [19]. The idea of combining support information and signal estimation has appeared in CS literature with several assumptions [20–27]. For example, in [25], the authors employ as prior information an estimate T of the support of the signal and propose a truncated ℓ_1 minimization problem. Another piece of literature [28] considers a weighted ℓ_1 minimization with weights $w_i = -\log p_i$ where p_i is the probability that $x_i^\star = 0$.

In this paper, we propose an iterative soft support detection and estimation method for CS. It is worth remarking that in our setting prior information on the support T or p_i is not available. The fundamental idea is to combine the *good geometric properties of the* ℓ_1 *cost function* associated to the Laplacian prior with a *good generative model* for sparse and compressible vectors [17]. For this purpose, we use a Laplace mixture model as the parametric representation of the prior distribution of the signal coefficients. Because of the partial symmetry of the signal sparsity, we know that each coefficient should have one out of only two distributions: a Laplace with small variance with high probability and a Laplace with large variance with low probability. We show empirically that this model fits better with the distribution of the Haar wavelet coefficients in test images. Then, we cast the estimation problem as a weighted ℓ_1 minimization method that incorporates the parametric representation of the signal.

We show that the proposed framework is able to improve a number of existing methods based on shrinkage-thresholding: by estimating the distribution of the components that are likely to be nonzero from signal estimates at each iteration (support detection), the shrinkage-thresholding step is tuned and optimized, thereby yielding better estimation. As opposed to other adaptive methods [10], we are able to prove, under suitable conditions, the convergence of the proposed tuned method. Moreover, we derive an upper bound on the reconstruction error. We apply this method to several algorithms, showing by numerical simulation that it improves recovery in terms of both speed of convergence and sparsity-undersampling trade-off, while preserving the implementation simplicity.

Compared to the literature on reconstruction methods that combine iterative support detection and weighted ℓ_1 minimization, the identification of the support is not nested or incremental over time as in [29–31]. Moreover, the choice of weights in ℓ_1 minimization is based on the Bayesian rules and a probabilistic model and not on greedy rules as in [19, 32]. This feature also marks the difference with respect to reweighted ℓ_1/ℓ_2 minimization, where the weights are chosen with the aim of approximating the ℓ_τ norm with $\tau \in (0, 1]$.

1.1 Outline

The paper is organized as follows. In Section 2, the basic CS theory and the classical methods based on ℓ_1 minimization for sparse recovery are reviewed. The proposed parametric model for sparse or highly compressible signals is described in Section 3 and compared with the related literature. In Section 4, the estimation problem based on Laplace mixture models is introduced and recast as a weighted ℓ_1 minimization problem. Then, in Section 5, the proposed approach is used to improve a number of existing methods based on shrinkage-thresholding. Numerical experiments are presented in Section 6 and some concluding remarks (Section 7) complete the paper. The theoretical results are rigorously proved in Appendices 1, 2, 3, and 4.

1.2 Notation

We conclude this introduction with some notation. We denote column vectors with small letters and matrices with capital letters. If $x \in \mathbb{R}^n$, we denote its jth element with x_j and, given $S \in [n] := \{1, \dots, n\}$, by $x|_S$ the subvector of x corresponding to the indexes in S. The support set of x is defined by $\mathrm{supp}(x) = \{i \in [n] : x_i \neq 0\}$, and we use $\|x\|_0 = |\mathrm{supp}(x)|$. Finally, the symbol $\|x\|$ with no subscript is to be understood as the Euclidean norm of the vector x. We denote as $r(x)$ the nonincreasing rearrangement of x

$$r(x) = \left(|x_{i_1}|, |x_{i_2}|, \dots, |x_{i_n}|\right)^\top,$$

where

$$|x_{i_\ell}| \geq |x_{i_{\ell+1}}|, \ \forall \ell = 1, \dots, n-1.$$

We denote with $\Sigma_s = \{x \in \mathbb{R}^n : |\mathrm{supp}(x)| \leq s\}$ and define

$$\sigma_s(x) = \arg\min_{z \in \Sigma_s} \|x - z\|.$$

It should be checked that

$$\sigma_s(x)_i = \begin{cases} x_i & \text{if } |x_i| > r(x)_{s+1} \\ 0 & \text{otherwise.} \end{cases}$$

Given a matrix A, A^T denotes its transpose.

2 Mathematical formulation

2.1 Sparse signal recovery from compressed measurements

Compressive sensing aims to recover a sparse signal $x^\star \in \mathbb{R}^n$ from $m \leq n$ random projections of the form

$$y = Ax^\star + \eta \tag{1}$$

where $y \in \mathbb{R}^m$ is the observation vector, $A \in \mathbb{R}^{m \times n}$ is the measurement matrix, and η is an additive noise. For example, in the transform domain compressive sig-

nal reconstruction [1], $A = \Phi\Psi$, where $\Psi \in \mathbb{R}^{n\times n}$ is the sparsifying basis (i.e., multiplying by Ψ corresponds to performing inverse transform), the entries of $x^\star$ is the transform coefficient vector that has k nonzero entries, and $\Phi \in \mathbb{R}^{m\times n}$ is the sensing matrix, whose rows are incoherent with the columns of Ψ.

Conventional reconstruction methods involve ℓ_1 regularization [4]. In particular, it has been shown that, in the absence of noise and under suitable assumptions on the matrix A, the basis pursuit problem (BP)

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \qquad \text{s.t.} \quad Ax = y \tag{2}$$

can exactly recover a k-sparse signal (i.e., with a number of nonzero coefficients not larger than k) from $m = O(k\log(n/k))$ measurements with high probability [33, 34]. In the presence of noise, one of the most popular convex relaxation methods is the least absolute shrinkage and selection operator (Lasso, [5]), which requires to solve the following unconstrained problem

$$\min_{x\in\mathbb{R}^n} \left[\lambda\|x\|_1 + \frac{1}{2}\|Ax-y\|_2^2\right] \tag{3}$$

where λ is a positive regularization parameter.

Even when the vector is not exactly sparse, under *compressibility* assumptions of the signals to be recovered, the ℓ_1 regularization provides estimates with a controlled error [33]. More formally, we recall the following definition [17].

Definition 1 (Compressible vectors) *A vector $x \in \mathbb{R}^n$ is compressible if, denoted $\varrho_k(x) := \inf_{\|z\|_0\le k}\|z-x\|$, the relative best k-term approximation error is $\overline{\varrho}_k(x) := \varrho_k(x)/\|x\| \ll 1$ for some $k \ll n$.*

If $x \in \mathbb{R}^n$ is not exactly sparse but compressible, the support is intended as the set of significant components $\mathrm{supp}(\varrho_k(x))$.

One drawback of classical ℓ_1 minimization is that it fails to penalize the coefficients in different ways. In this paper, we propose a new family of methods that incorporate two tasks: iterative support detection and signal recovery.

3 Discussion: learning sparsity models

3.1 A Bayesian view

In a Bayesian framework, if the noise in (1) is white Gaussian (we suppose for simplicity unitary standard deviation), BP and Lasso may be interpreted as a Bayesian MAP estimate [16]. In fact, imposing the ℓ_1 norm as penalty in the cost function is equivalent to modeling the signal coefficients $x_i^\star$ as independent and identically distributed as a Laplace distribution, namely

$$\widehat{x}^*(y) = \arg\max_{x\in\mathbb{R}^n} \log f^*_{x|y}(x|y)$$

where, using the Bayes rule, $f_{x|y}(x|y)$ is given by

$$f^{\mathrm{BP}}_{x|y}(x|y) = \frac{1}{Z}\prod_{i=1}^n \exp(-\lambda|x_i|)\prod_{j=1}^m \delta_{y_j=(Ax)_j},$$

in case of BP, and for Lasso by

$$f^{\mathrm{Lasso}}_{x|y}(x|y) = \frac{1}{Z}\prod_{i=1}^n \exp(-\lambda|x_i|) \times \prod_{j=1}^m \exp\left(-\frac{1}{2}\left(y_j-(Ax)_j\right)^2\right),$$

and Z is a normalization factor so that $\int f^*_{x|y}(x|y)\mathrm{d}x = 1$.

Despite the good geometric properties of the ℓ_1 cost function associated to such prior, that allow to select a sparse solution, the Laplace prior does not provide a relevant generative model for sparse or compressible signals [17]. In fact, if $x_n \in \mathbb{R}^n$ is distributed as i.i.d. with respect to Laplace distribution with scale parameter λ, then for any sequence k_n such that $\lim_{n\to\infty} k_n/n = \kappa \in [0,1]$, it holds almost surely

$$\varepsilon := \lim_{n\to\infty} \overline{\varrho}_{k_n}(x_n)^2 \overset{a.s.}{=} 1-\kappa\left(1+\log 1/\kappa + \frac{1}{2}(\log 1/\kappa)^2\right).$$

Therefore, the vectors generated from i.i.d. Laplace distribution are not compressible since we cannot have both κ and ε small at the same time.

Moreover, for a large class of real signals that have highly non-Gaussian statistics, the Laplace model does not provide a good fit to the empirical probability density function. We show this with a simple experiment (experiment 1) Fig. 1. We calculate a single vertical wavelet subband coefficients of several real images of size 256×256 pixels, and we compute for each image the best fitting of the double Laplace density function

Fig. 1 Test images for experiments 1 and 2: Lena, MRI-head, house, cameraman, pattern, Barbara, man, couple, plane (from left to right)

obtained by maximizing the likelihood of the data under that assumption. In Fig. 2, the empirical density function of the Haar wavelet coefficients using 256 bins is shown in the log domain (solid line) and the dashed line corresponds to the best fitting instance of the Laplace density function. It should be noticed that the Laplace density captures peaks at zero but is less accurate along the tails.

In [35], Lasso is proved to provide a robust estimation that is invariant to the signal prior. In sharp contrast, the Bayesian Lasso is able to provide an estimator with minimum mean squared error by incorporating the signal model in the estimation problem [14, 36], but the assumption that the signal prior is known in advance is not reasonable in most practical cases. Hence, it becomes crucial to incorporate in the recovery procedure new tools for adaptively learning sparsity models. Other models have been proposed for compressible signals [37–39] using more accurate probability density functions than double Laplace distribution. However, two issues generally appear when an accurate but complex signal prior is used: (1) it can be hard to estimate the model parameters and (2) the optimal estimators may not have simple closed form solution and their computation may require high computational work [40]. In fact, although the double Laplace prior is not the most accurate model, this is an especially convenient assumption since the MAP estimator has a simple and closed form [41].

Our goal is to use a compressible distribution as parametric representation of the signal coefficients, able to combine support detection and estimation, and to

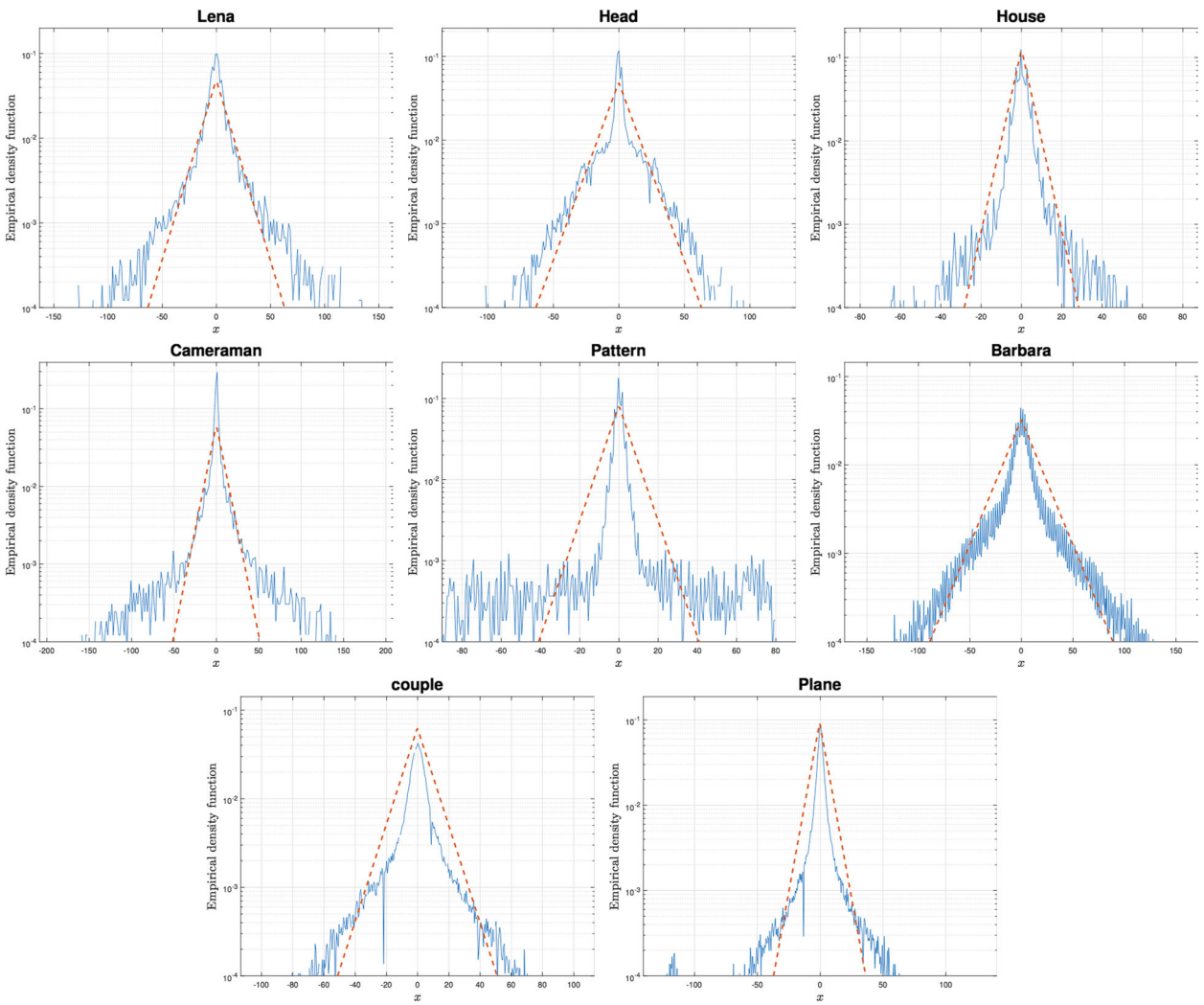

Fig. 2 Experiment 1: empirical density function in the log domain of a single vertical wavelet subband coefficients of several real images (solid) and the best fitting of the Laplace density function (dashed) obtained by maximizing the likelihood of the data

preserve the simplicity and advantages coming from the Laplace prior assumption.

3.2 Proposed approach: two-component Laplace mixture for support detection

We consider a two-state mixture model as a prior that describes our knowledge about the sparsity of the solution to (1). Because of the partial symmetry of the signal sparsity, we consider the case in which x is a random variable with components of the form

$$x_i = z_i u_i + (1 - z_i)v_i \quad i \in [n]$$

where u_i are identically and independently distributed (i.i.d.) as $\mathsf{Laplace}(0,\alpha)$, v_i are i.i.d. according to $\mathsf{Laplace}(0,\beta)$ and z_i are i.i.d. Bernoulli random variables with probability mass function $f(z_i = 1) = 1 - p$, with $p \ll 1/2$, $\alpha \approx 0$, and $\beta \gg 0$, in order to ensure that we have few large coefficients. We thus consider the conditional distribution of the data: let $\Theta = (\alpha, \beta)$

$$f(x|y;\Theta) = \frac{1}{Z}\prod_{i=1}^{n}\left[(1-p)f(x_i|z_i=1) + pf(x_i|z_i=0)\right] \times \prod_{j=1}^{m} f_j(y|x), \tag{4}$$

where

$$f(x_i|z_i=1) = \frac{1}{2\alpha}\exp\left(-\frac{|x_i|}{\alpha}\right) \tag{5a}$$

$$f(x_i|z_i=0) = \frac{1}{2\beta}\exp\left(-\frac{|x_i|}{\beta}\right), \tag{5b}$$

$$f_j(y|x) = \delta_{\{y_j=(Ax)_j\}}$$

in absence of noise or

$$f_j(y|x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y_j - (Ax)_j\right)^2\right)$$

in presence of noise, and Z is a normalization factor so that $\int f(x|y;\Theta)\mathrm{d}x = 1$.

This mixture model is completely described by three parameters: the sparsity ratio $p \ll 1/2$, α that is expected to be small, and $\beta > \alpha$ if the signal is sparse. It should be noticed that vectors generated from this distribution are typically compressible according to Definition 1.

Proposition 1 *Let $x_n \in \mathbb{R}^n$ be i.i.d. with respect to (4). Then, for any sequence k_n such that $\lim_{n\to\infty} k_n/n = \kappa \in [0,1]$, it holds almost surely*

$$\varepsilon := \lim_{n\to\infty} \overline{\varrho}_{k_n}(x_n)^2 \overset{a.s.}{=} \frac{(1-p)\left(\alpha^2 - \mathrm{e}^{-t/\alpha}\left(t^2/2+\alpha t+\alpha^2\right)\right) + p\left(\beta^2 - \mathrm{e}^{-t/\beta}\left(t^2/2+\beta t+\beta^2\right)\right)}{(1-p)\alpha^2 + p\beta^2}$$

where t is the unique solution of

$$(1-p)\mathrm{e}^{-t/\alpha} + p\mathrm{e}^{-t/\beta} = \kappa.$$

The proof is a consequence of proposition 1 in [17] and is deferred to Appendix 1.

In Fig. 3 (experiment 2), we compare the compressibility parameters (κ, ε) of the Laplace distribution and of 2-LMM distribution with $\alpha = 0.1, \beta = 10$ for several values of p. It should be noted that Laplace distribution is not a compressible distribution (we can not have κ and ε small at the same time), whereas 2-LMM distribution are compressible if parameter p is sufficiently small, as we can have both κ and ε small at the same time.

