

# OVERCOMPLETE REPRESENTATION IN A HIERARCHICAL BAYESIAN FRAMEWORK

MONICA PRAGLIOLA\*

University of Bologna, Department of Mathematics  
Piazza di Porta San Donato 5, Bologna, Italy

DANIELA CALVETTI AND ERKKI SOMERSALO

Case Western Reserve University, Department of Mathematics  
Applied Mathematics and Statistics  
10900 Euclid Avenue, Cleveland, OH 4410, USA

(Communicated by Martin Burger)

**ABSTRACT.** A common task in inverse problems and imaging is finding a solution that is sparse, in the sense that most of its components vanish. In the framework of compressed sensing, general results guaranteeing exact recovery have been proven. In practice, sparse solutions are often computed combining  $\ell_1$ -penalized least squares optimization with an appropriate numerical scheme to accomplish the task. A computationally efficient alternative for finding sparse solutions to linear inverse problems is provided by Bayesian hierarchical models, in which the sparsity is encoded by defining a conditionally Gaussian prior model with the prior parameter obeying a generalized gamma distribution. An iterative alternating sequential (IAS) algorithm has been demonstrated to lead to a computationally efficient scheme, and combined with Krylov subspace iterations with an early termination condition, the approach is particularly well suited for large scale problems. Here the Bayesian approach to sparsity is extended to problems whose solution allows a sparse coding in an overcomplete system such as composite frames. It is shown that among the multiple possible representations of the unknown, the IAS algorithm, and in particular, a hybrid version of it, is effectively identifying the most sparse solution. Computed examples show that the method is particularly well suited not only for traditional imaging applications but also for dictionary learning problems in the framework of machine learning.

**1. Introduction.** Sparsity promoting methods and algorithms for inverse problems and imaging applications have been extensively studied in the past decades, and they continue to be a very active field of research. The interest in compressed sensing has motivated a significant part of the works on the topic. The starting point of many sparse reconstruction problems is a *dictionary*, intended as a collection of elements in the ambient space referred to as *atoms* [17], used to represent the unknown quantity of interest. The dictionary may be selected according to some *a priori* information available on the problem of interest or, alternatively, its

---

2020 *Mathematics Subject Classification.* Primary: 62F15, 94A08.

*Key words and phrases.* Krylov subspace solver, sparsity, dictionary learning, composite frames.

MP acknowledges the National Group for Scientific Computation (GNCS-INDAM). The work of DC was partly supported by NSF grants DMS-1522334 and DMS-1951446, and the work of ES was partly supported by NSF grant DMS-1714617.

\*Corresponding author: Monica Pragliola.

formation can be data-driven - see, e.g., [18, 3, 14]. Recently, approaches aimed at learning the dictionary while jointly recovering the signal have also been developed [13]. Typically, the cardinality of the dictionary is significantly larger than the dimension of the ambient space. When the atoms in the dictionary do not form a basis for the ambient space, the dictionary is called *redundant* or *overcomplete*. The use of redundant dictionaries has proved to be a useful strategy in terms of artifact reduction, especially in the framework of signal denoising problems [20, 19].

Let  $x \in \mathbb{R}^n$  be an unknown signal, and let  $W = \{w_i\}_{i=1}^N$  be a dictionary with atoms  $w_i \in \mathbb{R}^n$ . We arrange the atoms as columns of the dictionary matrix,  $W \in \mathbb{R}^{n \times N}$ , with  $n \ll N$ , and refer to this matrix as the dictionary. In a synthesis perspective, a sparse reconstruction problem is the task of recovering a sparse vector  $\alpha \in \mathbb{R}^N$ , with most of its components vanishing, that represents the original signal in terms of  $W$ ,  $x = W\alpha$ , starting from a corrupted and possibly poorly sampled indirect observation  $b \in \mathbb{R}^m$  of  $x$ , with  $m \leq n$ . Assuming that the observation is linear in  $x$  and the noise is additive, the sparse dictionary representation can be formulated as an optimization problem of the form

$$(1) \quad \text{minimize } \|\alpha\|_0 \text{ such that } b = Ax + \varepsilon, \text{ with } x = W\alpha,$$

where  $\|\alpha\|_0 = \text{card}(\text{supp}(\alpha))$ ,  $A \in \mathbb{R}^{m \times n}$  is the forward model operator, and  $\varepsilon \in \mathbb{R}^m$  is an additive noise vector. Addressing the minimization problem (1) directly is a challenge due to its NP-hardness, thus explaining the need for alternative approaches. A strategy which has been widely explored and goes under the name of *basis pursuit* replaces the  $\ell_0$ -(semi)norm with its  $\ell_1$  convex relaxation [15].

