PAPER

Hierachical Bayesian models and sparsity: ℓ_2-magic

IOP Publishing Inverse Problems

Inverse Problems **35** (2019) 035003 (26pp) https://doi.org/10.1088/1361-6420/aaf5ab

Hierachical Bayesian models and sparsity: ℓ_2-magic

D Calvetti, E Somersalo and A Strang

Department of Mathematics, Applied Mathematics and Statistics,
Case Western Reserve University, Cleveland, OH, United States of America

E-mail: dxc57@case.edu, ejs49@case.edu and ags61@case.edu

Received 27 August 2018, revised 7 November 2018
Accepted for publication 3 December 2018
Published 18 January 2019

Abstract

Sparse recovery seeks to estimate the support and the non-zero entries of a sparse signal $x \in \mathbb{R}^n$ from possibly incomplete noisy observations $y = \mathsf{A}x_0 + \epsilon$, with $\mathsf{A} \in \mathbb{R}^{m\times n}$, $m \leqslant n$. It has been shown that under various restrictive conditions on the matrix A, the problem can be reduced to the ℓ_1 regularized problem

$$\min \|x\|_1 \text{ subject to } \|\mathsf{A}x - y\|_2 < \delta,$$

where δ is the size of the error ϵ, and the approximation error is well controlled by δ. A popular method for solving the above minimization problem is the iteratively reweighted least squares algorithm. Here we reformulate the question of sparse recovery as an inverse problem in the Bayesian framework, express the sparsity belief by means of a hierachical prior model and show that the maximum *a posteriori* (MAP) solution computed by a recently proposed iterative alternating sequential (IAS) algorithm, requiring only the solution of linear systems in the least squares sense, converges linearly to the unique minimum for any matrix A, and quadratically on the complement of the support of the minimizer. The values of the parameters of the hierarchical model are assigned from an estimate of the signal to noise ratio and *a priori* belief of the degree of sparsity of the underlying signal, and automatically take into account the sensitivity of the data to the different components of x. The approach gives a solid Bayesian interpretation for the commonly used sensitivity weighting in geophysics and biomedical applications. Moreover, since for a suitable choice of sequences of parameters of the hyperprior, the IAS solution converges to the ℓ_1 regularized solution, the Bayesian framework for inverse problems makes the ℓ_1-magic happen in the ℓ_2 framework.

Keywords: convergence rate, sensitivity weighting, Bayesian hypermodel, compressive sensing

(Some figures may appear in colour only in the online journal)

1. Introduction

Consider a linear discrete inverse problem of the form

$$\mathsf{A}x = b, \tag{1}$$

where $\mathsf{A} \in \mathbb{R}^{m\times n}$, $m \leqslant n$ is an ill-conditioned matrix and the right hand side vector is corrupt by noise. The classical approach for addressing the ill-posed problem is to replace it by a nearby well-posed one by considering the problem of finding a minimizer of a penalized functional, such as

$$F_p(x,\alpha) = \|b - \mathsf{A}x\|^2 + \alpha\|x\|_p^p,$$

where the second term penalizes solutions of large ℓ_p-norm and the regularization parameter $\alpha > 0$ determines the severity of the penalty, while the choice of the parameter p affects the properties of the solution. In particular, it is well known that $p \leqslant 1$ promotes sparsity of the solution, an observation that lies in the heart of compressive sensing. Several algorithms for effectively solving the above minimization problem when $p < 2$ have been proposed and analyzed in the literature, as well as, when $p = 2$, various versions of iteratively reweighted least squares (IRLS) algorithms that are relevant for the present work [9, 13].

An alternative for the regularization approach is to reformulate the inverse problem in the Bayesian framework by extending (1) to random variables characterized by probability distributions, and introducing a prior that encodes the available information about the unknown. The observation equation, defining the likelihood, is then used to update the prior based on the data, giving rise to the posterior distribution that represents the Bayesian solution to the inverse problem [2, 14].

In this work, we consider a particular hierarchical Bayesian model, previously introduced and analyzed in [2–6], and in particular, an iterative algorithm for computing the maximum *a posteriori* (MAP) estimate of the extended hierarchical posterior model. The hierarchical model postulates a conditionally Gaussian prior model with variable prior variances, and a hyperprior model from the family of Gamma distributions for the prior variances. The iterative algorithm, referred to as iterative alternating scheme (IAS), solves the MAP estimate by alternatingly updating the estimate of the unknown and its prior variance, and it can be interpreted as a Bayesian IRLS algorithm with the *a priori* belief about the sparsity of the solution. Indeed, in the previous works, it has been shown that with particular hyperparameter selections, the IAS algorithm is particularly suitable for estimating sparse signals, akin to the ℓ_1 regularization [2–4], and in the context of solving the magnetoencephalography (MEG) inverse problem in brain imaging, it was shown to be globally convergent with a unique global minimum [5]. In this work, we extend the analysis in different ways.

As pointed out, the iterative algorithm based on the hierarchical model has been demonstrated in practice to be an efficient alternative for the ℓ_1-penalized optimization for recovering sparse signals. In this work, we formally show that one of the hyperparameters, the shape parameter of the underlying Gamma distribution, controls the sparsity in the sense that at the limit, the solution of the IAS algorithm converges to the minimizer of the ℓ_1-penalized regularized solution. Thus, we can argue that while the convergence of the IAS is independent of

particular properties of the matrix A, the conditions guaranteeing sparse recovery with the ℓ_1 penalty can be applied to guarantee that the solution is close to a sparse solution.

In addition, we demonstrate that through the limiting process of the hypermodel, the regularization parameter α in the ℓ_1-penalized functional can be automatically selected based on the information about the signal-to-noise ratio (SNR) and *a priori* belief of the size of the support of the underlying signal.

Furthermore, it has been experimentally observed that when applied to data arising from underlying signal with small support, the IAS algorithm converges rapidly. In this work, the convergence rate is analyzed in detail. We show that regardless of the properties of the underlying signal, the convergence is at least linear, essentially quadratic for sparse or compressible signals.

Another important extension of the previous work is related to the issue of sensitivity weighting and hyperparameter selection. When the forward model represents a field measurement (acoustic, electromagnetic, gravitation potential) and x represents the discretized primary or secondary source, the data are much more sensitive to source components near the receivers than those far away from it. This is reflected by the fact that the columns of the matrix A scale typically as $1/r^{\alpha}$, where r is the characteristic distance between the source component and the receiver. The varying sensitivity of the data to the source components biases the minimizer towards solutions with all the active components close to the receivers: because of the ill-posedness of the problem, superficial sources explain the noisy data as well as deep sources, but since for the former the value of the penalty term is much lower, they are strongly favored. The remedy proposed in the literature to address this well-known problem, see, e.g. [16–20], is to replace the ℓ_p regularized problem with a *sensitivity-weighted* minimization problem

$$F_p(x, \alpha, w) = \|b - \mathsf{A}x\|^2 + \alpha \sum_j w_j |x_j|^p, \tag{2}$$

with the weights w_j chosen so as to compensate for the differences in sensitivity, defined precisely later in this article.

The observation model dependent sensitivity scaling has been difficult to justify in the Bayesian framework: in a Gibbs type prior

$$\pi_{\text{prior}}(x) \propto \exp\left(-\alpha \sum_j w_j |x_j|^p\right),$$

the selection of w_j would be tantamount to favoring sources far away from detectors based on the forward model rather than on *a priori* belief about the solution, a position that is hardly defendable. In this paper we will show that by combining hierarchical prior models and an exchangeability argument asserting that the signal to noise ratio should depend only on the cardinality of the support of the signal, independently of the actual locations, it is possible to provide a Bayesian interpretation of sensitivity scaling without violating the Bayesian principles of prior design.

The paper is organized as follows. In the section 2 we present a brief review of hierachical Bayesian models for linear observation models with Gaussian noise and the IAS algorithm for computing the corresponding MAP estimate. Section 3 addresses the sparsity promoting role of the shape hyperparameter and proposes an automatic way to assign the value of the scale parameters. After introducing the concept of exchangeability, we show a way to assign the value of the scale parameters that expresses the belief that the signal to noise ratio is a function of the cardinality of the support but not of its location. Furthermore, we show that as the shape parameter goes to zero, the MAP estimates computed with the IAS algorithm converges

to the minimizer of the ℓ_1 penalized functional. In section 4 we prove that the convergence rate of the IAS algorithm is at least linear, and at least quadratic on the complement of the support of the minimizer. Numerical examples in one and two dimensions are presented in section 5.

2. Hierarchical Bayesian model

We start by a brief review of the hierarchical Bayesian model discussed in previous articles [3–5] and further analyzed in this paper. Consider the linear observation model with additive Gaussian noise,

$$b = \mathsf{A}x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma), \tag{3}$$

where $\mathsf{A} \in \mathbb{R}^{m\times n}$ and $\Sigma \in \mathbb{R}^{m\times m}$ is a symmetric positive definite noise covariance matrix. Here we consider the case $m \leqslant n$, where the problem is underdetermined. Introducing the Cholesky decomposition of the noise precision matrix $\Sigma^{-1} = \mathsf{S}^{\mathsf{T}}\mathsf{S}$, the likelihood density can be expressed as

$$\pi_{b|x}(b \mid x) \propto \exp\left(-\frac{1}{2}(b - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(b - \mathsf{A}x)\right) = \exp\left(-\frac{1}{2}\|\mathsf{S}(b - \mathsf{A}x)\|^2\right),$$

where '$\propto$' stands for proportional up to a constant scaling factor.

We define a conditionally Gaussian prior model for x, postulating that for a prior variance vector $\theta \in \mathbb{R}^n_+ = \{\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n \mid \theta_j > 0\}$,

$$x \mid \theta \sim \mathcal{N}(0, \mathsf{D}_\theta), \quad \mathsf{D}_\theta = \mathrm{diag}(\theta),$$

yielding the prior density

$$\pi_{x|\theta}(x \mid \theta) = \frac{1}{(2\pi)^{n/2}\sqrt{\theta_1 \cdots \theta_n}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j}\right).$$

We remark that because θ is itself a random variable, the normalizing factor cannot be ignored, hence our aim is to estimate both x and θ based on the observation b. To this end we introduce a hypermodel for the variances θ_j, postulating that they are mutually independent and distributed according to a Gamma distribution,

$$\theta_j \sim \mathrm{Gamma}(\theta_j^*, \beta), \quad \pi_{\theta_j}(\theta_j) = \frac{1}{\Gamma(\beta)\theta_j^*}\left(\frac{\theta_j}{\theta_j^*}\right)^{\beta-1} \exp\left(-\frac{\theta_j}{\theta_j^*}\right).$$

Observe that to simplify slightly the model, we set the shape parameter $\beta > 0$ equal for all components, while assigning individually the values of the scaling parameters θ_j^*. We refer to [5] for the statistical motivation for choosing this type of hypermodel that, unlike a common practice in the statistical literature, is not conjugate to the prior model.