We now compute the empirical probability density function of the Haar wavelet coefficients of several images and the best fitting of the mixture of two Laplace density functions computed by maximizing the likelihood of the data. The computation has been carried out via expectation maximization algorithm [16]. In Fig. 4, we show the results for several images. In order to compare the two parametric representations of sparsity, in Table 1, the Kullback-Leibler divergence of the best fitting probability models and the empirical probability density function are computed for the two models. It can be noticed that a single Laplace is a poor model for the Haar

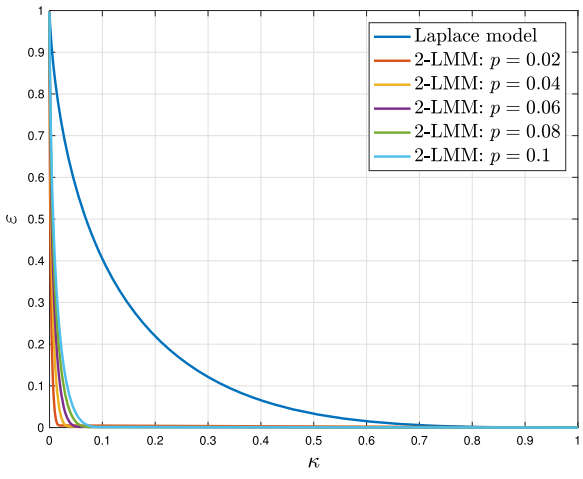

Fig. 3 Experiment 2: comparison of compressibility parameters (κ, ε) for several priors. Laplace distribution is not a compressible distribution, whereas for 2-LMM distribution if parameter p is small and $\alpha \ll \beta$ we can have both κ and ε small at the same time. In this case $\alpha = 0.1$, $\beta = 10$, and $p \in [0.02, 0.1]$

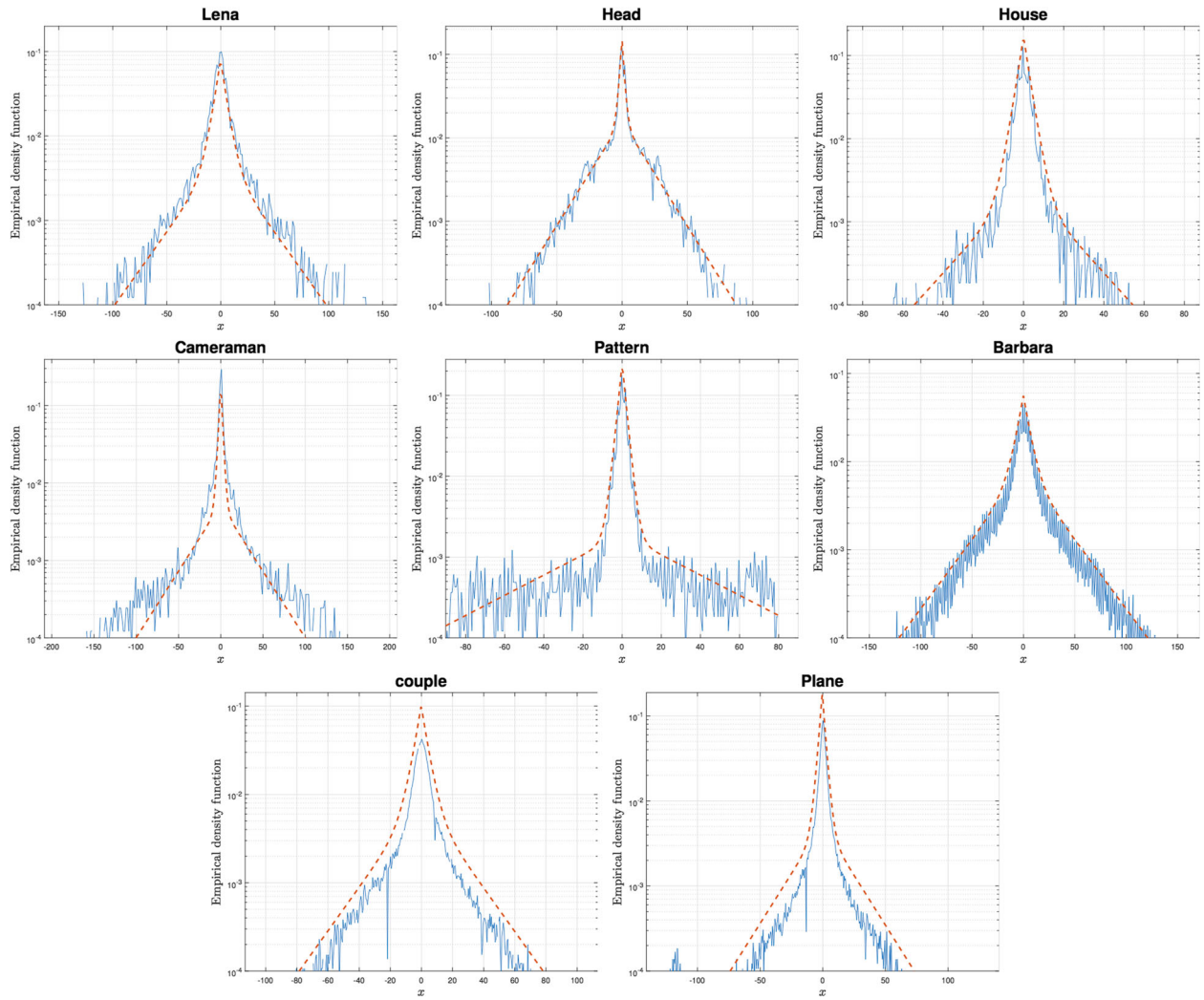

Fig. 4 Experiment 2: empirical density function in the log domain of a single vertical wavelet subband coefficients of natural images (solid). The dashed line represents the best fitting of the mixture of two Laplace density functions computed by maximizing the likelihood of the data

wavelet coefficients of natural images. The better accuracy obtained by the new parametric representation is evident.

4 Method: support detection and sparse signal estimation via the 2-LMM

4.1 Estimation using the 2-LMM generative model

Let us consider the logarithm of the conditional distribution in (4) in absence of noise (although similar consideration can be done in presence of noise):

$$L(x;\Theta) := \log\left[f(x|y;\Theta)\right] \tag{6}$$

For convenience, we consider p fixed as a guess of the degree of signal's sparsity, whereas $\Theta = (\alpha, \beta)$ will be unknown. The choice to keep p fixed does not entail a significant restriction to our analysis.

Proposition 2 *The following optimization problems are equivalent*

Table 1 The Kullback-Leibler divergence of the best fitting probability models and the empirical probability density function are computed for the two models

Image	Lena	MRI-head	House	Cameraman	Pattern	Barbara	Man	Couple	Plane
Lap	0.1266	0.2109	0.1593	0.4958	0.6756	0.1449	0.0941	0.1793	0.4044
2-LMM	0.0688	0.0339	0.0688	0.0435	0.0818	0.0651	0.0322	0.0558	0.0906

$$\max_{\Theta}\max_{x\in\mathbb{R}^n} L(x;\Theta) \tag{7}$$

$$\min_{\Theta}\min_{x\in\mathbb{R}^n}\min_{\pi\in[0,1]^n} \underbrace{J(x,\pi;\Theta)-\sum_{i=1}^{n} H(\pi_i)}_{V(x,\pi;\Theta)} \quad s.t. Ax=y \tag{8}$$

where

$$J(x,\pi;\Theta)=\sum_{i=1}^{n}\left[\frac{\pi_i|x_i|}{\alpha}+\pi_i\log\alpha-\pi_i\log(1-p)+\frac{(1-\pi_i)|x_i|}{\beta}+(1-\pi_i)\log\beta-(1-\pi_i)\log p\right], \tag{9}$$

and $H(t)=-t\log t-(1-t)\log(1-t)$ *is the natural entropy function with* $t\in[0,1]$.

Given y, A, instead of solving optimization problem in (7), we consider the minimization of the following modified cost function:

$$\min_{\Theta}\min_{x\in\mathbb{R}^n}\min_{\pi\in\Sigma_{n-K}} J_\epsilon(x,\pi;\Theta)-\sum_{i=1}^{n} H(\pi_i), \text{s.t.} Ax=y \tag{10}$$

where $K=\lfloor pn\rfloor$, and

$$\begin{aligned} J_\epsilon(x,\pi;\Theta) &:= J(x,\pi;\Theta)+\epsilon\left(\frac{1}{\alpha}+\frac{1}{\beta}\right) \\ &=\sum_{i=1}^{n}\left[\frac{\pi_i|x_i|+\epsilon/n}{\alpha}+\pi_i\log\alpha-\pi_i\log(1-p)\right. \\ &\quad+\frac{(1-\pi_i)|x_i|+\epsilon/n}{\beta}+(1-\pi_i)\log\beta \\ &\quad\left.-(1-\pi_i)\log p\right] \end{aligned} \tag{11}$$

Compared to (7), the optimization problem in (10)

1. Introduces the ϵ parameter, which is a regularization term used to avoid singularities whenever one of the Laplace components collapses onto a specific data point since we expect that $\alpha\approx 0$, since we seek a sparse solution; this fact will be clear later;
2. Introduces the constraint $\pi\in\Sigma_{n-K}$, which enforces a sparse solution.

Similar computations can be carried out for the case with noise, leading to

$$\min_{\Theta}\min_{x\in\mathbb{R}^n}\min_{\pi\in\Sigma_{n-K}} V(x,\pi,\alpha,\beta,\epsilon)$$

$$V(x,\pi,\alpha,\beta,\epsilon):=\frac{1}{2}\|y-Ax\|_2^2+\lambda J_\epsilon(x,\pi;\Theta)-\lambda\sum_{i=1}^{n} H(\pi_i) \tag{12}$$

It should be noted that there is not a closed form solution to problems (10) and (12).

However, partial minimizations of $V_\epsilon = V(x,\pi,\alpha,\beta,\epsilon)$ with respect to just one of the variables have simple representation. More precisely, we have the following expressions (see Lemma 3 in Appendix 3).

Proposition 3 *Let*

$$\begin{aligned} \widehat{\pi} &= \widehat{\pi}(x,\alpha,\beta,\epsilon)=\arg\min_{\xi\in\Sigma_{n-K}} V_\epsilon(x,\xi,\alpha,\beta) \\ \widehat{\alpha} &= \widehat{\alpha}(x,\alpha,\beta,\epsilon)=\arg\min_{\alpha\in\mathbb{R}} V_\epsilon(x,\xi,\alpha,\beta) \\ \widehat{\beta} &= \widehat{\beta}(x,\alpha,\beta,\epsilon)=\arg\min_{\beta\in\mathbb{R}} V_\epsilon(x,\xi,\alpha,\beta) \end{aligned}$$

then

$$\widehat{\pi}=\sigma_{n-K}\left(\frac{e^{-\frac{|x|}{\alpha}-\log(\alpha)+\log(1-p)}}{e^{-\frac{|x|}{\alpha}-\log(\alpha)+\log(1-p)}+e^{-\frac{|x|}{\beta}-\log(\beta)+\log p}}\right) \tag{13}$$

and

$$\widehat{\alpha}=\frac{\sum_{i=1}^{n}\pi_i|x_i|+\epsilon}{\sum_{j=1}^{n}\pi_j},\quad \widehat{\beta}=\frac{\sum_{i=1}^{n}(1-\pi_i)|x_i|+\epsilon}{\sum_{j=1}^{n}(1-\pi_j)}$$

In the following section, we present several iterative algorithms to approximately solve these optimization problems.

5 Proposed iterative methods and main results

5.1 Iterative shrinkage/thresholding algorithms

The literature describes a large number of approaches to address minimization of (2) and (3). Popular iterative methods belong to the class of iterative shrinkage-thresholding algorithms. These methods can be understood as a special proximal forward backward iterative scheme [42] and are appealing as they have lower computational complexity per iteration and lower storage requirements than interior-point methods. In fact, these types of recursions are a modification of the gradient method to solve a least square problem, where the dominant computational effort lies in a relatively cheap matrix-vector multiplication involving A and $A^\top$ and the only difference consists in the application of a shrinkage/soft thresholding operator, which promotes sparsity of the estimate at each iteration.

More precisely, let $\{\tau^{(t)}\}_{t\in\mathbb{N}}$ be a sequence in $(0,\infty)$ such that $\inf_{t\in\mathbb{N}}\tau^{(t)}>0$ and $\sup_{t\in\mathbb{N}}\tau^{(t)}<2\|A\|_2^{-2}$, and let $\{u^{(t)}\}_{t\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Then, for every $t\in\mathbb{N}$ let

$$x^{(t+1)}=\eta^{\mathsf{S}}_{\lambda\tau^{(t)}}\left[x^{(t)}+\tau^{(t)}A^{\mathsf{T}}\left(y-Ax^{(t)}\right)+u^{(t)}\right], \tag{14}$$

where η_γ^{S} is a thresholding function to be applied element-wise defined as

$$\eta_\gamma^{\mathsf{S}}[x] = \begin{cases} \operatorname{sgn}(x)(|x| - \gamma) & \text{if } |x| > \gamma \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The simplest form, known as iterative shrinkage-thresholding algorithms (ISTA, [6]), considers $u^{(t)} = 0$ and $\tau^{(t)} = \tau < 2\|A\|_2^{-2}$ for all $t \in \mathbb{N}$. This algorithm is guaranteed to converge to a minimizer of the Lasso [6]. Moreover, as shown in [43], if A fulfills the so-called finite basis injectivity condition, the convergence is linear. However, the factor determining the speed within the class of linearly convergent algorithms depends on local well conditioning of the matrix A, meaning that ISTA can converge arbitrarily slowly in some sense, which is also often observed in practice.

In order to speed up ISTA, alternative algorithms have exploited preconditioning techniques or adaptivity, combining a decreasing thresholding strategy with adaptive descent parameter. However, the lack of a model-based thresholding policy makes this algorithm very sensitive to the signal statistics and the accuracy is not always guaranteed. In [13], the thresholding and descent parameters are optimally tuned in terms of phase transitions, i.e., they maximize the number of nonzeros at which the algorithm can successfully operate. However, preconditioning can be very expensive and there is no proof of convergence for adaptive methods.

Finally, other variations update the next iterate using not only the previous estimation, but two or more previously computed iterates. Among all the proposed techniques with a significantly better rate of convergence and phase transitions, we recall (a) fast iterative shrinkage-thresholding algorithm (FISTA, [9]) obtained by (14) choosing $\tau^{(t)} = \tau < 2\|A\|_2^{-2}$ and

$$\begin{aligned} u^{(t)} &= \frac{\zeta^{(t-1)} - 1}{\zeta^{(t)}} \left(I - \tau A^{\mathsf{T}} A\right) \left(x^{(t)} - x^{(t-1)}\right) \\ \zeta^{(0)} &= 1, \quad \zeta^{(t+1)} = \frac{1 + \sqrt{1 + 4\left(\zeta^{(t)}\right)^2}}{2} \end{aligned} \tag{16}$$

and (b) approximate message passing (AMP, [14]) with threshold recursion proposed in [44]

$$\begin{aligned} u^{(t)} &= \left(1 - \tau^{(t)}\right) A^{\mathsf{T}} \left(A x^{(t)} - y\right) + \frac{\left\|x^{(t)}\right\|_0}{m} A^{\mathsf{T}} r^{(t-1)} \\ \tau^{(t)} &= \chi \frac{\left\|r^{(t)}\right\|_2}{\sqrt{m}}, \quad r^{(t)} = y - A x^{(t)} + r^{(t-1)}. \end{aligned} \tag{17}$$

In this section, we show how to adapt these numerical methods to solve the weighted minimization problem via 2-LMM.

5.2 2-LMM-tuned iterative shrinkage-thresholding

Let us consider the problem of minimizing (11). Since information about the locations of the nonzero coefficients of the original signal is not available a priori, the task of selecting the parameters α, β, and π is performed iteratively. We propose an alternating method for the minimization of (11), inspired by the EM algorithm [16]. The pseudocode of the algorithm is reported in Algorithm 1. The strategy can be summarized as follows.

1. Let $t := 0$ and set an initial estimate K for the sparsity level, $p = K/n$, a small value $\alpha^{(0)} \approx 0$ (e.g., $\alpha^{(0)} = 0.1$), the initial configuration $\pi^{(0)} = \mathbb{1}$, and $\epsilon^{(0)} = 1$. Since $\pi^{(0)} = \mathbf{1}$, $\beta^{(0)}$ can be arbitrary since it is not used in the first step of the algorithm.
2. Given the observed data y and current parameters π_i, α, andβ, a new estimation $x^{(t+1)}$ of the signal is obtained by moving in a minimizing direction of weighted Lasso

$$F(x) = \frac{1}{2}\|Ax - y\|^2 + \lambda \sum_{i=1}^{n} \omega_i^{(t+1)} |x_i| \tag{18}$$

   with $\omega_i^{(t+1)} = \pi_i/\alpha + (1 - \pi_i)/\beta$; in other terms $x^{(t+1)}$ is such that $F\left(x^{(t+1)}\right) \le F\left(x^{(t)}\right)$.
3. The posterior distribution of the signal coefficients is evaluated and thresholded by keeping its $n - K$ biggest elements and setting the others to zero. It is worth remarking that this step differs from the E-step of a classical EM algorithms as a thresholding operator σ_{n-K} is applied in order to promote the sparsity in the probability vector π.
4. Given the probabilities, we use them to re-estimate the mixture parameters $\alpha^{(t)}$ and $\beta^{(t)}$.
5. Set $t := t + 1$ and iterate until the stopping criteria is satisfied, e.g., until the estimate stops changing $\|x^{(t+1)} - x^{(t)}\|/\|x^{(t)}\| < \mathsf{tol}$ for some $\mathsf{tol} > 0$.