When the signal itself is known to be compressible, i.e.,  $\text{card}(\{x_i \mid |x_i| < \epsilon\}) \ll n$  with  $\epsilon > 0$  arbitrarily small, and the forward model operator  $A$  satisfies the *restricted isometry property* (RIP) condition, an optimal bound for the error  $\|x - \hat{x}\|_2$ , with  $\hat{x}$  denoting the recovered signal, has been derived. Moreover, if  $x$  is sufficiently sparse, the signal can be recovered exactly [11]. When the signal itself is not sparse, but it allows a sparse or compressible representation in a given dictionary, the exact recovery results still hold, provided that  $A$  satisfies a *restricted isometry property adapted to a dictionary* (D-RIP) condition [12]. Given the theoretical motivation, a significant amount of research is devoted to identifying classes of operators to which these results could be applied.

Besides the convex approaches,  $\ell_p$ -norms with  $p < 1$  have also been considered in place of the  $\ell_0$ -(semi)norm in problem (1), as they are known to promote sparsity more strongly than the case  $p = 1$ . Nonetheless, the presence of local minima is a clear limitation to the reliability of non-convex strategies.

The sparse reconstruction problem for linear inverse problems allows a natural formulation in the Bayesian computational framework, with the notion of sparsity promoting priors. In a number of previous works [10, 9, 8], the recovery of a sparse signal  $x$  - or of a signal admitting a sparse representation in a given basis - has been addressed by modeling its entries in a hierarchical Bayesian framework as conditionally Gaussian random variables with unknown variances, with a generalized gamma hyperprior distribution. The sparsity promotion and the convexity properties of the corresponding class of hypermodels have been studied in [10, 9]. The derived results, and in particular the considerations on the convexity properties of the resulting maximum a posteriori estimation problems, motivated the introduction of a hybrid hypermodel combining the strong sparsity promotion that typically characterizes non-convex settings with the convexity guarantees [8].

In this article, the hierarchical Bayesian framework outlined in the previous works, and in particular the use of the hybrid algorithm of [8], is extended to address sparse recovery problems in presence of redundant dictionaries. We consider a version of the iterated alternating sequential (IAS) algorithm that combines ideas from the Bayesian inference and iterative Krylov subspace methods, suitable for large scale problems, and is therefore particularly attractive for problems with large dictionaries. Numerical examples demonstrate the computational efficiency of the approach, and most importantly, show that for composite frame dictionaries, where each subdictionary would provide a sufficient representation of the signal, the method is capable of identifying an optimally sparse representation.

**2. Hierarchical Bayesian formulation.** Consider the linear inverse problem

$$b = Ax + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma) \quad \text{such that} \quad x = W\alpha,$$

where  $A \in \mathbb{R}^{m \times n}$ , with  $m \leq n$ , is the known forward model operator,  $x \in \mathbb{R}^n$  is the unknown of interest and  $\Sigma \in \mathbb{R}^{m \times m}$  is the symmetric positive definite covariance matrix of the additive Gaussian noise. In addition, we assume that  $x$  admits a representation in the redundant dictionary  $W \in \mathbb{R}^{n \times N}$ , with  $n \ll N$ , where the unknown vector  $\alpha \in \mathbb{R}^N$  is sparse, i.e.  $\|\alpha\|_0 \ll N$ , either because  $x$  can be naturally described by few atoms in the dictionary or because  $x$  needs to be compressed.