We may now combine the likelihood, prior, and hypermodel by Bayes' formula to obtain the posterior density for the pair (x, θ),

$$\begin{aligned}\pi_{(x,\theta)|b}(x, \theta \mid b) &= \frac{\pi_{x|\theta}(x \mid \theta)\pi_\theta(\theta)\pi(b \mid x, \theta)}{\pi_b(b)} \\ &\propto \exp\left(-\frac{1}{2}\|\mathsf{S}(b - \mathsf{A}x)\|^2 - \frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} - \sum_{j=1}^{n}\left[\frac{\theta_j}{\theta_j^*} - \left(\beta - \frac{3}{2}\right)\log\frac{\theta_j}{\theta_j^*}\right]\right).\end{aligned}$$

In the following we assume that $\beta > 3/2$ and denote $\eta = \beta - 3/2 > 0$. Furthermore, by scaling the forward mapping A and the data b as $(\mathsf{A}, b) \to (\mathsf{SA}, \mathsf{S}b)$, without loss of generality we can assume that $\mathsf{S} = \mathsf{I}_m$, the identity matrix of size $m \times m$, hence we write the Gibbs energy functional as

$$E(x,\theta) = \overbrace{\frac{1}{2}\|b - \mathsf{A}x\|^2}^{(a)} + \underbrace{\frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} + \sum_{j=1}^{n}\left[\frac{\theta_j}{\theta_j^*} - \eta\log\frac{\theta_j}{\theta_j^*}\right]}_{(b)}, \tag{4}$$

the braces identifying the x-dependent (a) and θ-dependent (b) portions of the functional. For the time being, we have suppressed the dependency on the hyperparameters.

2.1. The IAS algorithm

The maximum *a posteriori* (MAP) estimate of the pair (x, θ) is, by definition, a minimizer of the energy functional (4). In the articles [2–4], an iterative alternating sequential (IAS) algorithm was proposed, consisting of two separate minimization steps:

1. **Initialize:** Set $\theta = \theta^0$, $k = 0$.
2. **Iterate until convergence:**

   (i) Update x setting $x^{k+1} = \text{argmin}\{E(x, \theta^k)\}$.
   (ii) Update θ setting $\theta^{k+1} = \text{argmin}\{E(x^{k+1}, \theta)\}$.
   (iii) Increase $k \to k + 1$.

As pointed out in the cited articles, the algorithm is simple to implement because of the particular structure of the energy functional. Indeed, since in step (i) only the x-dependent part (a) in (4) needs to be considered, the corresponding minimization problems can be reduced to solving the linear system

$$\begin{bmatrix} \mathsf{A} \\ \mathsf{D}_{\theta^k}^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

in the least squares sense. In step (ii), on the other hand, where only the θ-dependent part (b) in (4) is minimized and the components are independent, the minimum can be explicitly computed as a critical point of the component functional, yielding

$$\theta_j^{k+1} = \theta_j^* \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{(x_j^{k+1})^2}{2\theta_j^*}} \right).$$

The similarity of the above iterative algorithm and the iteratively reweighted least squares algorithms (IRLS) is obvious, however, observe that the goal here is not to find a minimizer with the ℓ_p penalty. In IRLS, the idea of penalizing the components of x by weighing them individually, recomputing the weights iteratively can be traced back to the doctoral work of Lawson in 1961 [15] for the solution of uniform approximation problems. Extensions of this work eventually led to the FOCUSS algorithm for the reconstruction of sparse signals [13]. In general, since the new weights are expressed in terms of the corresponding components of the previous approximate solution, care must be taken to guarantee that the algorithm is well

defined and converges to the underlying sparse solution, often by requiring that the matrix A satisfies some limiting conditions. For instance, the convergence analysis of the IRLS algorithm proposed in [9] is restricted for classes of matrices satisfying conditions analogous to those for the ℓ_1 sparse recovery in compressed sensing [7, 8, 11, 12].

The extension of the minimization problem from $\mathbb{R}^n$ to $\mathbb{R}^n \times \mathbb{R}^n$ allows us to prove the following general convergence result [5].

Theorem 2.1. *For $\eta > 0$ and $\theta^* \in \mathbb{R}^n_+$, the energy functional (4) defined over $\mathbb{R}^n \times \mathbb{R}^n_+$ is strictly convex, thus having a unique global minimizer $\widehat{z} = (\widehat{x}, \widehat{\theta})$. The IAS algorithm produces a sequence $z^k = (x^k, \theta^k)$ that converges to the global minimum. Furthermore, the point $\widehat{x}$ is the global minimizer of the functional*

$$F(x) = E(x, f(x)), \quad f(x) = (f_1(x_1), \ldots, f_n(x_n)), \tag{5}$$

where

$$f_j(t) = \theta_j^* \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{t^2}{2\theta_j^*}} \right).$$

One of the main questions that we address in this article, not discussed in the cited work, is the convergence rate of the IAS algorithm. However, we start by analyzing further the role of the hyperparameters.

3. Hyperparameters

It has been pointed out in the literature how the hyperparameters β and θ^* affect on the MAP solution: the former controls the sparsity of the solution, while the second one can be related to sensitivity scaling, if properly interpreted. One of the aims of this work is to further analyze the role of the hyperparameters.

We begin with a simple limiting argument that will be developed further later on. Consider the function (5). It was shown in [5] that for fixed $x \in \mathbb{R}^n$,

$$F_0(x) = \lim_{\eta \to 0+} E(x, f(x)) = \frac{1}{2}\|b - \mathsf{A}x\|^2 + \sqrt{2} \sum_{j=1}^{n} \frac{|x_j|}{\sqrt{\theta_j^*}}, \tag{6}$$

where the sum extends only over the support of x,

$$S = \mathrm{supp}(x) = \{ j \mid x_j \neq 0 \}.$$

This preliminary observation serves two purposes. First, it suggests that the parameter $\eta > 0$ controls the sparsity of the solution, and second, it reveals that the parameter θ_j^* represents a weight that can be related to the sensitivity of the data. Below, we start by discussing the second observation.

3.1. Scale parameters and sensitivity weighting

As discussed in the Introduction, a common procedure in applied inverse problems is to use weighted penalty functions to compensate for the non-uniform sensitivity of the data to components of the unknown. In this section, we show how the current model provides a Bayesian

interpretation of the sensitivity scaling without the need to resort to questionable data-dependent priors.

The sensitivity of the linear forward model $x \mapsto \mathsf{A}x$ to the jth component of the vector x can be defined as

$$s_j = \left\| \frac{\partial(\mathsf{A}x)}{\partial x_j} \right\| = \|\mathsf{A}e_j\|,$$

where e_j is the jth canonical basis vector. Hence, s_j equals the norm of the jth column of A. In the inverse problems literature, it is common to choose the weights w_j in (2) to compensate for the variable sensitivity by setting $w_j \propto \|\mathsf{A}e_j\|^p$. Equivalently, this is tantamount to scaling the columns of A to have unit Euclidian norm by multiplying it from the right by a diagonal matrix $\mathsf{D} = \mathrm{diag}(1/s_1, \ldots, 1/s_n)$, and concurrently rescaling the variable x with the corresponding inverse scaling,

$$\mathsf{A}x = (\mathsf{AD})(\mathsf{D}^{-1}x) = \widetilde{\mathsf{A}}w, \qquad \widetilde{\mathsf{A}} = \mathsf{AD},\ w = \mathsf{D}^{-1}x,$$

amounting to a change of variables and concurrent rescaling the columns of the forward operator $\widetilde{\mathsf{A}}$. Here, we show that a similar scaling can be obtained through the hierarchical Bayesian model.

The following argument is a modification and extension of the discussion included in [6] in the context of interpreting MEG data: the key equation to help elucidate the role of θ_j^* as a weight factor is (6), although in the ensuing discussion, η may have any non-negative value.

Consider the observation model (3). We define the signal-to-noise ratio (SNR) by the formula

$$\mathrm{SNR} = \frac{\mathsf{E}\{\|b\|^2\}}{\mathsf{E}\{\|\epsilon\|^2\}},$$

where both x and ϵ are interpreted as random variables and the expectation of x is with respect to the prior distribution. We start with the following simple calculation in which we assume that the noise has not been whitened.

Lemma 3.1. *Assume a priori, that we have* $\mathrm{supp}(x) = S \subset \{1, 2, \ldots, n\}$. *Given the hierarchical model*

$$x \sim \mathcal{N}(0, \mathsf{D}_\theta), \quad \theta_j \sim \mathrm{Gamma}(\beta, \theta_j^*) \text{ for } j \in S,$$

and $\theta_j = 0$ *for* $j \notin S$, *the signal-to-noise ratio conditional on the support assumption is*

$$\mathrm{SNR}_S = \frac{\sum_{j\in S} \beta\theta_j^* \|\mathsf{A}e_j\|^2}{\mathrm{trace}(\Sigma)} + 1,$$

where $e_j \in \mathbb{R}^n$ *is the jth canonical coordinate vector.*

Proof. We start by observing that

$$\mathsf{E}\{\|\epsilon\|^2\} = \mathrm{trace}\big(\mathsf{E}\{\epsilon\epsilon^{\mathsf{T}}\}\big) = \mathrm{trace}(\Sigma)$$

and, from the independence of x and ϵ,

$$\begin{aligned}\mathsf{E}\{\|b\|^2\} &= \mathsf{E}\{\|\mathsf{A}x\|^2\} + \mathsf{E}\{\|\epsilon\|^2\} = \mathrm{trace}\big(\mathsf{E}\{(\mathsf{A}x)(\mathsf{A}x)^{\mathsf{T}}\}\big) + \Sigma \\ &= \mathrm{trace}\big(\mathsf{A}\mathsf{E}\{xx^{\mathsf{T}}\}\mathsf{A}^{\mathsf{T}}\big) + \Sigma.\end{aligned}$$

The hierarchical prior model implies that

$$\mathsf{E}\{xx^{\mathsf{T}} \mid \theta\} = \sum_{j\in S} \theta_j e_j e_j^{\mathsf{T}},$$

hence

$$\mathrm{trace}(\mathsf{A}\mathsf{E}\{xx^{\mathsf{T}} \mid \theta\}\mathsf{A}^{\mathsf{T}}) = \sum_{j\in S} \theta_j \|\mathsf{A}e_j\|^2.$$

The result follows from the observation that, since $\theta_j \sim \mathrm{Gamma}(\beta, \theta_j^*)$, its expectation is $\mathsf{E}\{\theta_j\} = \beta\theta_j^*$. □

Let $\|x\|_0 = \mathrm{card}(\mathrm{supp}(x))$ denote for the cardinality of the support of the vector x. The following definition is useful when trying to recover signals believed to have sparse support.