5.3 Relation to prior literature

As already observed, Algorithm 1 belongs to the more general class of methods for weighted ℓ_1 norm minimization [45–47] (see (18)). Common strategies for iterative reweighting ℓ_1 minimization (IRL1, [45]) that have been explored in literature re-compute weights at every iteration using the estimate at the previous iteration $\omega_i^{(t+1)} = \chi/(|x_i^{(t)}| + \epsilon)$ where χ and ϵ are appropriate positive constants. In Algorithm 1, the weights $\omega_i^{(t)}$ are chosen to jointly fit the signal prior and, consequently, depend on all components of the signal and not exclusively on the value $x_i^{(t)}$. Our strategy is also related to threshold-ISD [19] that incorporates support detection in the weighted ℓ_1 minimization and runs as fast as the basis

Algorithm 1 2-LMM-tuned iterative shrinkage-thresholding

Require: Data (y, A), set K, $p = K/n$

1: Initialization: $\alpha^{(0)} = \alpha_0$, $\pi^{(0)} = \mathbb{1}$, $\epsilon^{(0)} = 1$

2: **for** $t = 1, \dots,$ *StopIter* **do**

3: Computation of ℓ_1-weights:

$$\omega_i^{(t+1)} = \frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{1-\pi_i^{(t)}}{\beta^{(t)}}$$

4: Gradient/Thresholding step:

$$x^{(t+1)} = \eta^{\mathsf{S}}_{\lambda\tau^{(t)}\omega^{(t+1)}}\left(x^{(t)} + \tau^{(t)}A^\top\left(y - Ax^{(t)}\right) + u^{(t)}\right)$$

5: Posterior distribution evaluation:

$$a_i^{(t+1)} = \frac{\frac{(1-p)}{\alpha^{(t)}}\exp\left(-\frac{|x_i^{(t+1)}|}{\alpha^{(t)}}\right)}{\frac{(1-p)}{\alpha^{(t)}}\exp\left(-\frac{|x_i^{(t+1)}|}{\alpha^{(t)}}\right) + \frac{p}{\beta^{(t)}}\exp\left(-\frac{|x_i^{(t+1)}|}{\beta^{(t)}}\right)}$$

$$\pi^{(t+1)} = \sigma_{n-K}\left(a^{(t+1)}\right)$$

6: Regularization parameter:

$$\theta^{(t+1)} = \frac{1}{\log(t+1)} + c\left\|x^{(t+1)} - x^{(t)}\right\| + \frac{r\left(x^{(t+1)}\right)_{K+1}}{n}$$

$$\epsilon^{(t+1)} = \min\left(\epsilon^{(t)}, \theta^{(t+1)}\right)$$

where $r(x)$ is the non increasing rearrangement of x $r(x) = (|x_{i_1}|, |x_{i_2}|, \dots, |x_{i_n}|)^\top$, with $|x_{i_\ell}| \geq |x_{i_{\ell+1}}|$, $\forall \ell = 1, \dots, n-1$.

7: Parameters estimation:

$$\alpha^{(t+1)} = \frac{\sum_i \pi_i^{(t+1)}\left|x_i^{(t+1)}\right| + \epsilon^{(t+1)}}{\left\|\pi^{(t+1)}\right\|_1}$$

$$\beta^{(t+1)} = \frac{\sum_i \left(1-\pi_i^{(t+1)}\right)\left|x_i^{(t+1)}\right| + \epsilon^{(t+1)}}{\left\|\mathbb{1} - \pi^{(t+1)}\right\|_1}$$

8: **end for**

pursuit. Given a support estimate, the estimation is performed by solving a truncated basis pursuit problem. Also in [48], an iterative algorithm, called WSPGL1, is designed to solve a sequence of weighted LASSO using a support estimate, derived from the data, and updated at every iteration. Compared to threshold-ISD and WSPGL1, 2-LMM-tuned iterative shrinkage-thresholding does not use binary weights and is more flexible. Moreover, in threshold-ISD, like CoSaMP, the identification of the support is based on greedy rules and not chosen to optimally fit the prior distribution of the signal.

A prior estimation based on EM was incorporated within the AMP framework also in [49] where a Gaussian mixture model is used as the parametric representation of the signal. The key difference in our approach is that model fitting is used to estimate the support and to adaptively select the best thresholding function with the minimum mean square error. The necessity of selecting the best thresholding function is also proposed in parametric SURE AMP [50] where a class of parametric denoising functions is used to adaptively choose the best-in-class denoiser. However, at each iteration, parametric SURE AMP needs to solve a linear system and the number of parameters affects heavily both performance and complexity.

5.4 Convergence analysis

Under suitable conditions, we are able to guarantee the convergence of the iterates produced by Algorithm I and discuss sufficient condition for optimality.

Definition 2 *A point $(x^*, \pi^*, \alpha^*, \beta^*, \epsilon)$ is called a τ-stationary point of (12) if it satisfies the following relation*

$$x^* = \eta_{\omega^*\lambda\tau}\left(x^* + \tau A^\top(y - Ax^*)\right), \tag{19a}$$

$$\omega_i^* = \frac{\pi_i^*}{\alpha^*} + \frac{1-\pi_i^*}{\beta^*}, \tag{19b}$$

$$a_i^* = \frac{\frac{1-p}{\alpha^*}\exp\left(-\frac{|x_i^*|}{\alpha^*}\right)}{\frac{1-p}{\alpha^*}\exp\left(-\frac{|x_i^*|}{\alpha^*}\right) + \left(\frac{p}{\beta^*}\exp\left(-\frac{|x_i^*|}{\beta^*}\right)\right.} \tag{19c}$$

$$\pi^* = \sigma_{n-K}(a), \tag{19d}$$

$$\alpha^* = \sum_{i=1}^n \frac{\pi_i^*\left|x_i^*\right| + \epsilon}{\sum_{j=1}^n \pi_j^*}, \quad \beta^* = \sum_{i=1}^n \frac{\left(1-\pi_i^*\right)\left|x_i^*\right| + \epsilon}{\sum_{j=1}^n \left(1-\pi_j^*\right)}. \tag{19e}$$

Theorem 1 *If $(x^*, \pi^*, \alpha^*, \beta^*, \epsilon)$ is a minimizer of (12) then it is a τ-stationary point of (12) with $\tau < 2\|A\|_2^{-2}$. Viceversa, if $(x^*, \pi^*, \alpha^*, \beta^*, \epsilon)$ is a τ-stationary point of (12) with $\tau < 2\|A\|_2^{-2}$, then it is a local minimizer of (12).*

The proof can be obtained with similar techniques, devised in [51], and we omit the proof for brevity. This result provides a necessary condition for optimality and shows that, being the function in (12) not convex, τ-stationarity points are only local minima. The next theorem ensures that also the sequence $(\zeta^{(t)})$ converges to a limit point which is also a τ-stationary point of (12) of the algorithm and, from Theorem 1, a local minimum for (12). Moreover, in Theorem 3, we derive an upper bound on the reconstruction error.

Theorem 2 (2LMM-ISTA convergence) . *Let us assume that for every index set $\Gamma \subseteq [n] = K$, the columns of A associated with Γ are linearly independent, $\tau^{(t)} = \tau < 2\|A\|_2^{-2}$, $u^{(t)} = 0$. Then for any $y \in \mathbb{R}^m$, the sequence*

$\zeta^{(t)} = \left(x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}\right)$ *generated by Algorithm* 1 *converges to* $(x^\infty, \pi^\infty, \alpha^\infty, \beta^\infty, \epsilon^\infty)$ *which satisfies relations in* (19).

Definition 3 *Let A be an* $m \times n$ *matrix and let* $1 \leq s \leq n$ *be an integer. The matrix A is said to satisfy the s-restricted isometry property with restricted isometry constant* $\delta_s \in (0,1)$ *if, for every x with* $|supp(x)| \leq s$, *it holds*

$$(1-\delta_s)\|x\|_2^2 \leq \|A_{sx_s}\|_2^2 \leq (1+\delta_s)\|x\|_2^2.$$

Theorem 3 (2LMM-ISTA: upper bound on the error) *Suppose that A is an* $m \times n$ *sampling matrix with restricted isometry constant* δ_{2K}. *Let* x^∞ *be the output of Algorithm* 1 *with* $\tau^{(t)} = \tau < 2\|A\|_2^{-2}$, $u^{(t)} = 0$ *and* $\Lambda^\infty = supp(x^\infty)$. *Let* $e = x^\infty - x_\star$. *If* $r(x^\infty)_{K+1} = 0$ *and* $\|\sigma_K(A^\top_{(\Lambda^\infty)^c}(y - Ax^\infty))\| \leq \|A^\top_{\Lambda^\infty}(y - Ax^\infty)\| = c$, *then*

$$\|e_{\Lambda^\infty}\| \leq \frac{1}{1-2\delta_{2K}}\left(\frac{\lambda(1-\delta_{2K})\sqrt{K}}{\beta^\infty} + c\delta_{2K}\right)$$

and

$$\left\|e_{(\Lambda^\infty)^c \cap \Lambda^\star}\right\| \leq \frac{c}{1-\delta_{2K}} + \frac{\delta_{2K}}{1-\delta_{2K}}\|e_{\Lambda^\infty}\|.$$

It should be noticed that the result in Theorem 3 implies that the mean square error $\mathsf{MSE} = \|e\|^2/n = \left(c + \frac{\lambda\sqrt{K}}{\beta^\infty}\right)^2/n + O(\delta_{2K})$. In this sense, we have provided conditions verifiable a posteriori for convergence in a neighborhood of the solution. This is a common feature in shrinkage-thresholding methods. In [52], it is shown that, in the absence of noise, if certain conditions are satisfied, the error provided by Lasso is $O(\lambda)$, where λ is the regularization parameter. Since ISTA and FISTA converge to a minimum of the Lasso, we argue that the same estimate holds also for the error between the provided estimations and the true signal.

The proof of Theorems 2 and 3 are postponed to Appendices 3 and 4 and are obtained using arguments of variational analysis and analysis of τ-stationary point of (12), respectively.

Example 1 *Computing* δ_{2K} *is hard in practice. However, for i.i.d Gaussian and Rademacher matrices, the RIP holds with high probability when* $m \geq c_0(\delta/2)k\log(ne/k)$ *where* c_0 *is a function of isometry constant* δ [34]. *To give an example, if* $n = 10000$, $m = 8000$, *and* $k = 10$, *then the RIP property holds with probability 0.98 with isometry constant equal to* $\delta_{2k} = 0.4$. *Running the proposed iterative algorithms (ISTA or FISTA) with* $\lambda = 10^{-3}$, *we can empirically check that the condition* $\|\sigma_K(A^\top_{(\Lambda^\infty)^c}(y - Ax^\infty))\| \leq \|A^\top_{\Lambda^\infty}(y - Ax^\infty)\| \leq c$ *is always satisfied with* $c = 1.7176 \cdot 10^{-4}$. *The error of the provided estimate is* $\mathsf{MSE} = \|x^\star - x\|^2/n \approx 1.47 \cdot 10^{-10}$ *and the estimated upper bound, obtained using Theorem 3, is* $5.143 \cdot 10^{-10}$. *Using error bounds in* [52], *we are able to guarantee that the solutions provided by ISTA and FISTA are accurate with an error only proportional to* $\lambda = 10^{-3}$.

6 Numerical results, experiments, and discussion

In this section, we compare several first-order methods with their versions augmented by the support estimation, in terms of convergence times and empirical probability of reconstruction in the absence and in the presence of noise. It is worth remarking that this does not represent a challenge among all first-order methods for compressed sensing. Our aim is to show that the combination of support detection and estimation using an iterative reweighted first-order method can improve several iterative shrinkage methods. In other terms, we want to show that, given a specific algorithm for CS, the speed and the performance can be improved via its 2-LMM counterpart. Moreover, in order to show that the choice of the weights is important to obtain fast algorithms and good performance, we have employed an iterative shrinkage method for iterative reweighted ℓ_1 minimization algorithm (IRL1). In [45], IRL1 requires to solve at each step a weighted ℓ_1 minimization. This algorithm has computational complexity which is not comparable with the iterative shrinkage/thresholding algorithms since each iteration has complexity of order $O(n^3)$. We employ a shrinkage-thresholding method for IRL1 in the spirit of [53] and show that the performance are not as good as in the proposed methods. What we mean as IRL1 is summarized in Algorithm 2.

Algorithm 2 IRL1 - tuned iterative shrinkage-thresholding

Require: Data (y, A), set K

1: Initialization: $\epsilon^{(0)} = 1$

2: **for** $t = 1, \ldots,$ *StopIter* **do**

3: Computation of ℓ_1-weights:

$$\omega_i^{(t+1)} = \frac{1}{\left|x_i^{(t+1)}\right| + \epsilon^{(t)}}$$

4: Gradient/Thresholding step:

$$x^{(t+1)} = \eta^{\mathsf{S}}_{\lambda\tau^{(t)}\omega^{(t+1)}}\left(x^{(t)} + \tau^{(t)}A^\top\left(y - Ax^{(t)}\right) + u^{(t)}\right)$$

5: Regularization parameter:

$$\theta^{(t+1)} = \frac{1}{\log(t+1)} + c\left\|x^{(t+1)} - x^{(t)}\right\| + \frac{r\left(x^{(t+1)}\right)_{K+1}}{n}$$

$$\epsilon^{(t+1)} = \min\left(\epsilon^{(t)}, \theta^{(t+1)}\right)$$

6: **end for**

6.1 Reconstruction from noise-free measurements

6.1.1 Rate of convergence

As a first experiment, we consider Bernoulli-Gaussian signals [54]. More precisely, the signal to be recovered has length $n = 560$ with $k = 50$ nonzero elements drawn from a $\mathsf{N}(0,4)$, respectively. The sensing matrix A with $m = 280$ rows is sampled from the Gaussian ensemble with zero mean and variance $1/m$. We fix $\lambda = 10^{-3}$ and $\tau = 0.19$, and the mixture parameters are initialized $\alpha^{(0)} = 1, \pi^{(0)} = \mathbb{1}, K = k + 10$, and $p = K/n$.

In Fig. 5, we compare the convergence rate of ISTA, FISTA, IRL1, and AMP with the corresponding methods with 2-LMM-tuning (2-LMM-ISTA, 2-LMM-FISTA, and 2-LMM-AMP). In particular, the mean square error (MSE) of the iterates $\mathsf{MSE}(t) = \|x^{(t+1)} - x^\star\|^2/n$ averaged over 50 instances is depicted as a function of the iteration number.

A few comments are in order. The sparsity problem that ISTA, FISTA, and AMP are intended to approximately solve is the same (basis pursuit or Lasso problem). However, the convergence results are different for these iterative algorithms. More precisely, in the absence of noise,

- ISTA and FISTA, under certain conditions, are analytically proved to converge to a minimum of Lasso. This solution is shown to provide only an approximation of the sparse solution $x^\star$ which is controlled by Lasso parameter λ. More precisely, $\|x^\star - \widehat{x}\|_2 \leq C\lambda$ where $C \in \mathbb{R}$ and perfect reconstruction is not guaranteed.
- AMP instead is not proved to converge in a deterministic sense. In [14], only the average case performance analysis is carried out. The authors exploit the randomness of A and instead of calculating the limit of $\|x^t - x^\star\|^2$, they show the convergence in the mean square sense $\mathbb{E}\|x^t - x^\star\|^2 \to 0$.

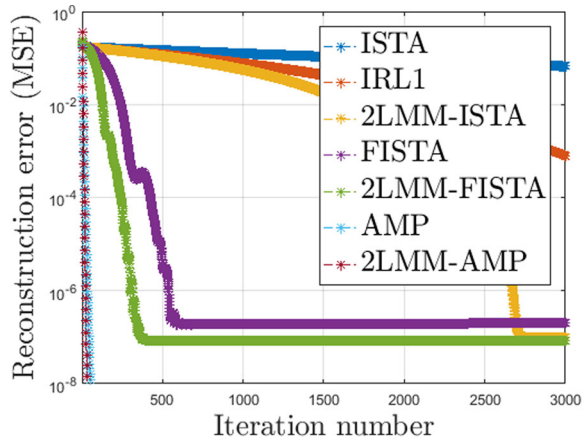

Fig. 5 Convergence rate: evolution of the MSE for classical thresholding method algorithms and the corresponding versions with 2-LMM-tuning for sparse Bernoulli-Gaussian signals

In Fig. 5 the accuracy of the solution provided by ISTA, FISTA, and AMP are different. The difference of AMP has been already explained. The difference between ISTA and FISTA is due to the fact that $\lambda = 0.005$ for ISTA (to speed up the algorithm) and $\lambda = 0.001$ for FISTA. As already observed, we are not interested in a challenge among all first-order methods for CS. Our aim is to show that the combination of support detection and estimation using an iterative reweighted first-order method can improve a series of iterative shrinkage methods. More precisely, given a specific algorithm for CS, the speed can be improved via its 2-LMM counterpart.

It should be noted that the proposed algorithms are much faster than classical iterative shrinkage-thresholding methods: there is about a 81,37, and 35% of reduction in the number of iterations needed for the convergence of 2-LMM-ISTA, 2-LMM-FISTA, and 2-LMM-AMP, respectively.

6.1.2 Effect of the prior distribution

We now show a second experiment: we fix $n = 512$ and take the fraction of the nonzero coefficients fixed to $\rho = k/n$ and we study the effect of the nonzero coefficients distribution on the empirical probability of reconstruction for different values of $k \in [1, 250]$. More precisely, $x_i^\star \sim (1-\rho)\delta_0(x_i^\star) + \rho g(x_i^\star)$ where g is a probability distribution function and δ_0 is the Dirac delta function. In Table 1, the acronyms of the considered distributions are summarized (see also [55]).

Figures 6, 7, 8, and 9 (left) show the empirical recovery success rate, averaged over 50 experiments, as a function of the signal sparsity for different signal priors (see Table 2). For all recovery algorithms, the convergence tolerance has been fixed to 10^{-4}. In this case, the elements of matrix A with $m = 350$ are sampled from a normal distribution with variance $1/m$. We have fixed a total number of iterations equal to 1000. The algorithm parameters have been initialized as in Table 3.

It should be noticed that for ISTA, IRL1, and 2-LMM-ISTA, we have chosen $\lambda = 0.005$, instead of $\lambda = 0.001$ as in FISTA and 2-LMM-FISTA, in order to speed up the convergence. The convergence of ISTA and IRL1 are extremely slow, and before 1000 iterations, we get only an approximation with an error of order 10^{-3} for sparsity larger than 50 in most of the cases.