In a number of previous contributions [10, 9, 8], the *a priori* sparsity belief on the unknown has been exploited by modeling its entries as independent random variables following a *conditionally Gaussian* distribution, i.e.,

$$\alpha_j \mid \theta_j \sim \mathcal{N}(0, \theta_j), \quad 1 \leq j \leq N,$$

or, in an equivalent compact form,

$$\alpha \mid \theta \sim \mathcal{N}(0, D_\theta), \quad D_\theta = \text{diag}(\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times N}.$$

The conditional Gaussian prior on  $\alpha$  given the vector  $\theta$  takes the form

$$\begin{aligned} \pi_{\alpha|\theta}(\alpha \mid \theta) &\propto \frac{1}{\prod_{j=1}^N \sqrt{\theta_j}} \exp\left(-\frac{1}{2} \|D_\theta^{-1/2} \alpha\|^2\right) \\ &= \exp\left(-\frac{1}{2} \|D_\theta^{-1/2} \alpha\|^2 - \frac{1}{2} \sum_{j=1}^N \log \theta_j\right). \end{aligned}$$

According to the Bayesian paradigm, the unknown vector of variances  $\theta$  is also modeled as a random variable. The *a priori* beliefs about  $\theta$  are encoded in the *hyperprior*  $\pi_\theta(\theta)$ , and the joint prior on the coupled vector of unknowns  $(\alpha, \theta)$  reads

$$(2) \quad \pi_{(\alpha, \theta)}(\alpha, \theta) = \pi_{\alpha|\theta}(\alpha \mid \theta) \pi_\theta(\theta).$$

In [9], the authors propose to model the unknown variances  $\theta_j$  as mutually independent random variables following a generalized gamma distribution,

$$(3) \quad \pi_\theta(\theta) = \pi_\theta(\theta \mid r, \beta, \vartheta) = \frac{|r|^n}{\Gamma(\beta)^n} \prod_{j=1}^N \frac{1}{\vartheta_j} \left(\frac{\theta_j}{\vartheta_j}\right)^{r\beta-1} \exp\left(-\left(\frac{\theta_j}{\vartheta_j}\right)^r\right),$$

where  $r \in \mathbb{R} \setminus \{0\}$ ,  $\beta > 0$ ,  $\vartheta_j > 0$ . This choice is motivated by the observation that generalized gamma distributions tend to favor values which are close to the expected value while also allowing for few outliers very far from the mean. Presumably, the

outlier variances give rise to the few non-zero values of  $\alpha$ , or values above a tiny threshold.

The information about the observation process is encoded in the likelihood distribution, which in view of the additive Gaussian noise model, takes the form

$$\pi_{b|\alpha}(b | \alpha) \propto \exp \left( -\frac{1}{2} \|\mathbf{S}(\mathbf{A}\mathbf{W}\alpha - b)\|_2^2 \right),$$

where  $\mathbf{S}$  is the Cholesky factor of the precision matrix  $\Sigma^{-1}$ , i.e.  $\Sigma^{-1} = \mathbf{S}^\top \mathbf{S}$ . If the matrix  $\Sigma$  and thereby  $\mathbf{S}$ , are known, without loss of generality we can assume the noise to be white, i.e.  $\Sigma = \mathbf{I}$ , because it can be whitened by a linear transform on  $\mathbf{A}$  and  $b$ , namely

$$\mathbf{A} \longrightarrow \mathbf{S}\mathbf{A}, \quad b \longrightarrow \mathbf{S}b.$$

Under the white normal noise assumption, the likelihood distribution is of the form

$$\pi_{b|\alpha}(b | \alpha) \propto \exp \left( -\frac{1}{2} \|\mathbf{A}\mathbf{W}\alpha - b\|_2^2 \right).$$

The conditional prior and the hyperprior are coupled to the posterior distribution via Bayes' formula, yielding the following expression for the posterior distribution

$$\pi_{(\alpha, \theta)|b}(\alpha, \theta | b) \propto \pi_{b|\alpha}(b | \alpha) \pi_{(\alpha, \theta)}(\alpha | \theta) \pi_\theta(\theta).$$