Definition 3.2. A problem satisfies the *exchangeability condition* if whenever $\mathrm{card}(S) = \mathrm{card}(S')$

$$\mathrm{SNR}_S = \mathrm{SNR}_{S'}.$$

We are now ready to prove the following theorem, that provides a criterion for setting the values of the scale parameters of the Gamma hyperprior.

Theorem 3.3. *Given the probability distribution of the cardinality of the source*

$$\mathsf{P}\{\|x\|_0 = k\} = p_k, \quad p_0 = p_n = 0,$$

and an estimate $\overline{\mathrm{SNR}}$ *of the signal-to-noise ratio, if the system satisfies the exchangeability condition, the values of the scale hyperparameters* θ_j^* *should be set to*

$$\theta_j^* = \frac{C}{\|\mathsf{A}e_j\|^2}, \quad C = \frac{(\overline{\mathrm{SNR}} - 1)\,\mathrm{trace}(\Sigma)}{\beta} \sum_{k=1}^{n-1} \frac{p_k}{k}. \tag{7}$$

Proof. Assume for the time being that the cardinality of the support of the source is k. Then from the previous lemma

$$\sum_{\ell=1}^{k} \beta\theta_{j_\ell}^* \|\mathsf{A}e_{j_\ell}\|^2 = \mathrm{trace}(\Sigma)(\mathrm{SNR} - 1)$$

for some index values $j_1, \ldots, j_k$, and because of the exchangeability condition this equation must be satisfied for any choice of k indices.

Let P_k be the $\binom{n}{k} \times n$ matrix where each row contains zeros in correspondence of the complement of the support, and ones in correspondence of the support. In other words, the rows of P_k represent all possible choices for the support of x. Then for the vector $\gamma \in \mathbb{R}^n$ such that $\gamma_j = \beta\theta_j^* \|\mathsf{A}e_j\|^2$, we have that

$$\mathsf{P}_k\gamma = \mathrm{trace}(\Sigma)(\mathrm{SNR} - 1)\mathbf{1},$$

where $\mathbf{1}$ is a $\binom{n}{k}$-vector of ones. Since P_k has rank n, we must have

$$\gamma_j = \beta\theta_j^* \|\mathsf{A}e_j\|^2 = \frac{1}{k}\mathrm{trace}(\Sigma)(\mathrm{SNR}-1), \quad 1 \leqslant j \leqslant n,$$

leading to

$$\theta_j^* \mid (\|x\|_0 = k) = \frac{1}{k}\mathrm{trace}(\Sigma)(\mathrm{SNR}-1)\frac{1}{\beta\|\mathsf{A}e_j\|^2}.$$

Finally, marginalizing over the cardinality of the support it follows that

$$\theta_j^* = \sum_{k=1}^{n-1} p_k\big(\theta_j^* \mid (\|x\|_0 = k)\big) = \frac{C}{\|\mathsf{A}e_j\|^2},$$

completing the proof. □

Revisiting (6) in the light of the last theorem, the limiting functional can be written as

$$F_0(x) = \frac{1}{2}\|b - \mathsf{A}x\|^2 + \sqrt{\frac{2}{C}}\sum_{j=1}^{n} w_j|x_j|, \quad w_j = \|\mathsf{A}e_j\|,$$

which is what would be obtained from sensitivity weighting with the column norms of the forward map. Based on this observation, we have the following result on how Tikhonov regularization parameter can be selected on the basis of the estimated noise level and prior information about the support.

Corollary 3.4. *Given an estimate* $\overline{\mathrm{SNR}}$ *of the signal to noise ratio and a priori probability distribution of* $\|x\|_0$*, a judicious choice of the Tikhonov regularization parameter* α *for*

$$x_\alpha = \arg\min\Big\{\|b - \mathsf{A}x\| + \alpha\sum_{j=1}^{n} w_j|x_j|\Big\}$$

is

$$\alpha = \sqrt{\frac{2}{C}} = \sqrt{\frac{2\beta}{(\overline{\mathrm{SNR}}-1)\,\mathrm{trace}(\Sigma)}}\sum_{k=1}^{n-1}\frac{p_k}{k},$$

which is always real because SNR is always greater than 1.

Remark 3.5. Often, one may have additional restrictive prior information about the size of the components of x, such as '$|x_j| < M$ with high probability', based on, e.g. physical considerations. A natural way to incorporate such information is to define the effective value

$$\theta_{\mathrm{eff},j}^* = \min\{\theta_j^*, (M/2)^2\}, \tag{8}$$

where θ_j^* is given by (7), the value M thus representing two times the standard deviation.

To simplify notations, we shall non-dimensionalize the problem by defining

$$\theta_j \to \frac{\theta_j}{\theta_j^*}, \quad x_j \to \frac{x_j}{\sqrt{\theta_j^*}}, \quad \mathsf{A} \to \mathsf{A}\mathrm{diag}(\sqrt{\theta^*}),$$

which is tantamount to rescaling each column of the matrix A by the square root of the corresponding element of θ^*. Therefore in the following discussion, without loss of generality, we assume that $\theta_j^* = 1$.

3.2. Shape parameter and sparsity

A significant amount of research continues to be devoted to the conditions for the exact—or near exact—recovery of sparse signals. The results on the optimality of the ℓ_1 penalized approximation have motivated the use of different norms to measure fidelity and favor sparsity, raising interest in how to approach the problem computationally. Hierarchical Bayesian models with suitable choice of hyperpriors have been shown to promote sparsity: in [5] it was shown that with the Gamma hyperprior, for x fixed, if the shape parameter η is small enough, the Gibbs energy converges towards the ℓ_1-penalized functional. Here we show that, in the limit as η goes to zero, the IAS solution $\widehat{x} = x_\eta$ corresponding to the parameter η converges to ℓ_1-penalized solution, thus the ℓ_1-magic can be attained as the limit of an all ℓ_2 procedure.

To simplify the notation, without loss of generality we assume that $\theta_j^* = 1$ and emphasize the dependence of the functional minimized by the IAS algorithm on η by writing

$$F_\eta(x) = \frac{1}{2}\|b - \mathsf{A}x\|^2 + \frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} + \sum_{j=1}^{n}(\theta_j - \eta\log\theta_j),$$

where

$$\theta_j = f_j(x_j,\eta) = f(x_j,\eta)$$

with

$$f(x,\eta) = \frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{x^2}{2}},$$

and let

$$x_\eta = \operatorname{argmin}\{F_\eta(x)\}$$

be the minimizer of this functional. The following lemma shows that as η goes to zero, the sequence of minimizers computed by the IAS algorithm remains bounded.

Lemma 3.6. *There is a constant $B > 0$ such that*

$$\|x_\eta\| \leqslant B,$$

for all η, $0 \leqslant \eta \leqslant \frac{1}{2}$.

Proof. The claim is proved by contradiction. Assume that we can find a sequence $\eta^1, \eta^2, \ldots$ such that $\|x^k\| > k$, where $x^k = x_{\eta^k}$. Then $\|x^k\|_\infty > k/n$ and from the formula for updating the components of θ,

$$\theta_j^k = f(x_j^k,\eta^k) = \frac{\eta^k}{2} + \sqrt{\frac{(\eta^k)^2}{4} + \frac{(x_j^k)^2}{2}} > \frac{|x_j^k|}{\sqrt{2}} \tag{9}$$

it follows that $\|\theta^k\|_\infty > k/(\sqrt{2}n)$. This implies that

$$\begin{aligned} F_{\eta^k}(x^k) &\geqslant \sum_{j=1}^{n}(\theta_j^k - \eta^k \log \theta_j^k) \\ &\geqslant \sum_{j=1}^{n}(\theta_j^k - \eta^k \theta_j^k) \geqslant \frac{1}{2}\sum_{j=1}^{n}\theta_j^k \\ &\geqslant \frac{1}{2}\|\theta^k\|_\infty \geqslant \frac{k}{2\sqrt{2}n} \to \infty. \end{aligned}$$

This is a contradiction since x^k is the minimizer of F_{η^k} and therefore, in particular,

$$F_{\eta^k}(x^k) \leqslant F_{\eta^k}(0) = \frac{1}{2}\|b\|^2 + n(\eta^k - \eta^k \log \eta^k) < \infty,$$

thus completing the proof. □

An immediate consequence of the previous lemma is that, for $0 \leqslant \eta \leqslant 1/2$, the components of θ are also bounded, since

$$\theta_{\eta,j} = f(x_{\eta,j}, \eta) = \frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{x_{\eta,j}^2}{2}} \leqslant \frac{1}{4} + \sqrt{\frac{1}{16} + \frac{B^2}{2}} = K.$$

Let us denote

$$x_0 = \operatorname{argmin}\{F_0(x)\}. \tag{10}$$

We are now ready to prove the following result.