It should be noticed that the 2-LMM-tuning improves the performance of iterative shrinkage-thresholding methods in terms of sparsity-undersampling trade-off. For example for 5P1, it turns out that the signal recovery with 2-LMM-tuning is possible with 30%, 63% sparsity level higher than FISTA, and AMP, respectively.

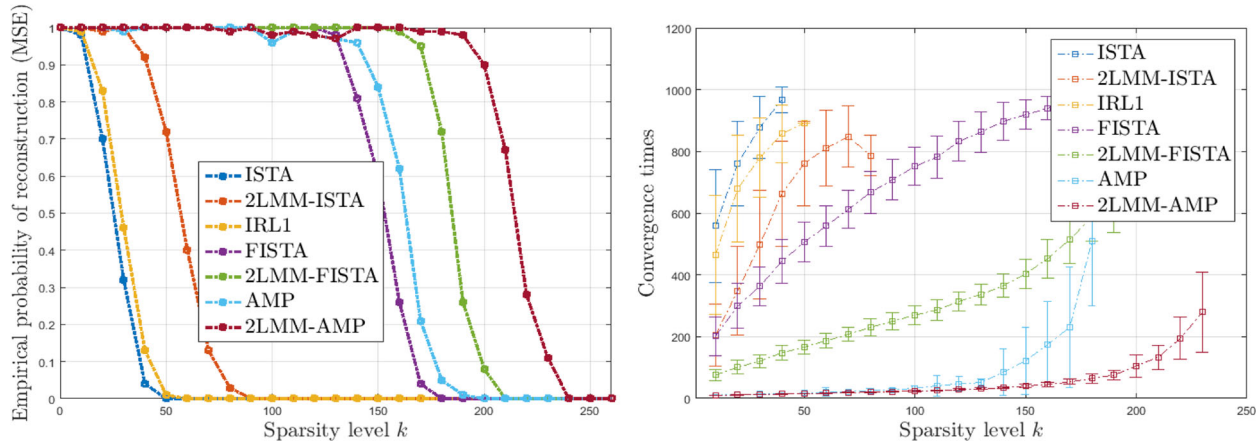

Fig. 6 Analysis of performance for L1 signals: empirical probability of successful recovery as a function of the sparsity value k with $n = 512$ and $m = 350$ sparse Bernoulli-Laplace signals (L1)

Figures 6, 7, 8, and 9 (right) show the average running times (CPU times in seconds) of the algorithms computed over the successful experiments, and the error bar represents the standard deviation of uncertainty for different signal priors. These graphs demonstrate the benefit of 2-LMM-tuning for iterative shrinkage/thresholding methods. Not only 2-LMM-tuning shows better performance in the reconstruction but it also runs much faster than traditional methods. Despite the additional per iteration computational cost needed to update the mixture parameters, the gain of the 2-LMM-tuning ranges from 2 to over 6 times, depending on the signal prior. The algorithm efficiency can be attributed to the simple form of the model used as parametric representation of the signal and the improved runtime performance comes from the effective denoising so that fewer iterations are required to converge.

6.2 Reconstruction in imperfect scenarios

In this section, we compare the first-order methods with their versions augmented by the support estimation for recovery of signals in imperfect scenarios where the signal is not exactly sparse or the measurements are noisy.

6.2.1 Not exactly sparse signals

In this experiment, we investigate the performance of first-order methods with 2-LMM-tuning for signals that are not strictly sparse. We consider signals of the form $x^\star = x_0 + \xi$ where x_0 is drawn i.i.d from the ensemble of Bernoulli-Gaussian signals (see G1 in Table 2) of length $n = 512$ with sparsity level $k = 56$ and $\xi \in \mathbb{R}^n$ is a vector whose components are distributed as $\mathsf{N}\left(0, \sigma^2\right)$ with $\sigma = 0.01$. Here, k can be interpreted as the compressibility level of the signal $x^\star$. The sensing matrix A with $m \in [160, 360]$ rows is sampled from the Gaussian ensemble

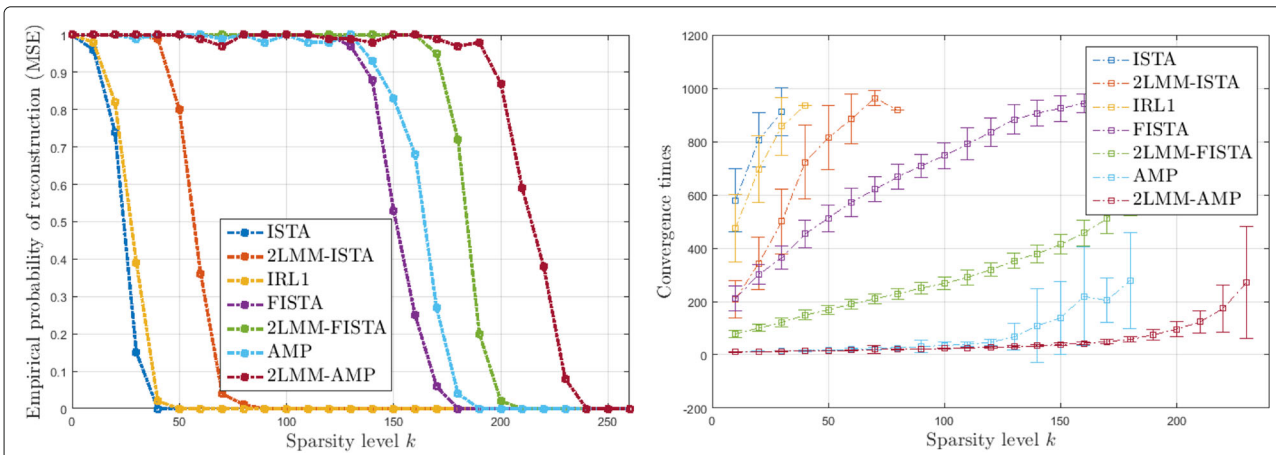

Fig. 7 Analysis of performance for G1 signals: empirical probability of successful recovery as a function of the sparsity value k with $n = 512$ and $m = 350$ for sparse Bernoulli-Gaussian signals (G1)

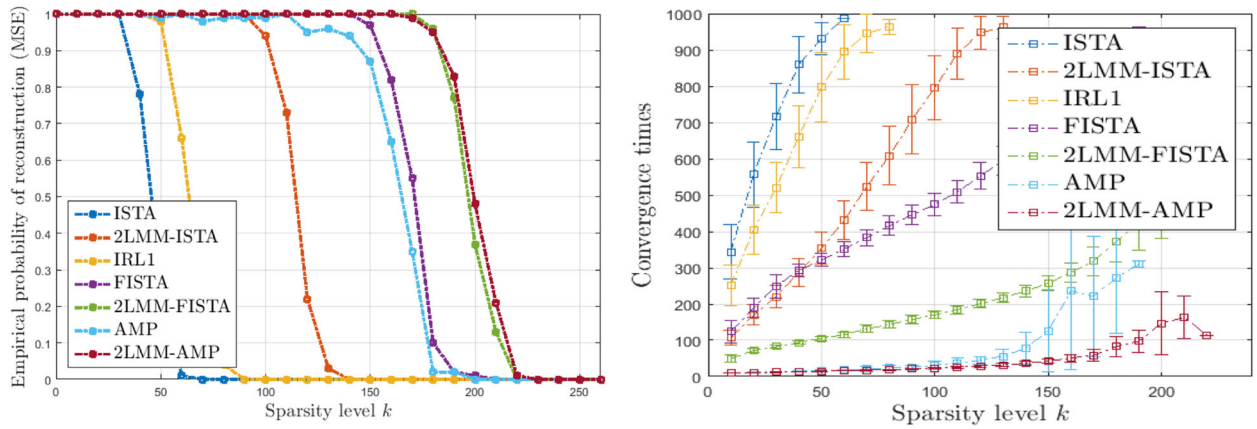

Fig. 8 Analysis of performance for U1 signals: empirical probability of successful recovery as a function of the sparsity value k with $n = 512$ and $m = 350$ for sparse Bernoulli-Uniform signals (U1)

with zero mean and variance $1/m$. Then, the reconstruction is performed using measurements $y = Ax^\star$. The first-order methods are compared with their versions augmented by 2-LMM-tuning: the parameters have been initialized as follows: $\lambda = 10^{-3}, \tau = 0.1, \alpha^{(0)} = 1, \pi^{(0)} = \mathbb{1}$, $K = k + 10$, and $p = K/n$. Figure 10 shows the MSE achieved by 3000 iterations of the algorithms as a function of measurements used for the reconstruction. As we can see, the algorithms with 2-LMM-tuning have similar reconstruction performance and outperform significantly their traditional counterpart (for example 2-LMM-AMP needs 62% of measurements required by AMP to reach a similar accuracy of MSE $= 10^{-4}$). In this case the improved performance comes from the effective denoising so that fewer iterations are required to achieve a better accuracy.

6.2.2 Reconstruction with noisy measurements

We now fix $n = 512$ and we study the performance of the proposed methods in scenarios with inaccurate measurements according to (1). In this case, $x^\star$ is a random Bernoulli-Uniform signal (see U1 in Table 2) with sparsity degree $k = 56$ and the noise η is white Gaussian noise with standard deviation $\sigma = 0.01$. In this case, the parameters are set as follows: $\lambda = 10^{-3}, \tau = 0.1, \alpha^{(0)} = 1, \pi^{(0)} = \mathbb{1}, K = k + 10$, and $p = K/n$. The MSE achieved by 3000 iterations is depicted as a function of the number of measurements used in the reconstruction. Also in this case, the best results are obtained by methods with 2-LMM-tuning. The efficiency of the proposed algorithms allows to reduce the number of measurements required to achieve a satisfactory level of accuracy. As can be noticed from Fig. 10 (right), 2-LMM-FISTA and 2-LMM-AMP need fewer observations (about 180 measurements) than FISTA and AMP (about 270 measurements) to achieve MSE $= 10^{-4}$.

In Fig. 11, we show that the proposed methods are robust against noise. More precisely, the mean square error, averaged over 50 runs, and obtained after 3000

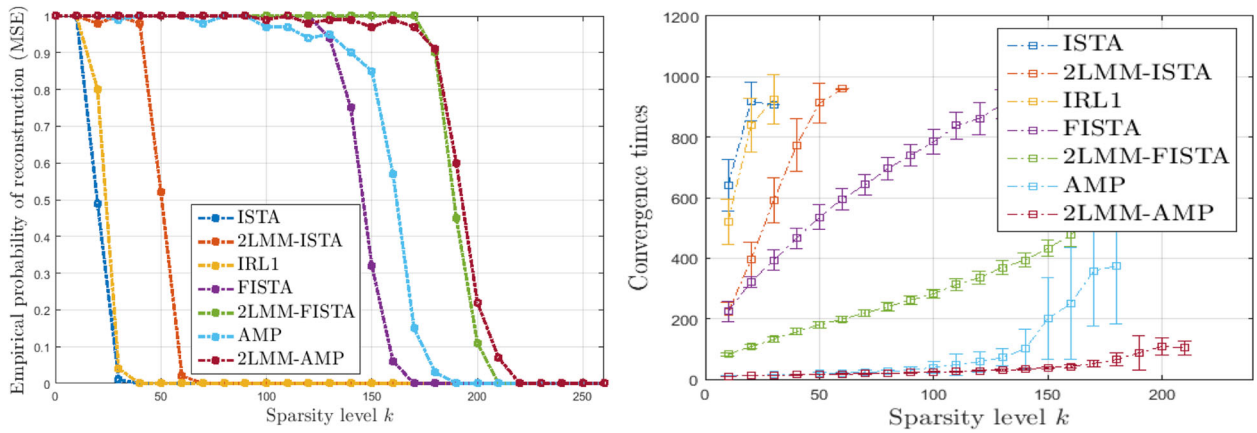

Fig. 9 Analysis of performance for 5P1 signals: empirical probability of successful recovery as a function of the sparsity value k with $n = 512$ and $m = 350$ for sparse Bernoulli-Binomial signals (5P1)

Table 2 Nonzero coefficients distribution

Notation	g
L1	Lap(0, 4)
G1	$N(0,4)$
U1	$U[0,4]$
5P1	$\mathbb{P}(x=-1)=\mathbb{P}(x=1)=0.3$ $\mathbb{P}(x=-5)=\mathbb{P}(x=5)=0.2$

iterations, is depicted as a function of signal-to-noise ratio (SNR), defined as follows

$$\mathrm{SNR}=\frac{\mathbb{E}\left[\|Ax\|^2\right]}{m\sigma^2}.$$

As to be expected, as the SNR increases the MSE goes to zero. Moreover, the MSE of the proposed algorithms are smaller than those obtained via classical iterative thresholding algorithms. As already observed, the MSE of ISTA is very high compared to the other methods. This is due to the fact that the algorithm is very slow and the number of iterations are not enough to reach a good recovery error.

In our setup, we have considered only Gaussian Noise and the robustness against multiplicative noise is out of our scope. This would require a drastic modification of the proposed algorithms and will be subject of future research. For example, when the underlying sparse or approximately sparse signal is a vector of nonnegative intensities whose measurements are corrupted by Poisson noise, standard CS techniques cannot be applied directly, as the Poisson noise is signal-dependent [56, 57]. In this case, the rationale of our method can be adapted combining the use of mixtures models with exponential distribution instead Laplace distribution with penalized negative Poisson log-likelihood objective function with nonnegativity constraints. We refer to [58] for more details on the model and on the implementations of related iterative thresholding algorithms.

If the multiplicative noise is due to hardware's amplification and is not signal-dependent, we can model the measurements as follows

Table 3 Parameters of several shrinkage-thresholding algorithms

	λ	$\tau^{(t)}$	$\pi^{(0)}$	$\alpha^{(0)}$	K
ISTA	0.005	0.2	–	–	–
IRL1	0.005	0.2	–	–	66
FISTA	0.001	0.2	–	–	66
2-LMM-ISTA	0.005	0.2	$\mathbb{1}$	0.1	66
2-LMM-FISTA	0.001	0.2	$\mathbb{1}$	0.1	66
AMP	0.9	Eq. (17), $\chi=0.9$	–	–	–
2-LMM-AMP	0.9	Eq. (17), $\chi=0.9$	$\mathbb{1}$	0.1	66

$$y=DAx^\star+\eta$$

where $D=\mathrm{diag}(\exp(\mathcal{N}(0,\sigma^2)))$, where D is a diagonal matrix of noise and σ is the parameter governing the amplitude of decalibration. To address this problem, the most standard existing approach is the blind calibration for compressed sensing [59]. More precisely, the sparse regularization is exploited considering A as an inaccurate estimate of the true measurement system $A^\star=DA$ and $y=Ax^\star+(\widehat{A}-A)x^\star+\eta\approx Ax^\star+\varepsilon+\eta$ with ε an estimate of the magnitude of this added noise $\left\|\left(\widehat{A}-A\right)x^\star\right\|$. We refer to [59] for more sophisticated approaches of blind supervised calibration and adaptations of classical methods that perform both calibration and reconstruction.

6.3 Comparison with structured sparsity-based Bayesian compressive sensing

Many authors have recently developed structured sparsity-based Bayesian compressive sensing methods in order to deal with different signals arising in several applications and adaptively explore the statistical structure of nonzero pattern. We refer the interested reader to the repository http://people.ee.duke.edu/~lcarin/BCS.html for an introduction to Bayesian compressive sensing (BCS) methods and to structured sparsity-based Bayesian compressive sensing.

For example, [60] proposes a spatio-temporal sparse Bayesian method to recover multichannel signals simultaneously, not only exploiting temporal correlation within each channel signal but also exploiting inter-channel correlations among different signals. This method has been shown to provide several advantages in applications in brain computer interface and electroencephalography-based driver's drowsiness estimation in terms of measurements for reconstruction and computational load. In [61], using a new bilinear time-frequency representation, a redesigned BCS approach is developed for the problem of spectrum estimation of multiple frequency-hopping signals, arising in various communication and radar applications in the context of multiple-input multiple-output (MIMO) operations in the presence of random missing observations. Another example of structured sparsity-based Bayesian compressive sensing comes from the context of reconstruction of signals and images that are sparse in the wavelet basis [62] or in DCT basis with applications to image compression. More precisely, in [62], the statistical structure of the wavelet coefficients is exploited explicitly using a tree-structured Bayesian compressive sensing approach. This tree structure assumption shows several advantages in terms of number of measurements required for reconstruction.