In the Bayesian framework, the posterior distribution is the complete solution to the inverse problem, that can be used to produce representative estimates of the unknown of interest, and quantify the uncertainty. Here, we chose to summarize the posterior with the Maximum A Posteriori (MAP) estimate,

$$(\alpha^*, \theta^*) \in \arg \max_{\alpha, \theta} \{ \pi_{(\alpha, \theta)|b}(\alpha, \theta | b) \}$$

or equivalently, by taking the negative logarithm of the density and ignoring the additive constants,

$$(4) \quad (\alpha^*, \theta^*) \in \arg \min_{\alpha, \theta} \{ \mathcal{F}(\alpha, \theta) \},$$

where

$$(5) \quad \mathcal{F}(\alpha, \theta) = \mathcal{F}(\alpha, \theta | r, \vartheta, \beta) \\ = \underbrace{\frac{1}{2} \|b - \mathbf{A}\mathbf{W}\alpha\|^2 + \frac{1}{2} \sum_{j=1}^N \frac{\alpha_j^2}{\theta_j} - \eta \sum_{j=1}^N \log \frac{\theta_j}{\vartheta_j} + \sum_{j=1}^N \left( \frac{\theta_j}{\vartheta_j} \right)^r}_{\mathcal{P}(\alpha, \theta | r, \beta, \vartheta)}, \quad \eta = \left( r\beta - \frac{3}{2} \right).$$

Echoing the terminology of classical regularization schemes, we refer to  $\mathcal{P}(\alpha, \theta | r, \beta, \vartheta)$  as the *penalty term*.

**3. The IAS algorithm.** The search for the minimizer of the MAP objective function in (5) is carried out with the *global hybrid* scheme introduced in [8], based on the iterative alternating sequential (IAS) algorithm described below. Details of the hybrid scheme that ensues are reviewed in Section 5.

Given a suitable initialization of the variances  $\theta^0$ , at each iteration step the IAS algorithm updates the iterates  $\alpha^t, \theta^t$  by solving the minimization problem in alternating directions, that is

$$\alpha^{t+1} \in \arg \min_{\alpha} \{ \mathcal{F}(\alpha, \theta^t) \}, \quad \theta^{t+1} \in \arg \min_{\theta} \{ \mathcal{F}(\alpha^{t+1}, \theta) \}.$$

Because of the particular form of the objective function, both variables can be updated efficiently as follows.

**Update of  $\alpha$ .** The  $\alpha$ -update reduces to solving a quadratic minimization problem, i.e.,

$$\alpha^{t+1} \in \arg \min_{\alpha} \left\{ \|b - A W \alpha\|_2^2 + \|D_{\theta}^{-1/2} \alpha\|^2 \right\}, \quad \theta = \theta^t,$$

or, equivalently, finding the solution in the least squares sense of the linear system

$$(6) \quad \begin{bmatrix} A W \\ D_{\theta}^{-1/2} \end{bmatrix} \alpha = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

After performing the change of variable

$$D_{\theta}^{-1/2} \alpha = \gamma,$$

we can write (6) as

$$\begin{bmatrix} A W D_{\theta}^{1/2} \\ I \end{bmatrix} \gamma = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

where  $I$  is an  $n \times n$  unit matrix. The solution of this least squares problem is also Tikhonov regularized solution of

$$A W D_{\theta}^{1/2} \gamma = b, \quad \alpha = D_{\theta}^{1/2} \gamma,$$

with regularization parameter equal to one. An alternative to Tikhonov regularization yielding a similar solution is to solve the underlying linear system with an iterative solver equipped with an early stopping criterion. The stopping condition is usually based on a variant of Morozov discrepancy principle, whereas the iterations terminate as soon as the discrepancy is of the order of the observation noise. In the statistical framework, under the Gaussian noise assumption, the noise level can be expressed in terms of the standard deviation of the noise. In our case, where we assume  $m$ -dimensional white noise, this quantity is equal to  $\sqrt{m}$ . Following [7, 9], we solve the linear system using the Conjugate Gradient for Least Squares (CGLS) algorithm with the early stopping at noise level  $\sqrt{m}$ ; see [7] for more details.