Theorem 3.7. *Assume that the matrix* A *is such that the minimizer (10) is unique. Then, as* $\eta \to 0+$*, the minimizers* x_η *converge to the minimizer* x_0 *of the* ℓ_1*-penalized functional* F_0*.*

Proof. The proof is by contradiction. Assume that there is a sequence of η^i converging to 0 such that for some $\delta > 0$,

$$\|x_{\eta^i} - x_0\| > \delta > 0.$$

It follows from the boundedness of the x_η established in the lemma 3.6 and the compactness of the ball $\{\|x\| \leqslant B\}$ that there is a convergent subsequence $\eta^{i_k} \to 0$ such that

$$x^k = x_{\eta^{i_k}} \to \bar{x}_0, \qquad \|x_0 - \bar{x}_0\| > \delta.$$

We denote $\theta_j^k = f(x_j^k, \eta^{i_k})$ and write

$$|F_{\eta^{i_k}}(x^k) - F_0(\bar{x}_0)| \leqslant |F_{\eta^{i_k}}(x^k) - F_0(x^k)| + |F_0(x^k) - F_0(\bar{x}_0)|.$$

By continuity of F_0, the second term on the right tends to zero as k increases and the first term on the right can be estimated as

$$
\begin{aligned}
|F_{\eta^{i_k}}(x^k) - F_0(x^k)| &\leqslant \frac{1}{2}\sum_{j=1}^n \left| \frac{(x_j^k)^2}{\theta_j^k} - \sqrt{2}|x_j^k| \right| + \sum_{j=1}^n \left| \theta_j^k - \frac{|x_j^k|}{\sqrt{2}} \right| + \sum_{j=1}^n \eta^{i_k} |\log \theta_j^k| \\
&= \frac{1}{2}\sum_{j=1}^n \frac{|x_j^k|}{\theta_j^k} \left| |x_j^k| - \sqrt{2}\theta_j^k \right| + \sum_{j=1}^n \left| \theta_j^k - \frac{|x_j^k|}{\sqrt{2}} \right| + \sum_{j=1}^n \eta^{i_k} |\log \theta_j^k| \\
&\leqslant \sqrt{2}\sum_{j=1}^n \left| |x_j^k| - \sqrt{2}\theta_j^k \right| + \sum_{j=1}^n \eta^{i_k} |\log \theta_j^k|,
\end{aligned}
$$

where we used the inequality (9) for x_j^k and θ_j^k. Since $|x_j^k| \to |i\,\bar{x}_{0,j}|$ and $\theta_j^k \to |\bar{x}_{0,j}|/\sqrt{2}$, the first term converges to zero. Furthermore, from

$$
\eta^{i_k} \leqslant \theta_i^k \leqslant K,
$$

it follows that

$$
\eta^{i_k} \sum_{j=1}^n |\log \theta_j^k| \to 0, \quad \text{as } k \to \infty,
$$

proving that

$$
\lim_{k\to\infty} F_{\eta^k}(x^k) = F_0(\bar{x}_0).
$$

Since x^k is the minimizer of F_{η^k}, we have

$$
F_{\eta^k}(x^k) \leqslant F_{\eta^k}(x_0),
$$

and, at the limit,

$$
F_0(\bar{x}_0) \leqslant F_0(x_0), \quad \bar{x}_0 \neq x_0.
$$

which contradicts the uniqueness of the minimizer x_0, completing the proof. □

The convergence implies, in particular, that if A is a matrix such that the ℓ_1 regularized solution of the minimization problem (10) is sparse, then the solution of the IAS algorithm with $\eta > 0$ small can be made arbitrarily small outside the support of x_0. Likewise, if the minimization problem above is compressible, that is, the components of x_0 are smaller than a known threshold outside a set $S \subset \{1, 2, \ldots, n\}$, when $\eta > 0$ is small enough, the same is true for the IAS solution x_η with a slightly larger threshold.

In the following section, we establish some results about the rate of convergence of the IAS algorithm and show that if the underlying signal is sparse, the convergence is quadratic on the complement of the support.

4. Convergence rate

For the sake of simplifying the notation, in this section we combine x and θ in the new variable $z = (x, \theta) \in \mathbb{R}^{2n}$, and, for given $\eta > 0$ fixed, denote the objective function (4) to be minimized by $E(z) = E(x, \theta)$. We partition the Hessian of E into four $n \times n$ blocks,

$$\mathsf{H}(z) = \begin{bmatrix} \mathsf{H}^{11}(z) & \mathsf{H}^{12}(z) \\ \mathsf{H}^{21}(z) & \mathsf{H}^{22}(z) \end{bmatrix}, \tag{11}$$

where

$$\mathsf{H}^{11}(z) = \left[\frac{\partial^2 E}{\partial x_j \partial x_k}\right] = \mathsf{A}^{\mathsf{T}}\mathsf{A} + \mathrm{diag}(1/\theta), \tag{12}$$

$$\mathsf{H}^{12}(z) = (\mathsf{H}^{21}(z))^{\mathsf{T}} = \left[\frac{\partial^2 E}{\partial x_j \partial \theta_k}\right] = -\mathrm{diag}(x/\theta^2), \tag{13}$$

and

$$\mathsf{H}^{22}(z) = \left[\frac{\partial^2 E}{\partial \theta_j \partial \theta_k}\right] = \mathrm{diag}(x^2/\theta^3 + \eta/\theta^2), \tag{14}$$

with the powers of vectors, the division by θ, or by its powers, understood in the componentwise sense. Introducing the matrices $\mathsf{Q}_1,\ \mathsf{Q}_2 \in \mathbb{R}^{2n\times n}$,

$$\mathsf{Q}_1 = \begin{bmatrix} \mathsf{I}_n \\ \mathsf{O}_n \end{bmatrix}, \quad \mathsf{Q}_2 = \begin{bmatrix} \mathsf{O}_n \\ \mathsf{I}_n \end{bmatrix},$$

where I_n and O_n are, respectively, the unit matrix and the zero matrix of size $n \times n$, the updating steps in each iteration of the IAS algorithm can be expressed in the following unified form.

Given the current $z^c \in \mathbb{R}^{2n}$,

$$\text{minimize } E(z^c + \mathsf{Q}y), y \in \mathbb{R}^n, \tag{15}$$

where $\mathsf{Q} \in \{\mathsf{Q}_1, \mathsf{Q}_2\}$.

Before stating the main result about the rate of convergence of the IAS algorithm, we prove that, in a neighborhood of the minimizer $\widehat{z}$ of (4), the norm of the error at the next iteration can be bounded in terms of the norm of the error in the current iteration.

Lemma 4.1. *Let* Ω *be an open connected neighborhood of the minimizer* $\widehat{z}$ *of (4) where the Hessian* $\mathsf{H}(z)$ *of* E *is Lipschitz continuous with Lipschitz constant* γ*, the condition number of* H *is bounded above by* $\kappa > 0$ *and* $\|\mathsf{H}(z)^{-1}\| \leqslant \nu$*. Then the error* $\varepsilon^+ = z^+ - \widehat{z}$ *in approximating* $\widehat{z}$*, with* $z^+ = z^c + \mathsf{Q}y^+$ *where* y^+ *is the minimizer of (15), can be written as*

$$\varepsilon^+ = \mathsf{J}\varepsilon^c + e,$$

where $\varepsilon^c = z^c - \widehat{z}$ *and, letting* $\widehat{\mathsf{H}} = \mathsf{H}(\widehat{z})$,

$$\mathsf{J} = \mathsf{I}_{2n} - \mathsf{Q}\left(\mathsf{Q}^{\mathsf{T}}\widehat{\mathsf{H}}\mathsf{Q}\right)^{-1}\mathsf{Q}^{\mathsf{T}}\widehat{\mathsf{H}}, \tag{16}$$

and

$$\|e\| \leqslant 2\nu\gamma(1+\kappa)^2\|\varepsilon^c\|^2.$$

Proof. Given the current iterate z^c, consider the function

$$g(y) = E(z^c + \mathsf{Q}y), \quad y \in \mathbb{R}^n,$$

with gradient

$$\nabla g(y) = \mathsf{Q}^{\mathsf{T}} \nabla E(z^c + \mathsf{Q}y).$$

At the minimizer $y = y^+$ of g,

$$\nabla g(y^+) = \mathsf{Q}^{\mathsf{T}} \nabla E(z^c + \mathsf{Q}y^+) = \mathsf{Q}^{\mathsf{T}} \nabla E(z^+) = 0 \tag{17}$$

must hold. Denote the local quadratic model of E based at $\widehat{z}$ by

$$M(z) = E(\widehat{z}) + \frac{1}{2}(z - \widehat{z})^{\mathsf{T}} \widehat{\mathsf{H}}(z - \widehat{z})$$

and let

$$D(z) = E(z) - M(z)$$

be the approximation error. Then, under the mild regularity conditions on E, the following bound on the approximation error,

$$|D(z)| \leqslant \frac{\gamma}{6}\|z - \widehat{z}\|^3,$$

and on its gradient,

$$\|\nabla D(z)\| \leqslant \frac{\gamma}{2}\|z - \widehat{z}\|^2,$$

hold uniformly in Ω, see [10]. Substituting

$$\nabla E(z) = \widehat{\mathsf{H}}(z - \widehat{z}) + \nabla D(z)$$

in (17) yields

$$\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}}(z^+ - \widehat{z}) + \mathsf{Q}^{\mathsf{T}} \nabla D(z^+) = 0,$$

which upon the substitution $z^+ = z^c + \mathsf{Q}y^+$ becomes

$$\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}}(z^c - \widehat{z} + \mathsf{Q}y^+) + \mathsf{Q}^{\mathsf{T}} \nabla D(z^+) = 0.$$

Solving the last equation for y^+ we obtain

$$y^+ = -\left(\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}} \mathsf{Q}\right)^{-1} \left(\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}} \varepsilon^c + \mathsf{Q}^{\mathsf{T}} \nabla D(z^+)\right),$$

therefore, $\varepsilon^+ = z^+ - \widehat{z}$ satisfies

$$\begin{aligned}
\varepsilon^+ &= z^c - \widehat{z} - \mathsf{Q}\left\{\left(\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}} \mathsf{Q}\right)^{-1} \left(\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}} \varepsilon^c + \mathsf{Q}^{\mathsf{T}} \nabla D(z^+)\right)\right\} \\
&= \left\{\mathsf{I}_{2n} - \mathsf{Q}\left(\mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}} \mathsf{Q}\right)^{-1} \mathsf{Q}^{\mathsf{T}} \widehat{\mathsf{H}}\right\} \varepsilon^c + R(z^+) \\
&= \mathsf{J}\varepsilon^c + R(z^+).
\end{aligned}$$

To complete the proof, we need to estimate the remainder in term $R(z^+)$,

$$R(z^+) = -\mathsf{Q}\left(\mathsf{Q}^\mathsf{T}\widehat{\mathsf{H}}\mathsf{Q}\right)^{-1}\mathsf{Q}^\mathsf{T}\nabla D(z^+)$$

in terms of the error ε^c. From the observation that

$$\mathsf{Q}_j^\mathsf{T}\widehat{\mathsf{H}}\mathsf{Q}_j = \widehat{\mathsf{H}}^{jj}, \quad j = 1, 2$$

it follows that

$$\|(\mathsf{Q}^\mathsf{T}\widehat{\mathsf{H}}\mathsf{Q})^{-1}\| \leqslant \|\widehat{\mathsf{H}}^{-1}\| \leqslant \nu,$$

hence

$$\|R(z^+)\| \leqslant \|\widehat{\mathsf{H}}^{-1}\|\|\nabla D(z^+)\| \leqslant \frac{\nu\gamma}{2}\|\varepsilon^+\|^2.$$