It is worth remarking that in our approach, we do not use any prior information on the structure of the sparsity pattern and we expect that structured

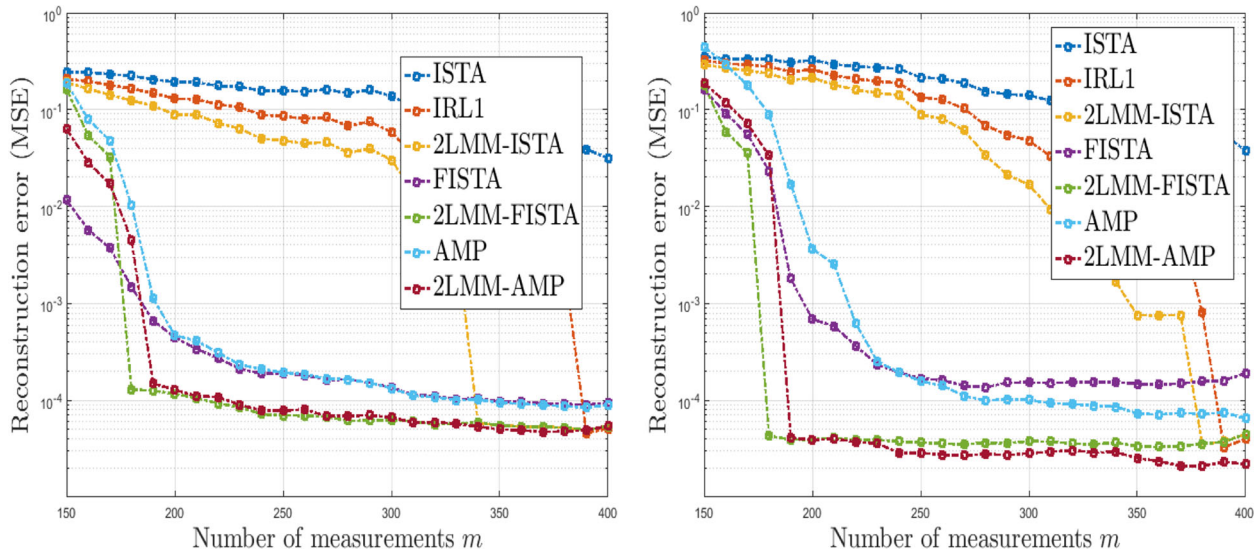

Fig. 10 Analysis of performance in imperfect scenarios: MSE versus number of measurements for recovery of not exactly sparse signals (left) and of exactly sparse signals in presence of noise (right), $n = 512$, $k = 56$

sparsity-based methods outperform our approach. A detailed comparison and an ad hoc adaptation of our approach to all specific frameworks mentioned above is out of the scope of this paper. However, in this section, we propose a numerical comparison of 2-LMM-tuning FISTA with tree-structured wavelet-based Bayesian compressive sensing (WBCS). The implementation of WBCS algorithm used for comparison are implemented via a hierarchical Bayesian framework, with the tree structure incorporated naturally in the prior setting. See TS-BCS for wavelet via MCMC in the repository `http://people.ee.duke.edu/~lcarin/BCS.html` for a detailed description of the code.

For the comparison, we consider the setting in [62] with a signal of length $n = 512$ that are sparse in the Haar wavelet basis and whose coefficients are not independent as in classical compressed sensing framework. Specifically, under the wavelet basis, if a parent node in a wavelet

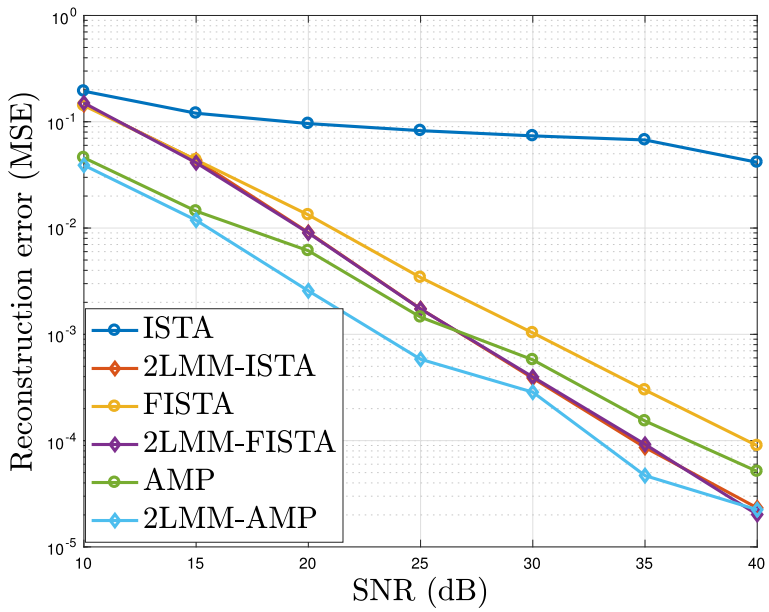

Fig. 11 Analysis of robustness against noise: the mean square error is depicted as a function of the SNR

tree is zero or close to zero, with a very large probability, its children nodes are also zero or close to zero. We refer to [62] for details on the generation of the signal. The sparsity of the considered signal is $k = 63$. In this case, the parameters are set as follows: $\lambda = 10^{-3}, \tau = 0.05, \alpha^{(0)} = 1, \pi^{(0)} = \mathbb{1}$, $K = k + 10$, and $p = K/n$. In Fig. 12 the reconstruction error achieved by 10000 iterations is depicted as a function of the number of measurements used in the reconstruction. It is worth remarking that WBCS explores the statistical structure of the wavelet coefficients to reduce the number of CS measurements and goes beyond simply assuming that the data are compressible in a wavelet basis. As to be expected, since more a priori knowledge is employed, WBCS shows better performance in terms of reconstruction accuracy. However, the gap is not large and the 2-LMM-FISTA tuning is able to learn the sparsity model and, as soon as the number of measurements is larger than 200, we obtain a good reconstruction accuracy, of order 10^{-4} for 2-LMM-FISTA and of order 10^{-5} for WBCS.

6.4 Deblurring images

In order to show the effectiveness of the 2-LMM-tuning, in this section, we repeat the same experiment proposed in [9] for deblurring two test images (Lena and cameraman). In [9], it has been shown that FISTA significantly outperforms ISTA and other first-order methods in terms of the number of iterations required to achieve a given accuracy. For this reason, we compare the performance of FISTA with our proposed algorithm 2-LMM-FISTA.

In the considered setting, both images have equal size 256×256 and all pixels of the original images are scaled into the range between 0 and 1. A Gaussian blur of size 9×9 and standard deviation 4 are applied to both images and an additive zero mean white Gaussian noise with standard deviation 10^{-4} is added. The original and observed images are given in Figs. 13 and 14, respectively. We then test FISTA and 2-LMM-FISTA for recovery, where y represents the (vectorized) observed image, and $A = RW$, where R is the matrix representing the blur operator and W is the inverse of a three-stage Haar wavelet transform. The regularization parameter is fixed as in [9] $\lambda = 2 \cdot 10^{-5}$, and the blurred image is used as initial condition. For 2-LMM-ITA, the parameters are set as follows: $\alpha^{(0)} = 1$, $\pi^{(0)} = \mathbb{1}, K = 10000$, and $p = K/n$.

In Fig. 15, the evolution of the error in dB is depicted as a function of the number of iterations. In particular, the images produced by 2-LMM-FISTA exhibit better quality than those obtained by using the classical version of FISTA. In Fig. 13 and 14, the reconstructions obtained by the competing methods are shown for Lena and cameraman, respectively. As can be seen in Figs. 13 and 14, 2-LMM-FISTA achieves significantly better visual quality, as the amount of noise is minimized and visual artifacts are greatly reduced. This is also reflected by the reconstruction PSNR, which is significantly higher for 2-LMM-FISTA.

7 Conclusions

In this paper, we proposed a new method to perform both support detection and sparse signal estimation in compressed sensing. Combining MAP estimation with the parametric representation of the signal with a Laplace mixture model, we formulated the problem of

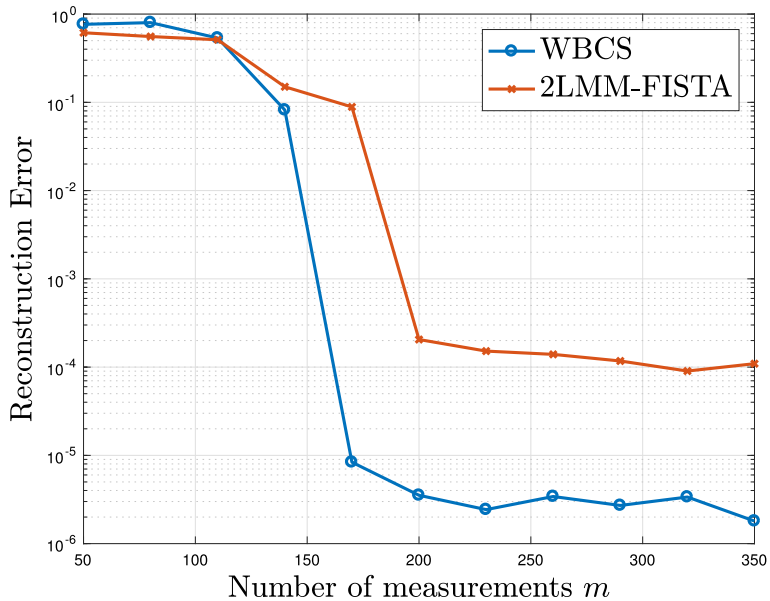

Fig. 12 Comparison between WBCS and 2-LMM: reconstruction error of a signal with $n = 512$ and $k = 63$ as a function of number of measurements

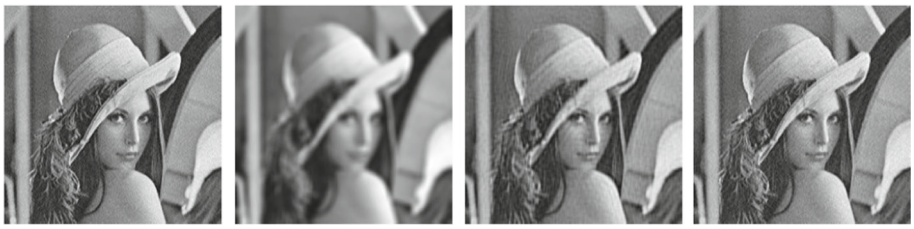

Fig. 13 Analysis of performance for Lena image deblurring: **a** original image, **b** acquired image, **c** FISTA reconstruction MSE (dB) = − 10.718 (after 1000 iterations), **d** 2-LMM-FISTA reconstruction MSE (dB) = − 13.3958 (after 1000 iterations)

reconstruction as a reweighted ℓ_1 minimization. Our contribution includes theoretical derivation of necessary and sufficient conditions for reconstruction in the absence of noise. Then, 2-LMM-tuning has been proposed to improve the performance of several iterative shrinkage-thresholding algorithms. Iterative procedures have been designed by combining EM algorithm with classical iterative thresholding methods. Numerical simulations show that these new algorithms are faster than classical ones and outperform them in terms of phase transitions. Topics of our current research is to use similar technique based on Laplace mixture models for robust compressed sensing, where measurements are corrupted by outliers (see [63] and reference therein).

8 Appendix 1: proof of Proposition 1

The proof of Proposition 1 is a direct consequence of a more general result on compressible prior result, formally stated in Proposition 1 in [17].

Lemma 1 (Proposition 1.1 in [17]) *Suppose $x_n \in \mathbb{R}^n$ is i.i.d. with respect to a distribution $p(x)$. Denote $p(x) := 0$ for $x < 0$, and $\overline{p}(x) := p(x) + p(-x)$ for $x \geq 0$ as the probability density function of $|X_n|$, and $\overline{F}(t) := \mathbb{P}(|X| \leq t)$ as its cumulative density function. Assume that $\overline{F}$ is continuous and strictly increasing on some interval $[a, b]$, with $\overline{F}(a) = 0$ and $\overline{F}(b) = 1$, where $0 \leq a \leq b \leq \infty$. For any $0 < \kappa \leq 1$, define the following function*

$$G_2[p](\kappa) := \frac{\int_0^{\overline{F}^{-1}(1-\kappa)} x^2 \overline{p}(x)\mathrm{d}x}{\int_0^{\infty} x^2 \overline{p}(x)\mathrm{d}x}.$$

If $\mathbb{E}[|X|^q] < \infty$ for some $q \in (0, \infty)$. Then, $G_q[p](\kappa)$ is also well defined for $\kappa = 0$, and for any sequence k_n such that $\lim_{n\to\infty} k_n/n = \kappa \in [0, 1]$ the following holds almost surely

$$\lim_{n\to\infty} \overline{\varrho}_{k_n}(x_n)^2 = G_2[p](x)$$

Proof of Proposition 1. For 2-LMM, we have

$$p(x) = \frac{(1-p)}{2\alpha} \mathrm{e}^{-|x|/\alpha} + \frac{p}{2\beta} \mathrm{e}^{-|x|/\beta},$$

from which we get

$$\overline{F}(t) = \int_0^t \overline{p}(x)\mathrm{d}x = 1 - (1-p)\mathrm{e}^{-t/\alpha} - p\mathrm{e}^{-t/\beta}.$$

Then, let $t = t(\kappa) = \overline{F}^{-1}(1-\kappa)$, i.e., be the solution

$$(1-p)\mathrm{e}^{-t/\alpha} + p\mathrm{e}^{-t/\beta} = \kappa.$$

We now compute

$$\int_0^t x^2 \overline{p}(x)\mathrm{d}x = (1-p)\left[\alpha^2 - \left(\alpha^2 + \alpha t + \frac{t^2}{2}\right)\mathrm{e}^{-t/\alpha}\right] + p\left[\beta^2 - \left(\beta^2 + \beta t + \frac{t^2}{2}\right)\mathrm{e}^{-t/\beta}\right]$$

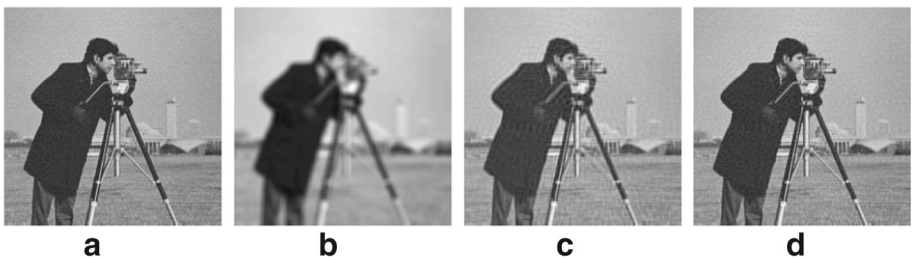

Fig. 14 Analysis of performance for cameraman image deblurring: **a** original image, **b** acquired image, **c** FISTA reconstruction MSE (dB) = − 11.4731 (after 1000 iterations), **d** 2-LMM-FISTA reconstruction MSE (dB) = − 14.4731 (after 1000 iterations)

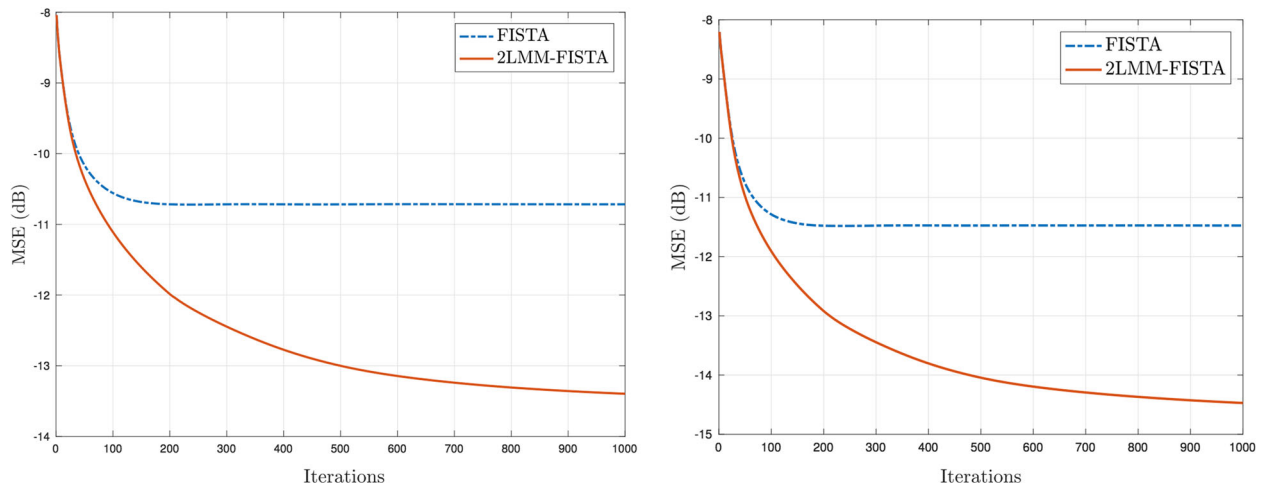

Fig. 15 Analysis of rate of convergence in image deblurring: Lena (left) and cameraman (right): MSE (dB) versus number of iterations

and

$$\int_0^\infty x^2 \overline{p}(x)\mathrm{d}x = (1-p)\alpha^2 + p\beta^2.$$

Then, the assertion is proved by applying Lemma 1 and we obtain

$$\lim_{n\to\infty} \overline{\varrho}_{k_n}(x_n)^2 = \frac{(1-p)\left[\alpha^2 - \left(\alpha^2+\alpha t+\frac{t^2}{2}\right)\mathrm{e}^{-t/\alpha}\right] + p\left[\beta^2 - \left(\beta^2+\beta t+\frac{t^2}{2}\right)\mathrm{e}^{-t/\beta}\right]}{(1-p)\alpha^2+p\beta^2}$$

9 Appendix 2: proof of Proposition 2

Recall that

$$f(x|y;\Theta) = f(x;\Theta)f(y|x;\Theta)$$
$$f(x;\Theta) = \prod_{i=1}^n f(x_i;\Theta)$$
$$f(x_i;\Theta) = \frac{(1-p)}{2\alpha}\mathrm{e}^{-|x_i|/\alpha} + \frac{p}{2\beta}\mathrm{e}^{-|x_i|/\beta}$$

and define

$$f(x,z;\Theta) = \prod_{i=1}^n f(x_i,z_i;\Theta)$$
$$f(x_i,z_i;\Theta) = \frac{(1-p)z_i}{2\alpha}\mathrm{e}^{-|x_i|/\alpha} + \frac{p(1-z_i)}{2\beta}\mathrm{e}^{-|x_i|/\beta}$$
$$f(z|x;\Theta) = \prod_{i=1}^n f(z_i|x_i;\Theta)$$
$$f(z_i|x_i;\Theta) = \frac{\frac{(1-p)z_i}{2\alpha}\mathrm{e}^{-|x_i|/\alpha} + \frac{p(1-z_i)}{2\beta}\mathrm{e}^{-|x_i|/\beta}}{\frac{(1-p)}{2\alpha}\mathrm{e}^{-|x_i|/\alpha} + \frac{p}{2\beta}\mathrm{e}^{-|x_i|/\beta}}$$

Lemma 2 *The log-likelihood function defined in* (6) *is given by*

$$L(x;\Theta) = \sum_{i=1}^n \sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta)\log\left(f(x_i,z_i;\Theta)\right) - \sum_{i=1}^n \sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta)\log\left(f(z_i|x_i;\Theta)\right) + \log(f(y|x;\Theta))$$