**Update of  $\theta$ .** Due to the mutual independence of the entries of  $\theta$ , each variance  $\theta_j$  can be updated separately by imposing the component-wise first order optimality condition on (5). More specifically,  $\theta_j^{t+1}$  is the solution of the non-linear equation

$$(7) \quad \frac{\partial \mathcal{F}}{\partial \theta_j} = -\frac{1}{2} \frac{\alpha_j^2}{\theta_j^2} - \left( r\beta - \frac{3}{2} \right) \frac{1}{\theta_j} + r \frac{\theta_j^{r-1}}{\vartheta_j^r} = 0, \quad \alpha = \alpha^{t+1}.$$

For some values of  $r$ , e.g.,  $r = \pm 1$ , (7) admits an analytic solution. However, in general we need to solve it numerically. It was shown in [9] that after the changes of variables  $\theta_j = \vartheta_j \xi_j$ ,  $\alpha_j = \sqrt{\vartheta_j} z_j$ , we may write  $\xi_j = \varphi(|z_j|)$ , and via implicit differentiation, the function  $\varphi$  satisfies the initial value problem

$$(8) \quad \varphi'(z) = \frac{2z\varphi(z)}{2r^2\varphi(z)^{r+1} + z^2}, \quad \varphi(0) = \left( \frac{\eta}{r} \right)^{1/r}.$$

Therefore the updated value of  $\theta_j$  can be computed by a numerical time integrator. Since the same type of differential equation is satisfied by all components, an efficient way to update  $\theta$  is to sort the current values  $z_j$  in an ascending order, and integrate sequentially over the gaps between the values by a suitable time integrator.

We point out that unlike in the formally similar alternating direction method for multipliers (ADMM) algorithm [2], that is often used to solve regularized inverse problems with sparsity promoting priors, the IAS algorithm does not require the

introduction of an artificial decoupling term of the fidelity and penalty terms, as the partial decoupling in IAS is automatic and exact.

From the point of view of statistical analysis, the proposed algorithm is not aiming at exploring the posterior density, and the MAP estimate might not be the best single point estimate to characterize the posterior. Other strategies of interest include the marginalization of the posterior density with respect to the hyperparameter  $\theta$ , or estimating an optimal  $\theta$  by first marginalizing  $\alpha$ . These alternative strategies have been discussed in literature, see, e.g., [21] for a recent and comprehensive contribution.

**4. Parameter selection strategies.** Before presenting the details of the hybrid scheme used in the numerical tests, we briefly review some of the main results related to the selection of the hyperparameters  $(r, \beta, \vartheta)$  appearing in the expression of the hyperprior in (3).

We start recalling a theorem, whose proof can be found in [9], summarizing how  $r$  and  $\beta$  affect the convexity properties of the functional  $\mathcal{F}$ .

**Theorem 4.1.** *Let  $\beta > 0$  and  $r \neq 0$ , and let  $\mathcal{F}(\alpha, \theta)$  be the objective function for the minimization problem in (4).*

- (a) *If  $r \geq 1$  and  $\eta = r\beta - 3/2 > 0$ , the function  $\mathcal{F}(\alpha, \theta)$  is globally convex.*
- (b) *If  $0 < r < 1$  and  $\eta = r\beta - 3/2 > 0$ , or, if  $r < 0$  and  $\beta > 0$ , the function  $\mathcal{F}(\alpha, \theta)$  is convex provided that*

$$(9) \quad \theta_j < \bar{\theta} = \vartheta_j \left( \frac{\eta}{r|r-1|} \right)^{1/r}.$$

The convexity of the MAP objective function, guaranteed for  $r \geq 1$ , is very convenient, however some of the configurations attained for  $r < 1$  can be very attractive in terms of sparsity promotion and rate of convergence. To better understand the connection between the parameter of the hyperprior and sparsity, consider the updating formula (8), expressing  $\theta_j$  as a function of  $\alpha_j$  as

$$\theta_j = g_j(\alpha_j) = \vartheta_j \varphi \left( \frac{|\alpha_j|}{\sqrt{\vartheta_j}} \right).$$

We review some recent results [4, 10, 9] about the connections that can be drawn between the generalized gamma hyperpriors and classical sparsity promoting penalty terms, assuming that  $(\theta_j, \alpha_j)$  satisfies the above identity.