Furthermore, from the estimate

$$\|\mathsf{J}\| \leqslant 1 + \|\mathsf{Q}(\mathsf{Q}^\mathsf{T}\widehat{\mathsf{H}}\mathsf{Q})^{-1}\mathsf{Q}^\mathsf{T}\widehat{\mathsf{H}}\| \leqslant 1 + \kappa,$$

it follows that

$$\|\varepsilon^+\| \leqslant \|\mathsf{J}\varepsilon^c\| + \|R(z^+)\| \leqslant (1+\kappa)\|\varepsilon^c\| + \frac{\nu\gamma}{2}\|\varepsilon^+\|^2,$$

and, if we are close enough to the minimizer that $\|\varepsilon^+\| < 1/\nu\gamma$,

$$\|\varepsilon^+\| \leqslant \frac{1+\kappa}{1-(\nu\gamma/2)\|\varepsilon^+\|}\|\varepsilon^c\| \leqslant 2(1+\kappa)\|\varepsilon^c\|.$$

Combining the above estimates we have

$$\|\varepsilon^+ - \mathsf{J}\varepsilon^c\| \leqslant 2\nu\gamma(1+\kappa)^2\|\varepsilon^c\|^2$$

which completes the proof. □

We remark that when $\mathsf{Q} = \mathsf{Q}_1$ the solution of (15) is the updated x, and in terms of the block partitioning (11) of the Hessian,

$$\mathsf{Q}_1\left(\mathsf{Q}_1^\mathsf{T}\widehat{\mathsf{H}}\mathsf{Q}_1\right)^{-1}\mathsf{Q}_1^\mathsf{T}\widehat{\mathsf{H}} = \begin{bmatrix} \mathsf{I}_n & (\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12} \\ 0 & 0 \end{bmatrix},$$

implying that the matrix J in (16) updating the error vector is

$$\mathsf{J}_1 = \begin{bmatrix} \mathsf{O}_n & -(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12} \\ \mathsf{O}_n & \mathsf{I}_n \end{bmatrix}. \tag{18}$$

Similarly, when $\mathsf{Q} = \mathsf{Q}_2$, the solution of (15) gives us the updated θ, and the corresponding error updating matrix is

$$\mathsf{J}_2 = \begin{bmatrix} \mathsf{I}_n & \mathsf{O}_n \\ -(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21} & \mathsf{O}_n \end{bmatrix}. \tag{19}$$

Each IAS iteration solves two minimization problems, one with respect to x and the other with respect to θ. Starting from the current error ε^c and denoting by ε' the error after the first minimization step, the error ε^+ at the end of the iteration is obtained in two steps as

$$\begin{aligned}\varepsilon' &= \mathsf{J}_1\varepsilon^c + e_1,\\ \varepsilon^+ &= \mathsf{J}_2\varepsilon' + e_2\\ &= \mathsf{J}_2\mathsf{J}_1\varepsilon^c + \mathsf{J}_2 e_1 + e_2,\end{aligned}$$

where

$$\begin{aligned}\mathsf{J}_{21} = \mathsf{J}_2\mathsf{J}_1 &= \begin{bmatrix} \mathsf{I}_n & \mathsf{O}_n \\ -(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21} & \mathsf{O}_n \end{bmatrix}\begin{bmatrix} \mathsf{O}_n & -(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12} \\ \mathsf{O}_n & \mathsf{I}_n \end{bmatrix}\\ &= \begin{bmatrix} \mathsf{O}_n & -(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12} \\ \mathsf{O}_n & (\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21}(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12} \end{bmatrix}.\end{aligned}$$

If we switch the order of the updates inside the IAS iteration, that is, first update θ then x, the error propagation matrix becomes

$$\mathsf{J}_{12} = \mathsf{J}_1\mathsf{J}_2 = \begin{bmatrix} (\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12}(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21} & \mathsf{O}_n \\ -(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21} & \mathsf{O}_n \end{bmatrix}.$$

Consider the error propagation of the leading term in the iteration, ignoring the second order term,

$$\cdots \xrightarrow{\mathsf{J}_2} \varepsilon_j \xrightarrow{\mathsf{J}_1} \varepsilon'_j \xrightarrow{\mathsf{J}_2} \varepsilon_{j+1} \xrightarrow{\mathsf{J}_1} \varepsilon'_{j+1} \xrightarrow{\mathsf{J}_2} \varepsilon_{j+2} \xrightarrow{\mathsf{J}_1} \cdots,$$

and partitioning the error vectors $\varepsilon_j, \varepsilon'_j \in \mathbb{R}^{2n}$ as

$$\varepsilon_j = \begin{bmatrix} \varepsilon_{j,x} \\ \varepsilon_{j,\theta} \end{bmatrix}, \quad \varepsilon'_j = \begin{bmatrix} \varepsilon'_{j,x} \\ \varepsilon'_{j,\theta} \end{bmatrix},$$

we find that the updating formula for the leading term for the x-component of the error term is

$$\varepsilon'_{j+1,x} = (\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12}(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21}\varepsilon'_{j,x}.$$

Substituting the actual expression of the blocks of the Hessian (12)–(14), and writing

$$\mathsf{A}^{\mathsf{T}}\mathsf{A} + \mathrm{diag}(1/\widehat{\theta}) = \mathrm{diag}(1/\widehat{\theta}^{1/2})\big[\mathrm{diag}(\widehat{\theta}^{1/2})\mathsf{A}^{\mathsf{T}}\mathsf{A}\mathrm{diag}(\widehat{\theta}^{1/2}) + \mathsf{I}\big]\mathrm{diag}(1/\widehat{\theta}^{1/2})$$

we obtain

$$\begin{aligned}(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12}(\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21} &= (\mathsf{A}^{\mathsf{T}}\mathsf{A} + \mathrm{diag}(1/\widehat{\theta}))^{-1}\mathrm{diag}(\widehat{x}/\widehat{\theta}^2)(\mathrm{diag}(\widehat{x}^2/\widehat{\theta}^3 + \eta/\widehat{\theta}^2))^{-1}\mathrm{diag}(\widehat{x}/\widehat{\theta}^2)\\ &= \mathrm{diag}(\widehat{\theta}^{1/2})\big[\mathrm{diag}(\widehat{\theta}^{1/2})\mathsf{A}^{\mathsf{T}}\mathsf{A}\mathrm{diag}(\widehat{\theta}^{1/2}) + \mathsf{I}_n\big]^{-1}\mathrm{diag}(\widehat{x}^2/(\widehat{x}^2 + \eta\widehat{\theta}))\mathrm{diag}(1/\widehat{\theta}^{1/2}),\end{aligned}$$

hence, up to a second order term,

$$\|\mathrm{diag}(1/\widehat{\theta}^{1/2})\varepsilon'_{j+1,x}\| \leqslant \mu\|\mathrm{diag}(1/\widehat{\theta}^{1/2})\varepsilon'_{j,x}\|, \tag{20}$$

for some $\mu < 1$. Furthermore, in light of the form of $\mathsf{J}_2 : \varepsilon'_j \to \varepsilon_{j+1}$, up to a second order correction term

$$\varepsilon'_{j,x} = \varepsilon_{j+1,x}, \quad \varepsilon'_{j+1,x} = \varepsilon_{j+2,x},$$

hence the ordering of the two updates is irrelevant for the error estimate. This proves that the IAS algorithm converges at least $\widehat{\theta}$ *-linearly*, that is, linearly with respect to the $\widehat{\theta}$ -weighted norm,

$$\|z\|_{\widehat{\theta}}^2 = z^{\mathsf{T}} \mathsf{D}_{\widehat{\theta}}^{-1} z. \tag{21}$$

We point out that this is the Mahalanobis norm with respect to the prior distribution at the MAP value.

Assume now that the global minimizer $\widehat{x}$ has support $S \subset \{1, 2, \ldots, n\}$. Without loss of generality, we may assume that $S = \{1, 2, \ldots k\}$, where $k = \|\widehat{x}\|_0 < n$. Consider the updating matrix J_{21}. We observe that up to a second order term,

$$\begin{aligned}\varepsilon_{j+1,\theta} &= (\widehat{\mathsf{H}}^{22})^{-1}\widehat{\mathsf{H}}^{21}(\widehat{\mathsf{H}}^{11})^{-1}\widehat{\mathsf{H}}^{12}\varepsilon_{j,\theta} \\ &= \mathrm{diag}(\widehat{\theta}\widehat{x}/(\widehat{x}^2 + \eta\widehat{\theta}))(\mathsf{A}^{\mathsf{T}}\mathsf{A} + \mathrm{diag}(1/\widehat{\theta}))^{-1}\mathrm{diag}(\widehat{x}/\widehat{\theta}^2)\varepsilon_{j,\theta},\end{aligned}$$

hence

$$\mathsf{D}_{\widehat{\theta}}^{-1/2}\varepsilon_{j+1,\theta} = \mathrm{diag}(\widehat{\theta}\widehat{x}/(\widehat{x}^2 + \eta\widehat{\theta}))[\mathrm{diag}(\widehat{\theta}^{1/2})\mathsf{A}^{\mathsf{T}}\mathsf{A}\mathrm{diag}(\widehat{\theta}^{1/2}) + \mathsf{I}]^{-1}\mathrm{diag}(\widehat{x}/\widehat{\theta}^2)\mathsf{D}_{\widehat{\theta}}^{-1/2}\varepsilon_{j,\theta}. \tag{22}$$

In particular, this formula shows that, up to a second order term,

$$\big(\varepsilon_{j+1,\theta}\big)_\ell = 0 \text{ for } \ell > k,$$

that is, outside the support of $\widehat{x}$, the convergence is quadratic.

We collect the main results of this section in the following theorem.

Theorem 4.2. *In the IAS algorithm, the updates of x converge at least* $\widehat{\theta}$ *-linearly, that is, linearly in the Mahalanobis norm (21) evaluated at the MAP estimate. Moreover, if* $\mathrm{supp}(\widehat{x}) \subsetneq \{1, 2, \ldots, n\}$*, the convergence of* θ *in the complement of the support is quadratic.*

The 'ℓ_1-magic' results state that for matrices A satisfying certain conditions the ℓ_1-penalized solution is the exact solution of the ℓ_0-penalized problem when the support of the signal is appropriately smaller than the number of observations. The exact recovery is based on the assumption that there is no noise in the data. With noisy data, the reconstruction is not necessarily exact, however, the discrepancy between the exact and recovered signal is bounded by a small multiple of the norm of the noise. In the compressed sensing literature, a signal in $\mathbb{R}^n$ whose components outside a set $S \subsetneq \{1, 2, \ldots, n\}$ are below a given small threshold is referred to as compressible. The Bayesian framework at the foundation of the IAS algorithm is based on the assumption that the data is corrupt by additive Gaussian noise, hence even in the limit as the shape parameter η tends to zero, the best that we can expect is to recover a signal whose distance from the underlying sparse source is bounded above by a small multiple of the norm of the error in the data. Observe that since $\theta_j \geqslant \eta$, the prior assumption for $\eta > 0$ small is that the signal is not necessarily sparse but compressible.