Proof We have the following series of equalities:

$$L(x;\Theta) - \log(f(y|x;\Theta)) = \log\left[\prod_{i=1}^n f(x_i;\Theta)\right] = \sum_{i=1}^n \log\left[f(x_i;\Theta)\right] = \sum_{i=1}^n \log\left(\sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta)\frac{f(x_i,z_i;\Theta)}{f(z_i|x_i;\Theta)}\right)$$

from which we conclude

$$L(x;\Theta) - \log(f(y|x;\Theta)) \geq \sum_{i=1}^n \sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta) \times \left[\log\left(\frac{f(x_i,z_i;\Theta)}{f(z_i|x_i;\Theta)}\right)\right]$$

where the last inequality follows from Jensen's inequality ($\mathbb{E}(\phi(x)) \leq \phi(\mathbb{E}(x))$ and $\phi(x) = \log(x)$ concave function) and "$S_i \sim f(z_i|x_i;\Theta)$" subscripts above indicate that the expectations are with respect to S drawn from $f(z_i|x_i;\Theta)$ distribution. By noticing that $f(x_i,z_i;\Theta)/f(z_i|x_i;\Theta) = f(x_i|\Theta)$, i.e., constant with respect to z_i, we notice that the above inequality is actually an equality. The proof is then concluded by using the logarithm properties:

$$L(x;\Theta) - \log(f(y|x;\Theta))$$
$$= \sum_{i=1}^{n} \sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta) \log\big(f(x_i,z_i;\Theta)\big)$$
$$- \sum_{i=1}^{n} \sum_{z_i\in\{0,1\}} f(z_i|x_i;\Theta) \log\big(f(z_i|x_i;\Theta)\big)$$

□

Proposition 4 *Given y, A, Θ,*

$$-L(x;\Theta) = \begin{cases} J(x,\widehat{\pi};\Theta) - \sum_{i=1}^{n} H(\widehat{\pi}_i) & \text{if } y = Ax \\ +\infty & \text{if } y \neq Ax \end{cases} \tag{20}$$

where

$$J(x,\widehat{\pi};\Theta) = \sum_{i=1}^{n} \Bigg[\frac{\widehat{\pi}_i|x_i|}{\alpha} + \widehat{\pi}_i \log\alpha - \pi_i \log(1-p) + \frac{(1-\widehat{\pi}_i)|x_i|}{\beta} + (1-\widehat{\pi}_i)\log\beta - (1-\widehat{\pi}_i)\log p \Bigg], \tag{21}$$

$H(t) = -t\log t - (1-t)\log(1-t)$ is the natural entropy function with $t \in [0,1]$ and $\widehat{\pi} = \widehat{\pi}_i(x_i) = f(z_i = 1|x_i;\Theta)$.

Proof From Lemma 2 and defining $\widehat{\pi}_i = \widehat{\pi}_i(x_i;\Theta) = f(z_i = 1|x_i;\Theta)$

$$-L(x;\Theta) = \sum_{i=1}^{n} \Bigg[\widehat{\pi}_i \frac{|x_i|}{\alpha} + \widehat{\pi}_i \log\alpha - \widehat{\pi}_i \log(1-p) + \frac{(1-\widehat{\pi}_i)|x_i|}{\beta} + (1-\widehat{\pi}_i)\log\beta - (1-\widehat{\pi}_i)\log p \Bigg] + n\log 2 + \sum_{i=1}^{n} \big(-\widehat{\pi}_i \log(\widehat{\pi}_i) - (1-\widehat{\pi}_i)\log(1-\widehat{\pi}_i)\big) + \log\delta_{\{y=Ax\}}$$

we obtain

$$L(x;\Theta) = -J(x,\widehat{\pi};\Theta) + \sum_{i=1}^{n} H(\widehat{\pi}_i) + \log\delta_{\{y=Ax\}}$$

which gives (9). □

Proof of Proposition 2 . Let us consider $J(x,\pi;\Theta)$ and minimize with respect to π_i by taking x all the other variable fixed. By imposing the constraint

$$\frac{\partial J(x,\pi;\Theta)}{\partial \pi_i} - \sum_{i=1}^{n} \frac{\partial H(\pi_i;\Theta)}{\partial \pi_i} = 0$$

we get

$$\log\frac{1-\pi_i}{\pi_i} = \frac{|x_i|}{\beta} - \frac{|x_i|}{\alpha} - \log\left(\frac{\beta}{\alpha}\frac{1-p}{p}\right)$$

and the minimizing value is given by

$$\widehat{\pi}_i = \frac{1}{1 + \mathrm{e}^{-|x_i|\left(\frac{1}{\alpha}-\frac{1}{\beta}\right)} \frac{\beta}{\alpha}\frac{1-p}{p}} = f(z_i = 1|x_i;\Theta)$$

for which $\frac{\partial^2 J(x,\pi;\Theta)}{\partial \pi_i^2}(\widehat{\pi}_i) - \sum_{i=1}^{n} \frac{\partial^2 H(\pi_i;\Theta)}{\partial \pi_i^2} \geq 0$. From last equality and from Proposition 4, we conclude the thesis.

10 Appendix 3: proof of Theorem 2

In this section, we prove rigorously Theorem 2, which guarantees the convergence of 2-LMM-ISTA to a limit point. We start from the following preliminary results.

Let $V : \mathbb{R}^n \times \Sigma_{n-K} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the function defined in (12)

$$V(x,\pi,\alpha,\beta,\epsilon) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda J_\epsilon(x,\pi;\Theta) - \lambda \sum_{i=1}^{n} H(\pi_i) \tag{22}$$

where $H : [0,1] \to \mathbb{R}$ is the natural entropy function $H(\xi) = -\xi\log\xi - (1-\xi)\log(1-\xi)$. and J_ϵ is defined in (11).

Lemma 3 (Partial minimizations)

$$\widehat{\pi} = \widehat{\pi}(x,\alpha,\beta,\epsilon) = \arg\min_{\pi\in\Sigma_{n-K}} V(x,\pi,\alpha,\beta,\epsilon)$$
$$\widehat{\alpha} = \widehat{\alpha}(x,\pi,\beta,\epsilon) = \arg\min_{\alpha>0} V(x,\pi,\alpha,\beta,\epsilon)$$
$$\widehat{\beta} = \widehat{\beta}(x,\pi,\alpha,\epsilon) = \arg\min_{\beta>0} V(x,\pi,\alpha,\beta,\epsilon)$$

Then, it holds true that

$$\widehat{\alpha} = \frac{\sum_i \pi_i|x_i| + \epsilon}{\|\pi\|_1},$$
$$\widehat{\beta} = \frac{\sum_i (1-\pi_i)|x_i| + \epsilon}{\|\mathbb{1} - \pi\|_1},$$

and

$$\widehat{\pi} = \sigma_{n-K}(a)$$

with

$$a_i = \frac{\frac{(1-p)}{\alpha}\exp\left(-\frac{|x_i|}{\alpha}\right)}{\frac{(1-p)}{\alpha}\exp\left(-\frac{|x_i|}{\alpha}\right) + \frac{p}{\beta}\exp\left(-\frac{|x_i|}{\beta}\right)}$$

Proof By differentiating $V(x,\pi,\alpha,\beta,\epsilon)$ with respect to α and imposing the first-order condition, we obtain

$$\frac{\partial V}{\partial \alpha} = -\frac{\epsilon}{\alpha^2} - \frac{\sum_{i=1}^{n}\pi_i|x_i|}{\alpha^2} + \frac{\sum_{i=1}^{n}\pi_i}{\alpha} = 0$$

from which

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n}\pi_i|x_i| + \epsilon}{\sum_{i=1}^{n}\pi_i} = \frac{\sum_i \pi_i|x_i| + \epsilon}{\|\pi\|_1}.$$

Checking $\frac{\partial^2 V}{\partial \alpha^2}(\widehat{\alpha}) \geq 0$, we conclude that $\widehat{\alpha}$ is the minimizing value of $V(x,\pi,\alpha,\beta,\epsilon)$.

In analogous way, the expression for $\widehat{\beta}$ can be derived.

We now show that $\widehat{\pi} = \sigma_{n-K}(a)$ is the minimizing value of $V(x,\pi,\alpha,\beta,\epsilon)$ for fixed x,α,β,ϵ, i.e., $V(x,\widehat{\pi},\alpha,\beta,\epsilon) \leq V(x,\pi,\alpha,\beta,\epsilon)$ for all $\pi \in \Sigma_{n-K}$.

Let a be the vector satisfying

$$\frac{\partial V}{\partial \pi_i} = |x_i|\left(\frac{1}{\alpha} - \frac{1}{\beta} + \log\frac{\alpha}{\beta}\frac{1-p}{p}\right) - \log\frac{1-\pi_i}{\pi_i} = 0,$$

given by

$$a_i = \frac{\frac{(1-p)}{\alpha}\exp\left(-\frac{|x_i|}{\alpha}\right)}{\frac{(1-p)}{\alpha}\exp\left(-\frac{|x_i|}{\alpha}\right) + \frac{p}{\beta}\exp\left(-\frac{|x_i|}{\beta}\right)}$$

and define $\gamma = -\log(1 - r(a)_{n-K+1})$ where $r(a)$ is the nonincreasing rearrangement of a. It should be noticed that

$$\arg\min_{\pi\in\Sigma_{n-K}} V(x,\widehat{\pi},\alpha,\beta,\epsilon) = \arg\min_{\pi\in\Sigma_{n-K}} V(x,\widehat{\pi},\alpha,\beta,\epsilon) + \gamma\|\pi\|_0 \tag{23}$$

being $\gamma\|\pi\|_0 = (n-K)\gamma$ just a constant for $\pi \in \Sigma_{n-K}$.

The minimum of (23) can be calculated by minimizing with respect to each π_i individually and

$$V(x,\widehat{\pi},\alpha,\beta,\epsilon) + \gamma\|\pi\|_0 = g(\pi_i) + C,$$

where C is independent of π_i and

$$g(\pi_i) = \pi_i|x_i|\left(\frac{1}{\alpha} - \frac{1}{\beta}\right) + \pi_i\log\left(\frac{\alpha}{\beta}\frac{1-p}{p}\right) - H(\pi_i) + \gamma|\pi_i|^0.$$

To derive the minimum, we distinguish two cases, $\pi_i = 0$ and $\pi_i \neq 0$. In the first case, the element-wise cost is (ignoring the constant terms) 0. In the second case, the minimum cost (again ignoring the constant terms) is attained for $\pi_i = a_i$ if $\pi_i \neq 0$. Comparing the cost for both cases, i.e, $g(a_i) < 0$, we obtain

$$\begin{aligned} &g(a_i) = a_i\log\left(\frac{1-a_i}{a_i}\right) - H(a_i) + \gamma < 0 \\ &\log(1-a_i) < -\gamma \\ &a_i > 1 - \mathrm{e}^{-\gamma} \end{aligned}$$

By definition of γ, we get

$$\arg\min_{\pi_i} g(\pi_i) = \begin{cases} a_i & \text{if } a_i > 1 - \mathrm{e}^{-\gamma} = r(a)_{n-K+1} \\ 0 & \text{otherwise.} \end{cases}$$

From this result and the fact that

$$\widehat{\pi}_i = \sigma_{n-K}(a)_i = \begin{cases} a_i & \text{if } a_i \geq r(a)_{n-K+1} \\ 0 & \text{if } a_i < r(a)_{n-K+1} \end{cases} \in \Sigma_{n-K}$$

we conclude that $\widehat{\pi}$ is the minimizing value of V for fixed x,α,β,ϵ. □

Lemma 4 *Define the surrogate functional*

$$V^{\mathcal{S}}(x,a,\pi,\alpha,\beta,\epsilon) = V(x,\pi,\alpha,\beta,\epsilon) + \frac{1}{2\tau}\|x-a\|_2^2 - \frac{1}{2}\|A(x-a)\|_2^2\Big], \tag{24}$$

then

$$\eta_{\lambda\tau\omega}(a + \tau A^\top(y - Aa)) = \arg\min_{x\in\mathbb{R}^n} V^{\mathcal{S}}(x,a,\pi,\alpha,\beta,\epsilon)$$

with $\omega_i = \frac{\pi_i}{\alpha} + \frac{1-\pi_i}{\beta}$.

Proof By developing the least squares in (12) is straightforward to show that

$$V^{\mathcal{S}}(x,a,\pi,\alpha,\beta,\epsilon) = \frac{1}{2}\|x - (a + \tau A^\top(y - Aa))\|_2^2 + \sum_{i=1}^n \omega_i|x_i| + \chi(y,A,a,\pi,\epsilon,\alpha,\beta)$$

where χ is a function independent of x. By differentiating the function with respect x, we obtain the thesis. □

Proposition 5 *The function V defined in (22) is not increasing along the iterates $\zeta^{(t)} = \left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)$.*

Proof From Lemma 3 and 4, it should be noticed that, for each time $t \in \mathbb{N}$, we have

$$\begin{aligned} \alpha^{(t+1)} &= \arg\min_{\alpha} V\left(x^{(t+1)},\pi^{(t+1)},\alpha,\beta^{(t+1)},\epsilon^{(t+1)}\right) \\ \beta^{(t+1)} &= \arg\min_{\beta} V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta,\epsilon^{(t+1)}\right). \end{aligned}$$

Then

$$\begin{aligned} V\left(\zeta^{(t+1)}\right) &= V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t+1)},\beta^{(t+1)},\epsilon^{(t+1)}\right) \\ &\leq V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t+1)},\epsilon^{(t+1)}\right) \\ &\leq V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t+1)}\right). \end{aligned}$$

Since $V(x,\pi,\alpha,\beta,\epsilon)$ is an increasing function in ϵ and being $\epsilon^{(t+1)} = \min\{\epsilon^{(t)},\theta^{(t+1)}\} \leq \epsilon^{(t)}$ by definition, we obtain

$$V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t+1)}\right) \leq V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)$$

and therefore, using

$$\pi^{(t+1)} = \arg\min_{\pi\in\Sigma_{n-K}} V\left(x^{(t+1)},\pi,\alpha^{(t)},\beta^{(t)},\epsilon^{(t+1)}\right).$$

see Lemma 3, we get

$$\begin{aligned} V\left(\zeta^{(t+1)}\right) &\leq V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\leq V\left(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right). \end{aligned}$$

It should be noticed that for all x

$$V^{\mathcal{S}}(x,x,\pi,\alpha,\beta,\epsilon) = V(x,\pi,\alpha,\beta,\epsilon)$$

and

$$V^{\mathcal{S}}(x,a,\pi,\alpha,\beta,\epsilon) \geq V(x,x,\pi,\alpha,\beta,\epsilon)$$

for all $a \neq x$. Then, we have

$$\begin{aligned} V\left(\zeta^{(t+1)}\right) &\leq V\left(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &= V^{\mathcal{S}}\left(x^{(t+1)},x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\leq V^{\mathcal{S}}\left(x^{(t+1)},x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\leq V^{\mathcal{S}}\left(x^{(t)},x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\leq V\left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \leq V\left(\zeta^{(t)}\right). \end{aligned}$$

□

The following lemma implies that these algorithms converge numerically when the number of iterations goes to infinity.

Lemma 5 *Let $(x^{(t)})$ be the sequence generated by 2-LMM-ISTA, then $x^{(t+1)} - x^{(t)} \to 0$ as $t \to \infty$.*

Proof If $\alpha^{(t)} \to 0$ or $\beta^{(t)} \to 0$ as $t \to \infty$, we have $\epsilon^{(t)} \to 0$ and, by definition of $\epsilon^{(t)} = \min\left\{\epsilon^{(t-1)}, \theta^{(t)}\right\}$ and $\theta^{(t)} = \frac{1}{\log(t+1)} + c\|x^{(t)} - x^{(t-1)}\| + r(x^{(t)})_{K+1}/n$, we get

$$\lim_{t\to\infty} c\left\|x^{(t+1)} - x^{(t)}\right\|_2 = 0$$

and the assertion is true.