- (i) For the gamma hypermodel, i.e.  $r = 1$ , as  $\eta = r\beta - \frac{3}{2} \rightarrow 0^+$  the penalty term approaches a weighted  $\ell_1$ -penalty term [4, 10],

$$\lim_{\eta \rightarrow 0^+} \mathcal{P} \left( \alpha, g(\alpha) \mid 1, \frac{3}{2} + \eta, \vartheta \right) = \sqrt{2} \sum_{j=1}^n \frac{|\alpha_j|}{\sqrt{\vartheta_j}}.$$

- (ii) If  $r\beta = \frac{3}{2}$ , the penalty term coincides with the weighted  $\ell_p$ -norm, with  $p = 2r/(r+1)$  [9],

$$\mathcal{P} \left( \alpha, g(\alpha) \mid r, \frac{3}{2r}, \vartheta \right) = C_r \sum_{j=1}^n \frac{|\alpha_j|^p}{\sqrt{\vartheta_j}^p}, \quad C_r = \frac{r+1}{(2r)^{r/(r+1)}}.$$

- (iii) For the inverse gamma hypermodel, corresponding to  $r = -1$ , the penalty term approaches the Student distribution, a prominently fat tailed distribution favoring large outliers, and leading to a greedy algorithm that strongly promotes sparsity [9].

To summarize, the above results indicate that the hyperpriors for which the global convexity of the corresponding hypermodel is not guaranteed ( $r < 1$ ) are expected to promote sparsity more effectively than the limit case  $r = 1$  that can be seen as a counterpart of the  $\ell_1$ -penalized case.

While the hyperparameters  $r$  and  $\beta$  determine the strength of the sparsity promotion and the convexity properties of the MAP objective function, the vector of the scale parameters  $\vartheta$  can be set automatically once the operator  $\mathbf{AW}$  is given. More specifically, for each  $j$ ,  $\vartheta_j$  can be related to the sensitivity of the data to  $x_j$ , given by the quantity  $\|\mathbf{AW}e_j\|^2$ , where  $e_j \in \mathbb{R}^N$  denotes the canonical  $j$ -th Cartesian unit vector. It was proven, for  $r = 1$  in [7, 10] and in more general settings in [9], that under the assumption that the signal-to-noise ratio is given, and that the prior satisfies an exchangeability condition guaranteeing that no particular sparse combinations of components of  $x$  are favored over others, the entries of  $\vartheta$  must be chosen as

$$(10) \quad \vartheta_j = \frac{C}{\|\mathbf{AW}e_j\|^2},$$

where  $C > 0$  is a constant encoding the expected sparsity on the solution and an estimate of the signal-to-noise-ratio. For details, we refer to the cited articles. We remark that, in general, sensitivity weights are introduced to compensate for the possible non-uniform design of the forward model operator  $\mathbf{A}$ . Sensitivity weights play an important role in, e.g., inverse source problems, in which sources near the observation points may be favored over far away sources unless the exchangeability condition is imposed. In the current setting, when the dictionary consists of sub-frames with possibly different column norms, we expect the different weights  $\vartheta_j$  to prevent the representation of the signal the frames with larger column norms to dominate.

**5. Local and global hybrid IAS.** In the following discussion, we write the penalty function  $\mathcal{P}(x, \theta \mid r, \beta, \vartheta)$  in terms of components,

$$\begin{aligned} \mathcal{P}(\alpha, \theta \mid r, \beta, \vartheta) &= \sum_{j=1}^N \left( \frac{1}{2} \frac{\alpha_j^2}{\theta_j} - \eta \log \frac{\theta_j}{\vartheta_j} + \left( \frac{\theta_j}{\vartheta_j} \right)^r \right) \\ &= \sum_{j=1}^N \mathcal{P}_j(\alpha_j, \theta_j \mid r, \beta, \vartheta_j). \end{aligned}$$