If $\widehat{x}$ is compressible, and for some threshold value $\delta > 0$,

$$|\widehat{x}_j| < \delta^{1/2}\eta^{3/2}, \quad j \notin S, \tag{23}$$

for $j \notin S$,

$$\left|\frac{\widehat{\theta}_j\widehat{x}_j}{\widehat{x}_j^2 + \eta\widehat{\theta}_j}\right| < \left|\frac{\widehat{x}_j}{\eta}\right| < \sqrt{\delta\eta},$$

and since $\widehat{\theta}_j > \eta$, we have

$$\left|\frac{\widehat{x}_j}{\widehat{\theta}_j^2}\right| \leqslant \sqrt{\frac{\delta}{\eta}}.$$

Therefore, if we denote by P^c the orthogonal projection on the complement of S, we find that

$$\|\mathsf{P}^c \mathrm{diag}(\widehat{\theta}\widehat{x}/(\widehat{x}^2+\eta\widehat{\theta}))[\mathrm{diag}(\widehat{\theta}^{1/2})\mathsf{A}^{\mathsf{T}}\mathsf{A}\mathrm{diag}(\widehat{\theta}^{1/2})+\mathsf{I}]^{-1}\mathrm{diag}(\widehat{x}/\widehat{\theta}^2)\mathsf{P}^c\| < \delta.$$

This estimate proves the following theorem.

Theorem 4.3. *Assuming that the MAP estimate $\widehat{x}$ is compressible and satisfies (23), the update of θ outside the support S is essentially quadratic,*

$$\|\mathsf{P}^c\varepsilon_{j+1,\theta}\|_{\widehat{\theta}} < \delta\|\mathsf{P}^c\varepsilon_{j,\theta}\|_{\widehat{\theta}} + \text{second order correction}.$$

We will illustrate the results of convergence and the effect of the scaling in the following section through computed examples.

5. Computed examples

In this section we elucidate some of the results with computed examples. In particular, examples 5.1 and 5.2 demonstrate the convergence rate of the IAS algorithm, while example 5.3 highlights the effect of the sensitivity weighting through the hyperparameter θ^*. Finally, example 5.4 elucidates the properties of the algorithm vis-à-vis sparsity properties of the recovered signal in the context of sparse recovery theory.

5.1. Example 1

We consider a one-dimensional deconvolution problem of the form

$$g(t) = \int_0^1 a_w(t-s)f(s)\mathrm{d}s, \quad a_w(t-s) = \frac{1}{\sqrt{2\pi w^2}}\mathrm{e}^{-(t-s)^2/2w^2}, \quad w = 0.01,$$

and its discretized version obtained by approximating the values of the integral at points $t_j = (j-1)/n$, $n = 128$, using a quadrature rule with $n = 128$ nodes $s_j = (j-1)/n, 1 \leqslant j \leqslant n$, and assuming that the data are contaminated by additive scaled white noise. The resulting $n \times n$ linear system is

$$b = \mathsf{A}x + \epsilon, \quad \mathsf{A}_{jk} = \frac{1}{n}\frac{1}{\sqrt{2\pi w^2}}\mathrm{e}^{-(t_j-s_k)^2/2w^2}, \quad 1 \leqslant j,k \leqslant n,$$

where $\epsilon \sim \mathcal{N}(0,\sigma^2\mathsf{I}_n)$. Our main aim here is to show the performance of the IAS for the recovery of a sparse signal and to verify the convergence rate. In line with the standard practice in compressive sensing literature, we ignore the fact that in actual applications the data may not arise from a model used to solve the inverse problem. The signal $x^\sharp$ used to generate the blurred noisy data shown in figure 1 is sparsely supported, with $\|x^\sharp\|_0 = 6$.

To compute θ^*, we assume that the estimated signal-to-noise ratio is $\mathrm{SNR} = 255$, which would correspond to noise variance

$$\sigma^2 = \frac{\|b_0\|^2}{n(\mathrm{SNR}-1)}, \quad b_0 = \mathsf{A}x^\sharp, \tag{24}$$

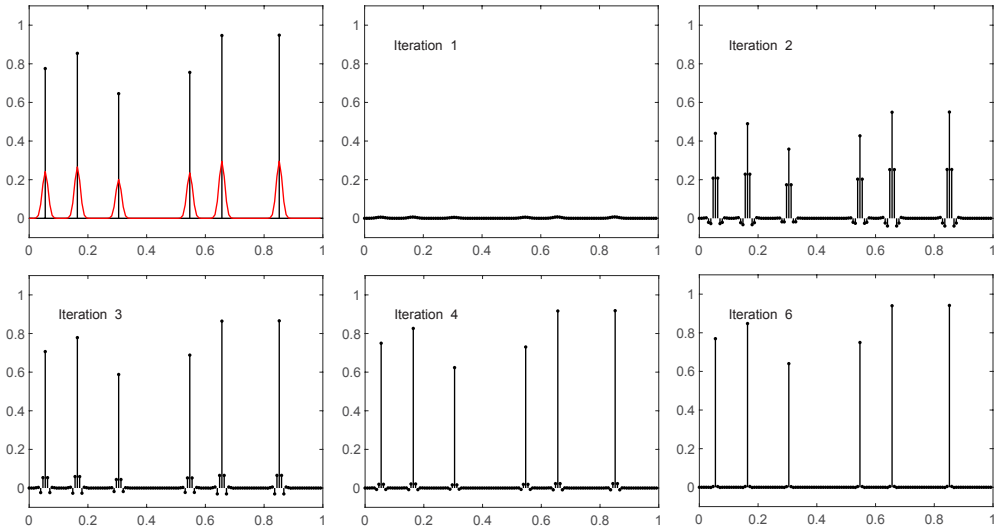

Figure 1. The one-dimensional deconvolution problem: top row, leftmost figure shows the true signal and the computed noisy data in red. The subsequent panels, in lexicographical order, show the progress of the iteration. After the sixth iteration, the changes in the approximate solution are visually indiscernible.

or $\sigma \approx 0.0053$, or a noise level of approximately 1.8% of the maximum of the noiseless signal. However, we run this example with data in which no artificial noise is added.

We compute the MAP estimate via the IAS algorithm with shape hyperparameter $\eta = 10^{-6}$, calculating the values of the scale hyperparameters θ_j^* from the formula (7) of theorem 3.3, assuming that, *a priori*, we expect to have at most $n_{\max} = 20$ non-zero entries, with uniform probability for the cardinality of the support,

$$p_j = \begin{cases} 0 & j = 0 \\ 1/n_{\max} & 1 \leqslant j \leqslant n_{\max} \\ 0 & j > n_{\max} \end{cases}.$$

Not surprisingly, the components of θ^* are almost constants, $\theta_j^* \approx 0.493$, except near the endpoints of the interval, where part of the Gaussian kernel leaks out of the interval. This is reflected in the fact that $\theta_1^* = \theta_n^* \approx 0.685$, $\theta_2^* = \theta_{n-1}^* \approx 0.514$.

Figure 1 shows how the IAS iterates approximate the underlying sparse signal, progressively flattening the signal in the complement of the support. After six IAS iterations, the solution stabilizes, correctly identifying six significant components in an almost vanishing background. The background level is not exactly zero, but of the order $\sim 10^{-6}$.

The leftmost panel of figure 2 shows the relative change in the norm of the variance parameter vector $\Delta_\theta^j = \|\theta^j - \theta^{j-1}\|/\|\theta^j\|$ from one iteration to the next, a quantity that can be used to design a stopping rule for the IAS iteration. The plot shows that after seven iterations, the relative change in θ has dropped below 10^{-2}, and after sixteen iterations, below 10^{-4}. The center left panel in figure 2 shows the logarithmic plot of the error $\|\varepsilon_{\theta,j}\| = \|\theta^j - \widehat{\theta}\|$ versus $\|\varepsilon_{\theta,j+1}\|$, using the final value of the iterations as an approximation of the minimizer $\widehat{\theta}$. The dashed lines in the plot can be used to identify linear (red) and quadratic (blue) convergence rates. For comparison, the center right panel shows the same convergence analysis, but using the $\widehat{\theta}$-weighted norm, where $\widehat{\theta}$ is approximated by the last iterate of θ. The convergence

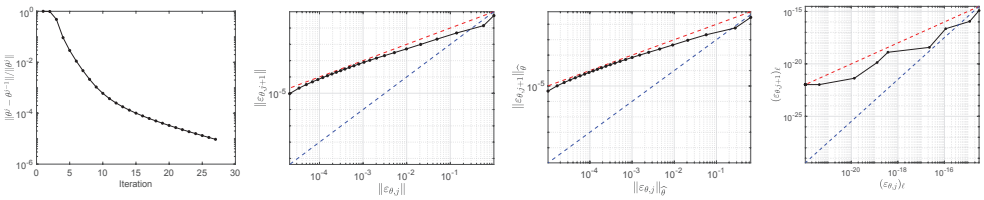

Figure 2. Left: the relative change $\Delta_\theta^j = \|\theta^j - \theta^{j-1}\|/\|\theta^j\|$ of the norm of the variance parameter vector as a function of iteration. Left middle: logarithmic plot of the norm of consecutive errors $\|\varepsilon_{\theta,j}\| = \|\theta^j - \widehat{\theta}\|$ in the variance parameter. The red dashed line indicates the linear rate of decrease, the blue one the quadratic rate. Right middle: the convergence rate measured by using the Mahalanobis norm at the MAP estimate, with $\widehat{\theta}$ approximated by the last iterate of θ. Right: convergence of θ_ℓ at a point corresponding to minimum value of the final estimate of θ.

rate of the norm of the error is at least linear, and at the beginning of the iterations close to quadratic. Also, the convergence rate with the weighted and non-weighted norm are identical, indicating that the latter can be used to estimate the convergence rate. The convergence of individual error components are qualitatively similar; exact quadratic convergence cannot be expected in practice since the minimizer $\widehat{x}$ has a small non-zero background value outside the outstanding peak values. The right panel in figure 2 shows the individual convergence history at the point where the estimated θ attains its minimum.