If instead neither $\alpha^{(t)}$ nor $\beta^{(t)}$ converge to zero, then there exists a constant $\tau > 0$ and a sequence of integers $\{T_\ell\} : \mathbb{N} \to \mathbb{N}$ such that $T_\ell \to \infty$, as $\ell \to \infty$ and $\min\left\{\alpha^{(T_\ell)}, \beta^{(T_\ell)}\right\} > \chi$ for all $\ell \in \mathbb{N}$. It holds true in general that from Proposition 5, we have

$$\begin{aligned} &\lambda\left[\sum_{i=1}^n \pi_i^{(t)} \log\alpha^{(t)} + \sum_{i=1}^n \left(1-\pi_i^{(t)}\right)\log\beta^{(t)} - n\log 2\right] \\ &\quad\leq \frac{1}{2}\left\|y - Ax^{(t)}\right\|_2^2 + \lambda\left[\sum_{i=1}^n \pi_i^{(t)}\log\alpha^{(t)}\right. \\ &\qquad + \sum_{i=1}^n \left(1-\pi_i^{(t)}\right)\log\beta^{(t)} - \sum_{i=1}^n H(\pi_i) \\ &\qquad \left. - \sum_{i=1}^n \pi_i^{(t)}\log(1-p) - \sum_{i=1}^n (1-\pi_i^{(t)})\log p\right] \\ &\quad\leq V\left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\quad\leq V\left(x^{(1)},\pi^{(1)},\alpha^{(1)},\beta^{(1)},\epsilon^{(1)}\right). \end{aligned} \tag{25}$$

Then, we have that also for the subsequence T_ℓ it holds

$$\begin{aligned} &V\left(x^{(T_\ell)},\pi^{(T_\ell)},\alpha^{(T_\ell)},\beta^{(T_\ell)},\epsilon^{(T_\ell)}\right) \\ &\geq \lambda\left[\sum_{i=1}^n \pi_i^{(T_\ell)}\log\alpha^{(T_\ell)} + \sum_{i=1}^n \left(1-\pi_i^{(T_\ell)}\right)\log\beta^{(T_\ell)} - n\log 2\right] \\ &\geq \lambda(n\log\chi - n\log 2). \end{aligned} \tag{26}$$

Since $\tau < \|A\|_2^{-2}$, we have

$$\begin{aligned} 0 &\leq \frac{1}{2\tau}\left(1-\tau\|A\|^2\right)\left\|x^{(t)} - x^{(t+1)}\right\|^2 \\ &\leq \frac{1}{2\tau}\left(x^{(t)} - x^{(t+1)}\right)^\top\left(I - \tau A^\top A\right)\left(x^{(t)} - x^{(t+1)}\right) \\ &= V^{\mathcal{S}}\left(x^{(t+1)},x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \\ &\quad - V\left(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right) \end{aligned} \tag{27}$$

Consider the following sum for all $\ell \in \mathbb{N}$

$$\begin{aligned} 0 &\leq \sum_{t=1}^{T_\ell} \frac{1}{2\tau}\left(x^{(t)} - x^{(t+1)}\right)^\top\left(I - \tau A^\top A\right)\left(x^{(t)} - x^{(t+1)}\right) \\ &\overset{(a)}{=} \sum_{t=1}^{T_\ell}\left[V^{\mathcal{S}}\left(x^{(t+1)},x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right. \\ &\quad \left. - V\left(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right] \\ &\overset{(b)}{\leq} \sum_{t=1}^{T_\ell}\left[V^{\mathcal{S}}\left(x^{(t)},x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right. \\ &\quad \left. - V\left(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right] \\ &\overset{(c)}{\leq} \sum_{t=1}^{T_\ell}\left[V\left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right. \\ &\quad \left. - V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right] \\ &\overset{(d)}{\leq} \sum_{t=1}^{T_\ell}\left[V\left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right. \\ &\quad \left. - V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t+1)}\right)\right] \\ &\overset{(e)}{\leq} \sum_{t=1}^{T_\ell}\left[V\left(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}\right)\right. \\ &\quad \left. - V\left(x^{(t+1)},\pi^{(t+1)},\alpha^{(t+1)},\beta^{(t+1)},\epsilon^{(t+1)}\right)\right] \\ &\overset{(f)}{=} V\left(x^{(1)},\pi^{(1)},\alpha^{(1)},\beta^{(1)},\epsilon^{(1)}\right) \\ &\quad - V\left(x^{(T_\ell+1)},\pi^{(T_\ell+1)},\alpha^{(T_\ell+1)},\beta^{(T_\ell+1)},\epsilon^{(T_\ell+1)}\right) \\ &\overset{(g)}{\leq} V\left(x^{(1)},\pi^{(1)},\alpha^{(1)},\beta^{(1)},\epsilon^{(1)}\right) - \lambda(n\log\chi - n\log 2) \\ &= C' \end{aligned}$$

where

(a) Follows from (25)
(b) Follows from the fact that

$$x^{(t+1)} = \arg\min_{x\in\mathbb{R}^n} V^{\mathcal{S}}\left(x, x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}\right)$$

(see Lemma 4)
(c) Is a consequence of the following relations

$$V^{\mathcal{S}}\left(x^{(t)}, x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}\right) = V\left(x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}\right)$$

and

$$\pi^{(t+1)} = \arg\min_{\pi\in\Sigma_{n-K}} V\left(x^{(t+1)}, \pi, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}\right)$$

(see Lemma 3);
(d) and (e) Are following from

$$\alpha^{(t+1)} = \arg\min_{\alpha} V\left(x^{(t+1)}, \pi^{(t+1)}, \alpha, \beta^{(t)}, \epsilon^{(t)}\right),$$

$$\beta^{(t+1)} = \arg\min_{\beta} V\left(x^{(t+1)}, \pi^{(t+1)}, \alpha^{(t+1)}, \beta, \epsilon^{(t)}\right)$$

(see Lemma 3) and from the fact that $V(x,\pi,\alpha,\beta,\epsilon)$ is an increasing function in ϵ and $\epsilon^{(t+1)} \le \epsilon^{(t)}$ by definition.

We conclude that for all $\ell \in \mathbb{N}$

$$\sum_{t=1}^{T_\ell} \frac{1}{2\tau}\left(x^{(t)} - x^{(t+1)}\right)^\top \left(I - \tau A^\top A\right)\left(x^{(t)} - x^{(t+1)}\right) \le C'.$$

By letting $\ell \to \infty$, we obtain that the series is convergent and, we obtain the necessary condition

$$0 \le \frac{1}{2\tau}\left(x^{(t)} - x^{(t+1)}\right)^\top \left(I - \tau A^\top A\right)\left(x^{(t)} - x^{(t+1)}\right) \to 0$$

as $t \to \infty$ and from inequality in (27) the assertion is proved.

(f) Follows by noticing that the series is telescopic
(g) Is a direct consequence bound in (26)

□

Lemma 6 *The sequence* $(x^{(t)})_{t\in\mathbb{N}}$ *is bounded*

Proof We now prove that both $\alpha^{(t)}$ and $\beta^{(t)}$ must be upper bounded. From (25), there exists a constant

$$C = \frac{V\left(x^{(1)}, \pi^{(1)}, \alpha^{(1)}, \beta^{(1)}, \epsilon^{(1)}\right)}{\lambda} + n\log 2$$

such that

$$\sum_{i=1}^{n}\left(1-\pi_i^{(t)}\right)\log\beta^{(t)} + \sum_{i=1}^{n}\pi_i^{(t)}\log\alpha^{(t)} \le C. \tag{28}$$

Suppose ad absurdum that $\beta^{(t)}$ is unbounded (similar consideration can be done if $\alpha^{()}$ is unbounded), then there exists a sequence $t_\ell : \mathbb{N} \to \mathbb{N}$ such that $\beta^{(t_\ell)} \to \infty$ as $\ell \to \infty$. By (28), we have $\alpha^{(t_\ell)} \to 0$. In fact, since $\pi^{t_\ell} \in \Sigma_{n-K}$ and $\beta^{(t_\ell)} > 1$ definitively (by unboundeness), we have

$$\begin{aligned} K\log\beta^{(t_\ell)} &+ \sum_{i\in\mathrm{supp}(\pi^{(t_\ell)})} \pi_i^{(t_\ell)}\log\alpha^{(t_\ell)} \\ &\le \sum_{i=1}^{n}\left(1-\pi_i^{(t)}\right)\log\beta^{(t)} + \sum_{i=1}^{n}\pi_i^{(t)}\log\alpha^{(t)} \\ &\le C \end{aligned}$$

and the inequality is satisfied if and only if $\alpha^{(t_\ell)} \to 0$. Consequently, by definition of $\alpha^{(t_\ell)}$, also $\epsilon^{(t_\ell)} \to 0$ and

$$r\left(x^{(t^\ell)}\right)_{K+1} \to 0, \tag{29}$$

where we recall that $r(x^{(t^\ell)})$ is the nonincreasing rearrangement of $x^{(t_\ell)}$. Let $\Delta := \{i \in [n] : \exists \epsilon > 0$ and $(t_\ell)_{q\in\mathbb{N}}$ for which $|x_i^{(t_\ell)}| > \epsilon\}$, then from (29), we have $|\Delta| \le K$. Since $|x_i| > |x_j|$ for all $i \in \Delta$ and $j \in \Delta^{\mathrm{c}}$, then

$$\begin{aligned} a_i^{(t_\ell)} &= \left(1 + \frac{p}{1-p}\frac{\alpha^{(t_\ell-1)}}{\beta^{(t_{\ell_1}-1)}}\mathrm{e}^{|x_i^{(t_\ell)}|\left(\frac{1}{\alpha^{(t_\ell-1)}} - \frac{1}{\beta^{(t_\ell-1)}}\right)}\right)^{-1} \\ &\ge a_j^{(t_\ell)} \end{aligned}$$

The application of hard thresholding yields $\pi^{(t_\ell)} = \sigma_{n-K}\left(a^{(t_\ell)}\right)$, and we have $\pi_i^{(t_\ell)} = 0$ and $\omega_i^{(t_\ell+1)} = 1/\beta^{(t_\ell)} \overset{q\to\infty}{\longrightarrow} 0$ for all $i \in \Delta$.

If $i \in \Delta^{\mathrm{c}}$, then $|x_i^{t_\ell}| \to 0$ as $\ell \to \infty$ and, consequently, from Lemma 5 also $|x_i^{t_\ell+1}| \to 0$ as $\ell \to \infty$.

Let now $\Lambda = \mathrm{supp}(x^\star)$. We thus have

$$\begin{aligned} \left\|x^{(t_\ell+1)} - x^\star\right\|_2^2 &= \left\|x_\Delta^{(t_\ell+1)} - x_\Delta^\star\right\|_2^2 + \left\|x_{\Delta^{\mathrm{c}}}^{(t_\ell+1)} - x_{\Delta^{\mathrm{c}}}^\star\right\|_2^2 \\ &\le \left\|\left(I - \tau A_\Delta^\top A\right)\left(x^{(t_\ell)} - x^\star\right)\right\|_2^2 \\ &\quad + K\lambda\tau\max_{i\in\Delta}\omega_i^{(t_\ell+1)} + \left\|x_{\Delta^{\mathrm{c}}}^{(t_\ell+1)} - x_{\Delta^{\mathrm{c}}}^\star\right\|_2^2. \end{aligned}$$

where the last inequality follows from the triangular inequality.

$$\begin{aligned} \left\|x^{(t_\ell+1)} - x^\star\right\|_2^2 &\le \left\|I - \tau A_\Delta^\top A_\Delta\right\|_2^2 \left\|x_\Delta^{(t_\ell)} - x_\Delta^\star\right\|_2^2 \\ &\quad + \left\|\tau A_\Delta^\top A_{\Delta^{\mathrm{c}}}\right\|_2^2 \left\|x_{\Delta^{\mathrm{c}}}^{(t_\ell)} - x_{\Delta^{\mathrm{c}}}^\star\right\|_2^2 \\ &\quad + K\lambda\tau\max_{i\in\Delta}\omega_i^{(t_\ell+1)} + \left\|x_{\Delta^{\mathrm{c}}}^{(t_\ell+1)} - x_{\Delta^{\mathrm{c}}}^\star\right\|_2^2. \end{aligned}$$

Since $|\Delta| \le K$ and the columns of A associated with Δ are linearly independent, then the matrix $A_\Delta^\top A_\Delta$ is nonsingular and $\left\|I - \tau A_\Delta^\top A_\Delta\right\|_2^2 = \gamma < 1$. As terms $\|x_{\Delta^{\mathrm{c}}}^{(t_\ell)} - x_{\Delta^{\mathrm{c}}}^\star\|_2^2$, $\|x_{\Delta^{\mathrm{c}}}^{(t_\ell+1)} - x_{\Delta^{\mathrm{c}}}^\star\|_2^2$, and $\max_{i\in\Delta}\omega_i^{(t_\ell+1)}$ are going to

zero when $\ell \to \infty$, there exists $t_0 \in \mathbb{N}$ and a constant $\chi \in \mathbb{R}$ such that if $t > t_0$ then

$$\left\| x^{(t_\ell+1)} - x^\star \right\|_2 \leq \gamma \left\| x^{(t_\ell)} - x^\star \right\|_2 + \chi$$

Iterating the argument and letting $\ell \to \infty$,

$$\lim_{\ell\to\infty} \left\| x^{(t_\ell+1)} - x^\star \right\|_2 \leq \frac{\chi}{1-\gamma}$$

and we conclude that the sequence $(x^{(t_\ell)})_{\ell\in\mathbb{N}}$ is bounded and so is $(\beta^{(t_\ell)})$ from which we get the contradiction. We conclude that $(\alpha^{(t)})_{t\in\mathbb{N}}$ and $(\beta^{(t)})_{t\in\mathbb{N}}$ are both upper bounded by a constant $\chi > 0$ and so $(x^{(t)})_{t\in\mathbb{N}}$:

$$0 \leq 2\epsilon^{(t)} + \left\| x^{(t)} \right\|_1 = \sum_{i=1}^n \pi_i \alpha^{(t)} + \sum_{i=1}^n (1-\pi_i)\beta^{(t)} \leq \chi n$$

□

Proposition 6 *Any accumulation point is a τ-stationary point of* (12) *of the algorithm and satisfies the equalities in* (19a)-(19e).

Proof Suppose that $\left(x^\sharp, \pi^\sharp, \alpha^\sharp, \beta^\sharp\right)$ is an accumulation point of the sequence $\left(x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}\right)_{t\in\mathbb{N}}$. Then, there exists a subsequence $\left(x^{(t_\ell)}, \pi^{(t_\ell)}, \alpha^{(t_\ell)}, \beta^{(t_\ell)}\right)_{\ell\in\mathbb{N}}$ that converges to $\left(x^\sharp, \pi^\sharp, \alpha^\sharp, \beta^\sharp\right)$ as $\ell \to \infty$. We now show (19d) and we let the reader verify the other conditions by continuity. Since $\lim_{\ell\to\infty} \pi^{(t_\ell)} = \pi^\sharp$, then there exists ℓ_0 such that, $\forall \ell > \ell_0$ and $\forall i \in \operatorname{supp}\left(\pi^\sharp\right)$, $\pi_i^{(t_\ell)} \neq 0$ and

$$\pi_i^{(t_\ell)} = a_i^{(t_\ell)} \overset{\ell\to\infty}{\to} \frac{\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right)\frac{1-p}{\alpha^\sharp}}{\frac{1-p}{\alpha^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right) + \frac{p}{\beta^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\beta^\sharp}\right)} = \pi_i^\sharp.$$

If $i \notin \operatorname{supp}\left(\pi^\sharp\right)$ we distinguish the following cases.

(a) There exists a sequence $\left(\ell_q\right)_{q\in\mathbb{N}}$ such that $0 \neq \pi_i^{\left(t_{\ell_q}\right)} = a_i^{\left(t_{\ell_q}\right)} \to \pi_i^\sharp = 0$. This means that there exists $q_0 \in \mathbb{N}$ such that $\forall q > q_0$ we have $a_i^{(t_{\ell_q})} < a_j^{\left(t_{\ell_q}\right)}$, $\forall j \in \operatorname{supp}\left(\pi^\sharp\right)$ and, by letting $q \to \infty$,

$$\begin{aligned} &\frac{\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right)\frac{1-p}{\alpha^\sharp}}{\frac{1-p}{\alpha^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right) + \frac{p}{\beta^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\beta^\sharp}\right)} \\ &\leq \frac{\exp\left(-\frac{|x_j^\sharp|}{\alpha^\sharp}\right)\frac{1-p}{\alpha^\sharp}}{\frac{1-p}{\alpha^\sharp}\exp\left(-\frac{|x_j^\sharp|}{\alpha^\sharp}\right) + \frac{p}{\beta^\sharp}\exp\left(-\frac{|x_j^\sharp|}{\beta^\sharp}\right)} \end{aligned} \tag{30}$$

(b) There exists $\ell_0 \in \mathbb{N}$ such that $\forall \ell > \ell_0\ \pi_i^{(t_\ell)} = 0$, from which $a_i^{(t_\ell)} < a_j^{(t_\ell)}$, $\forall j \in \operatorname{supp}\left(\pi^\sharp\right)$ and by letting $\ell \to \infty$ we get (30).

We conclude that for all $i \notin \operatorname{supp}(\pi^\infty)$

$$\pi_i^\sharp = 0 = \sigma_{n-K}\left(\frac{\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right)\frac{1-p}{\alpha^\sharp}}{\frac{1-p}{\alpha^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\alpha^\sharp}\right) + \frac{p}{\beta^\sharp}\exp\left(-\frac{|x_i^\sharp|}{\beta^\sharp}\right)}\right)_i.$$

□

Proof of Theorem 2: From Lemma 6 the sequence $\left(x^{(t)}\right)$ is bounded and by the Bolzano Weierstrass Theorem, there exists a subsequence $\left(x^{t_j}\right)_{j\in\mathbb{N}}$ such that

$$\lim_{j\to\infty} x^{(t_j)} = x^\infty \tag{31}$$

$\lim_{j\to\infty} \epsilon^{(t_j)} = \epsilon^\infty$, $\lim_{j\to\infty} \alpha^{(t_j)} = \alpha^\infty$, $\lim_{j\to\infty} \beta^{(t_j)} = \beta^\infty$, and $\lim_{j\to\infty} \pi^{(t_j)} = \pi^\infty$. We thus have

$$\begin{aligned} \lim_{j\to\infty} \left\| x^{(t_j+1)} - x^\infty \right\| &\leq \lim_{j\to\infty} \left\| x^{(t_j+1)} - x^{(t_j)} + x^{(t_j)} - x^{(t_j)} - x^\infty \right\| \\ &\leq \lim_{j\to\infty} \left\| x^{(t_j+1)} - x^{(t_j)} \right\| + \left\| x^{(t_j)} - x^\infty \right\| \\ &= 0 \end{aligned}$$

where the second inequality follows from triangular inequality and the last equality follows from Lemma 5 and (31). Since $\lim_{t\to\infty} x^{(t_j+1)} = \lim_{t\to\infty} x^{(t_j)} = x^\infty$, by continuity, we get $\lim_{t\to\infty} a^{(t_j+1)} = \lim_{t\to\infty} a^{(t_j)} = a^\infty$. This implies that $S := \operatorname{supp}\left(\pi^{(t_j+1)}\right) = \operatorname{supp}\left(\pi^{(t_j)}\right)$ definitely and we deduce that

$$\lim_{j\to\infty} \pi_i^{(t_j+1)} = 0 = \lim_{j\to\infty} \pi_i^{(t_j)} = \pi_i^\infty, \forall i \notin S$$

and

$$\lim_{j\to\infty} \pi_i^{(t_j+1)} = \lim_{j\to\infty} a_i^{(t_j+1)} = \lim_{j\to\infty} a_i^{(t_j)} = \pi_i^\infty, \forall i \in S.$$

Moreover

$$\lim_{j\to\infty} \epsilon^{(t_{j+1})} = \lim_{j\to\infty} \epsilon^{(t_j)} = \epsilon^\infty$$

($\epsilon^{(t)}$ is positive and monotonic). By continuity also $\alpha^{(t_j+1)} \to \alpha^\infty$ and $\beta^{(t_j+1)} \to \beta^\infty$ as $j \to \infty$. We conclude by induction that the sequence generated by Algorithm 1 converges to $(x^\infty, \pi^\infty, \alpha^\infty, \beta^\infty)$, which is also a τ-stationary point of (12) by Proposition 6.