5.2. Example 2

The second example, similar to the previous one but in two dimensions, confirms the results and shows that the algorithm retains its efficiency in larger scale problems. We want to recover a 'nearly black object', consisting of an image over the square $[0,1] \times [0,1]$ with only few non-zero pixels. The source image $x^\sharp$ of size $n \times n$, for $n = 128$, with $\|x^\sharp\|_0 = 50$ is shown in top left panel of figure 3. The data arises from the nearly black object blurred with a Gaussian kernel of width $w = 0.01$ to which Gaussian scaled white noise is added, using an estimated a signal-to-noise ratio $\mathrm{SNR} = 25$, and standard deviation $\sigma \approx 4.4 \times 10^{-3}$, corresponding to 2.3% of the maximum of the noiseless signal. The data are shown in the top center panel of figure 3. The computation of the hyperparameter vector θ^* assumes that the cardinality of the support is at most 100 pixels, with uniform probability for cardinality support between 1 and 100, that is $p_j = 1/100$ for $1 \leqslant j \leqslant 100$.

As in the previous example, we are interested in the convergence rate of the IAS algorithm. The quality of the approximation of the IAS iterates can be assessed visually by looking at the reconstructions shown in figure 3. Although after the tenth iteration the restored image remains visually unchanged, we carry out 50 iterations. The left panel of figure 4 shows the relative change in the norm of the variance parameter θ from one iteration to the next, while the right panel illustrates how the decrease in the norm of the error is in agreement with the theoretical results.

5.3. Example 3

The third example elucidates the importance of the sensitivity weights through an appropriate choice of the scale parameter vector θ^*. In this example, we consider an inverse source problem with an observation model

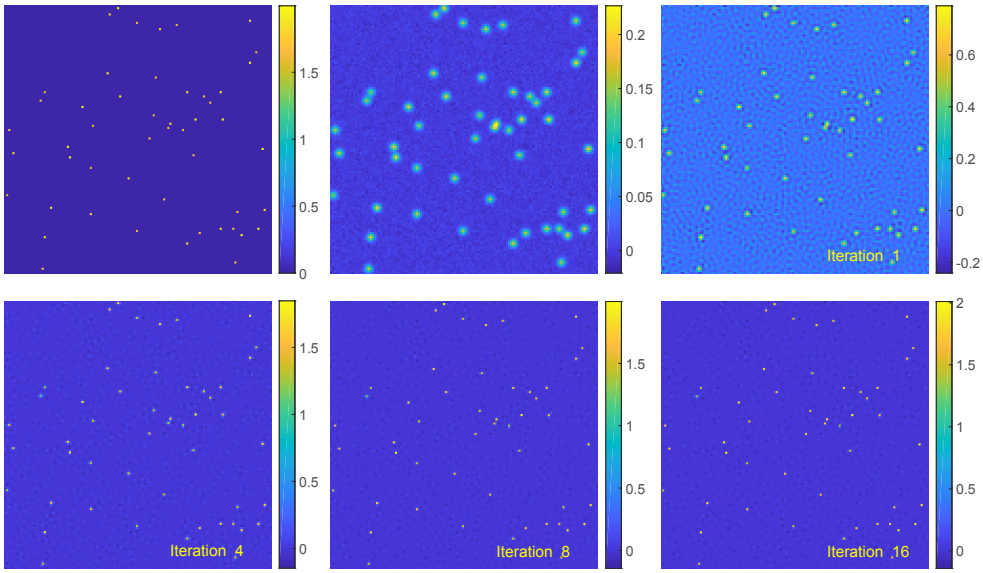

Figure 3. The underlying sparse image (top left) and the observed blurred and noisy version (top center). The remaining panels show, in lexicographical order, the reconstruction computed in the IAS iterates. After ten iterations, the results are visually unaltered.

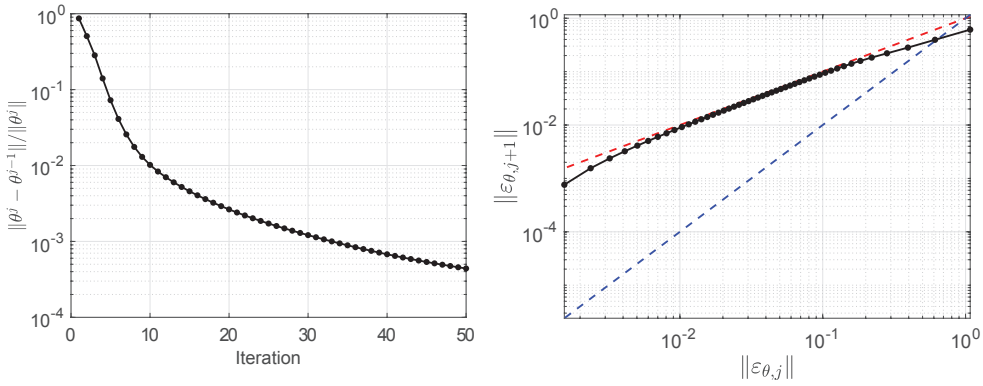

Figure 4. The relative change in the norm of the variance vector θ as a function of iterations for the two-dimensional deconvolution problem (left), and the plot of the error in θ versus the error in the previous iteration round. The dashed red line indicates the linear rate, the dashed blue line the quadratic rate.

$$b(R) = \int_{\Omega} \frac{\rho(r)}{|r-R|^2} \mathrm{d}r,$$

where $\Omega \subset \mathbb{R}^2$ is the source domain, and the observation points R are located outside the domain. We choose Ω to be the unit square, $\Omega = [0,1] \times [0,1]$, and a discrete set of observation points R_j, $1 \leqslant j \leqslant m$ chosen outside Ω but near three of the sides of it, see figure 5. To generate the data, we pick three points in Ω and place a point source at them.

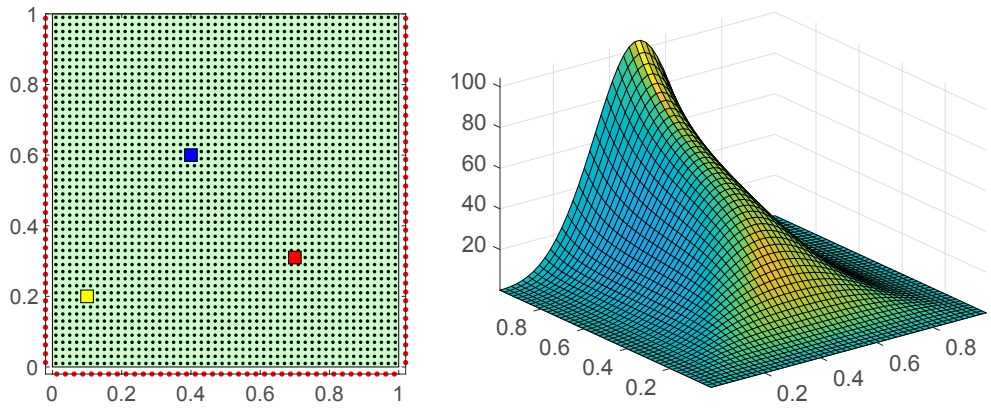

Figure 5. Left: the measurement configuration. The red dots are the observation points, $m = 120$, and the black dots denote the discretization points corresponding the forward model, $n = 2500$. The data are generated by placing three point sources, not in the grid points, with amplitudes $q_1 = q_2 = 5$ (blue and red), and $q_3 = 0.5$ (yellow). On the left, the computed θ^* is shown as a surface plot.

We discretize the forward model by assuming the source to be a sum of discrete point sources at fixed grid points, leading to a forward model

$$b_k = \sum_{j=1}^{n} \frac{q_j}{|r_j - R_k|^2}, \quad 1 \leqslant k \leqslant m,$$

and we set $m = 120$, $n = 2500$ as indicated in figure 5. In the same figure, the computed θ^* vector is shown as a surface plot, indicating that the data are highly sensitive to points near the three edges and much less sensitive to sources near the fourth edge. We generate the data, adding scaled white noise of standard deviation σ approximately 0.4% of the maximum of the noiseless signal, or $\mathrm{SNR} = 20\,000$. We run the algorithm using the focality parameter $\eta = 10^{-6}$. For comparison, we then run the same algorithm by choosing θ^* to be constant. We choose the constant value equal to the boundary values of θ^* in the previous case. Figure 6 shows the solutions after 100 IAS iterations. As expected, the solution without sensitivity weighting is completely concentrated on boundary pixels, while with the sensitivity weighting, one can reasonably identify the three sources.

5.4. Example 4

In his example, the underlying signal itself is not sparse, but it admits a sparse representation. More precisely, we consider the discrete linear observation model

$$b = \mathsf{A}x + \epsilon, \quad \mathsf{A} \in \mathbb{R}^{n\times n}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_n),$$

and assume *a priori* that there is an invertible matrix $\mathsf{L} \in \mathbb{R}^{n\times n}$ such that we can express the signal as

$$x = \mathsf{L}z,$$

in terms of a sparse vector $z \in \mathbb{R}^n$ that we model as a random variable with a hierarchical conditionally Gaussian distribution. In this example, we assume that the blurring kernel is an Airy kernel,

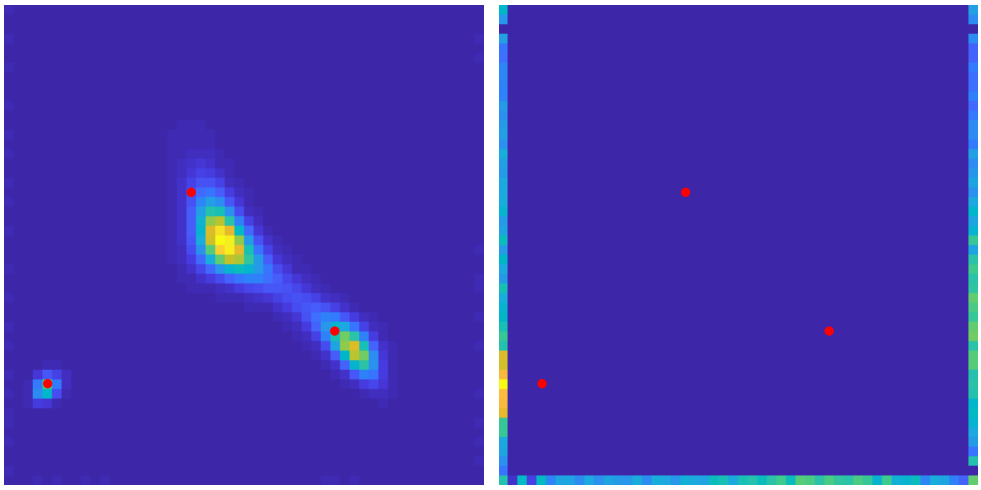

Figure 6. Reconstructions of the sources after 100 iterations using the automatic sensitivity weighting scheme (left), and a constant value (right) for the scaling vector θ^*. In the latter, the solution is entirely concentrated on the domain boundary near the observation points.

$$\mathsf{A}_{jk} = C\left(\frac{J_1(\lambda(t_j - s_k))}{\lambda(t_j - s_k)}\right)^2,$$

where J_1 is the Bessel function of the first kind of order 1, $\lambda = 40$ is a width parameter, the points $s_k, 1 \leqslant k \leqslant 128 = n$ are uniform grid points in the interval $[0, 1]$, while the observations are limited to every sixth grid point $t_j, 1 \leqslant j \leqslant 22 = m$, and, finally, $C > 0$ is a scaling factor. Let L be the backward differencing matrix,

$$\mathsf{L} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

and reformulate the problem in terms of z, whose posterior distribution is

$$\pi_{z,\theta|b}(z, \theta \mid b) \propto \exp\left(-\frac{1}{2\sigma^2}\|b - \mathsf{A}\mathsf{L}^{-1}z\|^2 - \frac{1}{2}\sum_{æ=1}^{n}\frac{z_j^2}{\theta_j} - \sum_{j=1}^{n}\left(\frac{\theta_j}{\theta_j^*} - \eta\log\frac{\theta_j}{\theta_j^*}\right)\right).$$

The believed sparsity of z_j is tantamount to assuming that x has sparse increments, since it is a straightforward matter to check that

$$x_j = \sum_{i=1}^{j} z_i, \quad 1 \leqslant j \leqslant n. \tag{25}$$

Moreover, we assume *a priori* that the absolute values of the jumps z_j are expected to be below a value $M = 1$ with high probability.