11 Appendix 4: Proof of Theorem 3:

From Theorem 2, Algorithm 1 converges to $(x^\infty, \pi^\infty, \alpha^\infty, \beta^\infty)$ which is also a τ-stationary point of (12). Let $\Lambda^\infty = \operatorname{supp}(x^\infty)$ and $k^\infty = |\Lambda^\infty| \leq K$ by assumption $r(x^\infty)_{K+1} = 0$. We thus have

$$a_i^\infty := \frac{1}{1 + \frac{\alpha^\infty}{\beta^\infty}\frac{p}{1-p}\exp\left(|x_i^\infty|\left(\frac{1}{\alpha^\infty} - \frac{1}{\beta^\infty}\right)\right)}$$

$$\begin{cases} = \frac{1}{1+\frac{\alpha^\infty}{\beta^\infty}\frac{p}{1-p}} & \text{for } i \in (\Lambda^\infty)^c \\ > \frac{1}{1+\frac{\alpha^\infty}{\beta^\infty}\frac{p}{1-p}} & \text{for } i \in (\Lambda^\infty) \end{cases}$$

Since $k^\infty \leq K$ then from (19d)

$$\pi_i^\infty = \sigma_{n-K}(a^\infty)_i = \begin{cases} \frac{1}{1+\frac{\alpha^\infty}{\beta^\infty}\frac{p}{1-p}} & \text{for } i \in T \\ 0 & \text{for } i \in T^c \end{cases}$$

for some $T \supseteq \Lambda^\infty$ with $|T| = K$. We deduce from (19e) that

$$\alpha^\infty = \frac{\sum_{i\in\Lambda^\infty} \pi_i^\infty |x_i|}{\sum_{i\in[n]} \pi_i} = 0$$

and $\pi_i^\infty = 1$ for all $i \in T$ and

$$\beta^\infty = \frac{\sum_{i\in\Lambda^\infty}(1-\pi_i^\infty)|x_i|}{\sum_{i\in[n]}(1-\pi_i^\infty)} = \frac{\|x\|_1}{K}.$$

Let $e := x^\star - x^\infty$. Since x^∞ is a τ-stationary point, it should be noticed that

$$\begin{aligned} \|e_{\Lambda^\infty}\| &\leq \left\| \left(A_{\Lambda^\infty}^\top A_{\Lambda^\infty}\right)^{-1} \right\| \left\| \frac{\lambda}{\beta^\infty} \operatorname{sgn}(x_{\Lambda^\infty}^\infty) \right\| \\ &\quad + \left\| \left(A_{\Lambda^\infty}^\top A_{\Lambda^\infty}\right)^{-1} \right\| \left\| A_{\Lambda^\infty}^\top A_{(\Lambda^\infty)^c} e_{(\Lambda^\infty)^c} \right\| \\ &\leq \left\| \left(A_{\Lambda^\infty}^\top A_{\Lambda^\infty}\right)^{-1} \right\| \frac{\lambda\sqrt{K}}{\beta^\infty} \\ &\quad + \left\| A_{\Lambda^\infty}^\top A_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| \left\| e_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| \\ &\leq \frac{1}{1-\delta_{2K}} \left(\frac{\lambda\sqrt{K}}{\beta^\infty} + \delta_{2K} \left\| e_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| \right). \end{aligned} \tag{32}$$

We have

$$\begin{aligned} \left\| A_{(\Lambda^\infty)^c\cap\Lambda^\star}^\top (y - Ax) \right\| &\leq \left\| \sigma_K \left(A_{(\Lambda^\infty)^c}^\top (y - Ax) \right) \right\| \\ &\leq \left\| A_{\Lambda^\infty}^\top (y - Ax) \right\| = c \end{aligned} \tag{33}$$

where the last inequality follows from hypothesis. Moreover, from the triangular inequality

$$\begin{aligned} &\left\| A_{(\Lambda^\infty)^c\cap\Lambda^\star}^\top (y - Ax) \right\| \\ &\quad \geq \left\| A_{(\Lambda^\infty)^c\cap\Lambda^\star}^\top A_{(\Lambda^\infty)^c\cap\Lambda^\star} e_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| \\ &\quad - \left\| A_{(\Lambda^\infty)^c\cap\Lambda^\star}^\top A_{\Lambda^\infty} e_{\Lambda^\infty} \right\| \\ &\quad \geq (1-\delta_{2K}) \left\| e_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| - \delta_{2K} \|e_{\Lambda^\infty}\| \end{aligned} \tag{34}$$

Combining (33) and (34), we get

$$\left\| e_{(\Lambda^\infty)^c\cap\Lambda^\star} \right\| \leq \frac{c}{1-\delta_{2K}} + \frac{\delta_{2K}}{1-\delta_{2K}} \|e_{\Lambda^\infty}\|.$$

Using (34) and (32), we obtain

$$\|e_{\Lambda^\infty}\| \leq \frac{1}{1-2\delta_{2K}} \left(\frac{\lambda(1-\delta_{2K})\sqrt{K}}{\beta^\infty} + c\delta_{2K} \right).$$

Abbreviations

2-LMM: Two-component Laplace mixture model; AMP: Approximate message passing; BCS: Bayesian compressive sensing; BP: Basis pursuit; CS: Compressed sensing; DCT: Discrete cosine transform; IRL1: Iterative reweighting ℓ_1 minimization; ISTA: Iterative shrinkage-thresholding algorithm; FISTA: Fast iterative shrinkage-thresholding algorithm; MAP: Maximum a posteriori; MCMC: Markov chain Monte Carlo; MSE: Mean square error; TS-BCS: Tree-structure Bayesian compressive sensing; WBCS: Wavelet-based Bayesian compressive sensing

Acknowledgements

The authors thank the European Research Council for the financial support for this research. They express their sincere gratitude to the Reviewers for carefully reading the manuscript and for their valuable suggestions.

Funding

This work has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 279848.

Availability of data and materials

Please contact the corresponding author (chiara.ravazzi@ieiit.cnr.it) for data requests.

Authors' contributions

Both authors contributed equally, read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]National Research Council of Italy, IEIIT-CNR, c/o Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. [2]Politecnico di Torino, DET, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

Received: 22 December 2017 Accepted: 13 June 2018
Published online: 13 July 2018

References

1. DL Donoho, Compressed sensing. IEEE Trans. Inform. Theory. **52**, 1289–1306 (2006)
2. AK Fletcher, S Rangan, VK Goyal, K Ramchandran, Denoising by sparse approximation: error bounds based on rate-distortion theory. EURASIP J. Adv. Sig. Process. **2006**(1), 026318 (2006). https://doi.org/10.1155/ASP/2006/26318
3. MA Davenport, MF Duarte, YC Eldar, G Kutyniok, *Introduction to Compressed Sensing*. (Cambridge University Press, Cambridge, 2012)
4. EJ Candès, JK Romberg, T Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pur. Appl. Math. **59**(8), 1207–1223 (2006)
5. R Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B. **58**, 267–288 (1994)
6. I Daubechies, M Defrise, C De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. **57**(11), 1413–1457 (2004)
7. ET Hale, W Yin, Y Zhang, A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. Technical report, Rice University (2007)
8. SJ Wright, RD Nowak, MAT Figueiredo, Sparse reconstruction by separable approximation. Trans. Sig. Proc. **57**(7), 2479–2493 (2009)
9. A Beck, M Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Img. Sci. **2**(1), 183–202 (2009)
10. M Fornasier, *Theoretical Foundations and Numerical Methods for Sparse Recovery*. (Radon Series on Computational and Applied Mathematics, Linz, 2010)
11. W Guo, W Yin, Edge guided reconstruction for compressive imaging. SIAM J. Imaging Sci. **5**(3), 809–834 (2012). https://doi.org/10.1137/110837309

12. W Yin, S Osher, D Goldfarb, J Darbon, Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imaging Sci. **1**(1), 143–168 (2008). https://doi.org/10.1137/070703983
13. A Maleki, DL Donoho, Optimally tuned iterative reconstruction algorithms for compressed sensing. IEEE J. Sel. Top. Sig. Process. **4**(2), 330–341 (2010). https://doi.org/10.1109/JSTSP.2009.2039176
14. DL Donoho, A Maleki, A Montanari, in *Proc. of 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*. Message passing algorithms for compressed sensing: I. motivation and construction, (2010), pp. 1–5
15. T Goldstein, C Studer, R Baraniuk, A field guide to forward-backward splitting with a FASTA implementation, 1–24 (2014). arXiv eprint. https://arxiv.org/abs/1411.3406
16. CM Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. (Springer, Secaucus, 2006)
17. R Gribonval, V Cevher, ME Davies, Compressible distributions for high-dimensional statistics. IEEE Trans. Inf. Theory. **58**(8), 5016–5034 (2012)
18. J Zhang, Y Li, Z Yu, Z Gu, Noisy sparse recovery based on parameterized quadratic programming by thresholding. EURASIP J. Adv. Sig. Proc. **2011** (2011). https://doi.org/10.1155/2011/528734
19. Y Wang, W Yin, Sparse signal reconstruction via iterative support detection. SIAM J. Img. Sci. **3**(3), 462–491 (2010)
20. C Hegde, MF Duarte, V Cevher, in *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Representations (SPARS)*. Compressive Sensing Recovery of Spike Trains Using Structured Sparsity, (Saint Malo, 2009). https://infoscience.epfl.ch/record/151471/files/Compressive%20sensing%20recovery%20of%20spike%20trains%20using%20a%20structured%20sparsity%20model.pdf
21. RG Baraniuk, V Cevher, MF Duarte, C Hegde, Model-based compressive sensing. IEEE Trans. Inf. Theor. **56**(4), 1982–2001 (2010)
22. R von Borries, CJ Miosso, C Potes, in *Computational Advances in Multi-Sensor Adaptive Processing, 2007. CAMPSAP 2007. 2nd IEEE International Workshop On, Compressed sensing using prior information.* (2007), pp. 121–124
23. N Vaswani, W Lu, in *Information Theory, 2009. ISIT 2009. IEEE International Symposium On, Modified-CS: Modifying compressive sensing for problems with partially known support.* (2009), pp. 488–492
24. OD Escoda, L Granai, P Vandergheynst, On the use of a priori information for sparse signal approximations. IEEE Trans. Sig. Process. **54**(9), 3468–3482 (2006)
25. MA Khajehnejad, W Xu, AS Avestimehr, B Hassibi, Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models. IEEE Trans. Sig. Process. **59**(5), 1985–2001 (2011)
26. MP Friedlander, H Mansour, R Saab, O Yilmaz, Recovering compressively sampled signals using partial support information. IEEE Trans. Inf. Theory. **58**(2), 1122–1134 (2012)
27. J Scarlett, JS Evans, S Dey, Compressed sensing with prior information: information-theoretic limits and practical decoders. IEEE Trans. Sig. Process. **61**(2), 427–439 (2013)
28. MA Khajehnejad, W Xu, AS Avestimehr, B Hassibi, Analyzing Weighted ℓ_1 Minimization for Sparse Recovery With Nonuniform Sparse Models. IEEE Trans. Sig. Process. **59**, 1985–2001 (2011)
29. JA Tropp, AC Gilbert, MJ Strauss, Algorithms for simultaneous sparse approximation: part i: greedy pursuit. Sig. Process. **86**(3), 572–588 (2006)
30. JA Tropp, AC Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory. **53**, 4655–4666 (2007)
31. D Needell, R Vershynin, Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. IEEE J. Sel. Top. Signal. Process. **4**(2), 310–316 (2010)
32. D Needell, JA Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2008). 0803.2392
33. EJ Candès, The restricted isometry property and its implications for compressed sensing. Compte Rendus de l'Academie des Sciences. Paris, France, ser. I. **346**, 589–592 (2008)
34. R Baraniuk, M Davenport, D DeVore, M Wakin, A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
35. M Bayati, A Montanari, The LASSO risk for Gaussian matrices. IEEE Trans. Inf. Theory. **58**(4), 1997–2017 (2012)
36. S Ji, Y Xue, L Carin, Bayesian compressive sensing. IEEE Trans. Signal Process. **56**(6), 2346–2356 (2008)
37. EP Simoncelli, Modeling the joint statistics of images in the wavelet domain. Proc. SPIE. **3813**, 188–195 (1999). https://doi.org/10.1117/12.366779
38. MAT Figueiredo, R Nowak, Wavelet-based image estimation: an empirical bayes approach using Jeffreys' noninformative prior. IEEE Trans. Image Process. **10**(9), 1322–1331 (2001)
39. H-Y Gao, Wavelet shrinkage denoising using the non-negative garrote. J. Comput. Graph. Stat. **7**(4), 469–488 (1998). https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474789
40. L Sendur, IW Selesnick, Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. Trans. Sig. Proc. **50**(11), 2744–2756 (2002). https://doi.org/10.1109/TSP.2002.804091
41. DL Donoho, De-noising by soft-thresholding. IEEE Trans. Inf. Theor. **41**(3), 613–627 (1995). https://doi.org/10.1109/18.382009
42. PL Combettes, VR Wajs, Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
43. K Bredies, D Lorenz, Linear convergence of iterative soft-thresholding. J. Fourier Anal. Appl. **14**(5-6), 813–837 (2008). https://doi.org/10.1007/s00041-008-9041-1
44. A Montanari, in *Compressed Sensing: Theory and Applications*, ed. by Y Eldar, G Kutyniok. Graphical models concepts in compressed sensing (Cambridge University Press, Cambridge, 2012), pp. 394–438
45. EJ Candes, MB Wakin, SP Boyd, Enhancing sparsity by reweighted L1 minimization. Technical report
46. MP Friedlander, H Mansour, R Saab, O Yilmaz, Recovering compressively sampled signals using partial support information. IEEE Trans. Inf. Theory. **58**(2), 1122–1134 (2012)
47. MA Khajehnejad, W Xu, AS Avestimehr, B Hassibi, in *Proceedings of the 2009 IEEE International Conference on Symposium on Information Theory - Volume 1, ISIT'09*. Weighted L1 Minimization for Sparse Recovery with Prior Information (IEEE Press, Piscataway, 2009), pp. 483–487. http://dl.acm.org/citation.cfm
48. H Mansour, in *2012 IEEE Statistical Signal Processing Workshop (SSP) (SSP'12). Beyond ℓ_1 norm minimization for sparse signal recovery* (IEEE, Ann Arbor, 2012). IEEE
49. JP Vila, P Schniter, Expectation-maximization Gaussian-mixture approximate message passing. IEEE Trans. Sig. Process. **61**(19), 4658–4672 (2013)
50. C Guo, ME Davies, Near optimal compressed sensing without priors: parametric sure approximate message passing. IEEE Trans. Sig. Process. **63**(8), 2130–2141 (2015)
51. C Ravazzi, SM Fosson, E Magli, Randomized algorithms for distributed nonlinear optimization under sparsity constraints. IEEE Trans. Sig. Process. **64**(6), 1420–1434 (2016)
52. MJ Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inf. Theory. **55**(5), 2183–2202 (2009)
53. C Chen, J Huang, L He, H Li, Fast iteratively reweighted least squares algorithms for analysis-based sparsity reconstruction. CoRR. 1–14 (2014). https://arxiv.org/abs/1411.5057
54. C Ravazzi, E Magli, Gaussian mixtures based IRLS for sparse recovery with quadratic convergence. IEEE Trans. Sig. Process. **63**(13), 1–16 (2015)
55. A Maleki. Approximate message passing algorithms for compressed sensing (Stanford University PhD thesis, 2011). https://www.ece.rice.edu/~mam15/suthesis-Arian.pdf
56. M Raginsky, S Jafarpour, ZT Harmany, RF Marcia, RM Willett, R Calderbank, Performance bounds for expander-based compressed sensing in poisson noise. IEEE Trans. Sig. Process. **59**(9), 4139–4153 (2011). https://doi.org/10.1109/TSP.2011.2157913
57. M Raginsky, RM Willett, ZT Harmany, RF Marcia, Compressed sensing performance bounds under poisson noise. IEEE Trans. Sig. Process. **58**(8), 3990–4002 (2010). https://doi.org/10.1109/TSP.2010.2049997
58. ZT Harmany, RF Marcia, RM Willett, This is spiral-tap: sparse poisson intensity reconstruction algorithms? Theory and practice. IEEE Trans. Image Process. **21**(3), 1084–1096 (2012). https://doi.org/10.1109/TIP.2011.2168410
59. R Gribonval, G Chardon, L Daudet, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Blind calibration for compressed*

sensing by convex optimization, (Kyoto, 2012). https://hal.inria.fr/hal-00658579. Accessed 25-30 Mar 2012
60. Z Zhang, T Jung, S Makeig, Z Pi, BD Rao, Spatiotemporal sparse Bayesian learning with applications to compressed sensing of multichannel physiological signals. IEEE Trans. Neural Syst. Rehabil. Eng. **22**, 1186–1197 (2014)
61. S Liu, YD Zhang, T Shan, S Qin, MG Amin, Structure-aware Bayesian compressive sensing for frequency-hopping spectrum estimation. Proc. SPIE. **9857**, 9857–9857 (2016)
62. L He, L Carin, Exploiting structure in wavelet-based Bayesian compressive sensing. IEEE Trans. Sig. Process. **57**, 3488–3497 (2009). https://doi.org/10.1109/TSP.2009.2022003
63. RE Carrillo, AB Ramirez, GR Arce, KE Barner, BM Sadler, Robust compressive sensing of sparse signals: a review. EURASIP J. Adv. Sig. Process. **2016**(1), 108 (2016). https://doi.org/10.1186/s13634-016-0404-5