An interesting observation is that, while the norms of the columns of A are essentially constant, except near the endpoints of the intervals, as pointed out in example 5.1, this is no

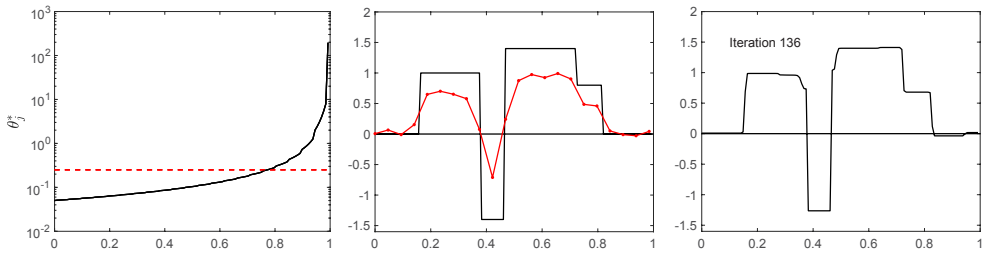

Figure 7. On the left, the plot of the scaling vector θ^*. In the middle, the original signal and the subsampled noisy observation of it using the Airy kernel. On the right, the IAS reconstruction with stopping criterion $\Delta_\theta^j < 10^{-3}$.

longer the case for AL^{-1}, due to the asymmetric roles of the endpoints $j = 1$ and $j = n$ for backwards finite differencing.

It follows from formula (25) that, since the observation b_j depends on all z_i, $i \leqslant j$, the data are more sensitive to the components of z near the left endpoint of the interval than to those near the right endpoint. In other words, even if A is symmetric and the convolution kernel is translation invariant, the sensitivity is not, and this fact is accounted for when the hyperparameter θ^* is set according to (7) with A replaced by AL^{-1}.

Let the underlying signal $x^\sharp$ be the piecewise constant function with five discontinuities shown in black in the left panel of figure 7, so that $\|z^\sharp\|_0 = 5$, we generate the noisy data by multiplying the discretized signal by the matrix A, and adding white Gaussian noise scaled so that the signal-to-noise ratio is approximately $\mathrm{SNR} = 15$. It follows from (24) that the noise level in this case is $\sigma \approx 0.22$, which is approximately 15% of the maximum of the noiseless signal. One random realization of the blurred noisy signal is shown in red in the left plot of figure 7. The right panel of figure 7 shows a plot of the sensitivity vector θ_j^* for z based on formula (7), increasing several orders of magnitude towards the end of the interval. To implement the prior belief that the jumps are not likely exceeding the value $M = 1$, we define the effective hyperparameter $\theta^*_{\mathrm{eff},j}$ using formula (8) of remark 3.5. The cut-off level is indicated in figure 7 by a red dashed line.

To put the algorithm in the context of sparse recovery, we compute the mutual coherence of the matrix, defined as

$$\mu(\widetilde{\mathsf{A}}) = \max_{j\neq k} \frac{|(a^{(j)})^{\mathsf{T}} a^{(k)}|}{\|a^{(j)}\|\|a^{(k)}\|},$$

where $a^{(j)}$ is the jth column of $\widetilde{\mathsf{A}}$, see [1, 12]. In the cited articles, it has been shown that the mutual coherence provides a lower bound of the spark of the matrix $\widetilde{\mathsf{A}} = \mathsf{AL}^{-1}$,

$$\mathrm{spark}(\widetilde{\mathsf{A}}) \geqslant 1 + \frac{1}{\mu(\widetilde{\mathsf{A}})},$$

defined as the smallest number of columns of the matrix that are linearly dependent. Vectors in the null space of $\widetilde{\mathsf{A}}$ must therefore satisfy $\|z\|_0 \geqslant \mathrm{spark}(\widetilde{\mathsf{A}})$. Mutual coherence is important in sparse recovery theory, as it can be shown that algorithms such as orthogonal greedy algorithm are guaranteed to find a sparse solution of the linear system provided that there is a solution satisfying

$$\|z\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\widetilde{A})}\right). \tag{26}$$

Consequently, the smaller the mutual coherence, the wider the class of sparse signals for which the success is guaranteed. In the present example, we have $\mu(\widetilde{A}) \approx 1 - 1.5 \times 10^{-5}$, that is, guarantees of sparse recovery exist only for signals with a single discontinuity.

The left panel of figure 7 shows a reconstruction from data corresponding to a true signal that does not satisfy the support condition (26) with noise level SNR $= 25$; the blurred noisy data are plotted over the underlying true signal. The right panel in the same figure displays the IAS reconstruction, where the locations of all discontinuities are correctly identified, and the amplitudes are found with reasonable precision, even when the columns of the matrix defining the forward map are highly coherent.

6. Conclusions

This article proposes an alternative approach to the sparse recovery problem based on Bayesian analysis of the inverse problem. Although the focus of this work is on an algorithm for computing the MAP estimate, and therefore does not take full advantage of properties of the posterior distribution, the analysis shows the usefulness of the probabilistic extension. The MAP estimation algorithm has a unique global minimum, the alternating algorithm is easy to implement, and as shown in this article, rapidly converging. Moreover, the Bayesian analysis of the signal-to-noise ratio combined with the statistically well-motivated exchangeability condition leads to a versatile sensitivity scaling that helps understanding, e.g. the depth weighting schemes used in geophysics and biomedical applications, putting them on a solid basis in terms of the assumptions about the noise and support of the source. The algorithm contains few user-supplied parameters, and the parameters have a clear interpretation. Numerical tests indicate that the algorithm is not very sensitive to the choice of the signal-to-noise ratio or the estimated support of the presumably sparse signal, however, the tests suggest that grossly underestimating or overestimating the support may lead the algorithm astray. Although the emphasis in this article is on sparse recovery, the IAS algorithm is not restricted to cases in which the source is sparse. In fact, using larger shape parameter values, numerical evidence points towards good recovery of distributed targets also. Such analysis is left for future work.

Acknowledgments

This work was partly supported by the NSF grants DMS-1522334(DC and AS), and DMS-1714617 (ES).

ORCID iDs

D Calvetti https://orcid.org/0000-0001-5696-718X

References

[1] Bruckstein A M, Donoho D L and Elad M 2009 From sparse solutions of systems of equations to sparse modeling of signals and images *SIAM Rev.* **51** 43–81
[2] Calvetti D and Somersalo E 2007 *An Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing* (New York: Springer)

[3] Calvetti D and Somersalo E 2007 A Gaussian hypermodel to recover blocky objects *Inverse Problems* **23** 733
[4] Calvetti D, Hakula H, Pursiainen S and Somersalo E 2009 Conditionally Gaussian hypermodels for cerebral source localization *SIAM J. Imaging Sci.* **2** 879–909
[5] Calvetti D, Pascarella A, Pitolli F, Somersalo E and Vantaggi B 2015 A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning *Inverse Problems* **31** 125005
[6] Calvetti D, Pascarella A, Pitolli F, Somersalo E and Vantaggi B 2018 Brain activity mapping from MEG data via a hierarchical Bayesian algorithm with automatic depth weighting *Brain Topography* 1–31 https://doi.org/10.1007/s10548-018-0670-7
[7] Candes E, Romberg J and Tao T 2006 Stable signal recovery from incomplete and inaccurate measurements *Commun. Pure Appl. Math.* **59** 1207–23
[8] Candes E J and Tao T 2005 Decoding by linear programming *IEEE Trans. Inf. Theory* **51** 4203–15
[9] Daubechies I, Devore R, Fornasier M and Güntürk C S 2010 Iteratively reweighted least squares minimization for sparse recovery *Commun. Pure Appl. Math.* **63** 1–38
[10] Dennis J E and Schnabel R B 1996 *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Philadelphia, PA: SIAM)
[11] Donoho D L 2006 For most large undetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution *Commun. Pure Appl. Math.* **59** 0907–34
[12] Donoho D L, Elad M and Temlyakov V 2006 Stable recovery of sparse overcomplete representations in the presence of noise *IEEE Trans. Inf. Theory* **52** 6–18
[13] Gorodnitsky I F and Rao B D 1997 Sparse signal reconstruction from limited data using FOCUSS, a re-weighted minimum norm algorithm *IEEE Trans. Signal Process.* **45** 600–16
[14] Kaipio J and Somersalo E 2004 *Statistical and Computational Inverse Problems* (New York: Springer)
[15] Lawson C L 1961 Contributions to the theory of linear least maximum approximation *PhD Thesis* University of California, Los Angeles
[16] Lin F H, Witzel T, Ahlfors S P, Stufflebeam S M, Belliveau J W and Hämäläinen M S 2006 Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates *NeuroImage* **31** 160–71
[17] Li Y and Oldenburg D W 1996 3D inversion of magnetic data *Geophysics* **61** 394–408
[18] Li Y and Oldenburg D W 1998 3D inversion of gravity data *Geophysics* **63** 109–19
[19] Phillips C, Rugg M D and Friston K J 2002 Systematic regularization of linear inverse solutions of the EEG source localization problem *NeuroImage* **17** 287–301
[20] Uutela K, Hämäläinen M and Somersalo E 1999 Visualization of magnetoencephalographic data using minimum current estimates *NeuroImage* **10** 173–80