In [8], two different hybrid strategies to speed up and enhance sparsity promotion in the IAS algorithm were proposed. In both versions, the IAS iterations are initiated by selecting a conservative set of hyperparameters for which the objective function is convex, thus guaranteeing global convergence to a unique minimizer. We denote this set of parameters by  $(r^{(1)}, \beta^{(1)}, \vartheta^{(1)})$ . For the second phase of the hybrid algorithm, we select another set of parameters,  $(r^{(2)}, \beta^{(2)}, \vartheta^{(2)})$ , for which the global convexity of the objective function is not valid. To match the models so that they express coherent prior beliefs, we adjust the scale parameters  $\vartheta^{(j)}$  so as

to satisfy the compatibility condition

$$(11) \quad \left( \frac{\eta^{(1)}}{r^{(1)}} \right)^{1/r^{(1)}} \vartheta_j^{(1)} = \left( \frac{\eta^{(2)}}{r^{(2)}} \right)^{1/r^{(2)}} \vartheta_j^{(2)},$$

that guarantees that the parameter  $\theta_j$  computed at  $\alpha_j = 0$  returns the same value regardless of the model. For further discussion, we refer to [8].

In the local hybrid version, the IAS algorithm is initially run with hyperparameters  $(r^{(1)}, \beta^{(1)}, \vartheta^{(1)})$ , and after each iteration step, we check which  $\theta_j$ , if any, satisfies the condition (9), where  $\bar{\theta}$  is computed using the hyperparameter set  $(r^{(2)}, \beta^{(2)}, \vartheta^{(2)})$ . In correspondence of those which do, we modify the local objective function so that

$$\mathcal{P}_j(\alpha_j, \theta_j \mid r^{(1)}, \beta^{(1)}, \vartheta_j^{(1)}) \rightarrow \mathcal{P}_j(\alpha_j, \theta_j \mid r^{(2)}, \beta^{(2)}, \vartheta_j^{(2)}).$$

The global hybrid scheme is based on the idea that after a number of IAS iteration rounds, the iterate of the globally convex objective function with hyperparameters  $(r^{(1)}, \beta^{(1)}, \vartheta^{(1)})$  is near the unique global minimum of that objective function. Restarting the IAS from the current point with the parameters  $(r^{(2)}, \beta^{(2)}, \vartheta^{(2)})$  may quickly find a local minimizer near the global minimizer of the original objective function. While the two minimizers are likely not far apart, the local minimizer is typically sparser, and the convergence to it is faster.

**6. Computed examples.** In this section, we demonstrate the viability of the hybrid IAS algorithm in the context of overcomplete representations. More specifically, we restrict ourselves to the global hybrid strategy, switching from the first to the second hyperprior after 10 iterations, if not differently specified. The main goal of the following examples is to demonstrate that the global hybrid IAS is capable of selecting from a dictionary of sub-frames, where several representations are admissible, a set of atoms that make the representation as sparse as possible.

**Signal restoration from convolution data.** The first test case is a one-dimensional deconvolution problem. The generative model is a piecewise constant signal  $f : [0, 1] \rightarrow \mathbb{R}$ ,  $f(0) = 0$ , and the data consist of a few discrete observations,

$$b_j = \int_0^1 A(s_j - t) f(t) dt + \varepsilon_j, \quad 1 \leq j \leq m, \quad A(s_j, t) = \frac{1}{\sqrt{2\pi w^2}} e^{-\frac{(s_j - t)^2}{2w^2}},$$

corrupted by Gaussian blur with  $w = 0.02$  and additive scaled white Gaussian noise, with standard deviation  $\sigma$  set to 2% of the maximum of the noiseless signal. The data has been generated using a discretization of the unit interval with  $n = n_{\text{dense}} = 1253$  nodes, while in the forward model used for solving the inverse problem, we set  $n = 500$ . The number of equidistant observation points in the signal domain is  $m = 46$ . The generative signal and the data are shown in Figure 1.

The generative signal admits a natural sparse representation in terms of its increments  $z_j = z_j - z_{j-1}$  over the interval of definition. Assuming  $x_0 = 0$ , then

$$(12) \quad z = Bx, \quad B = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n},$$



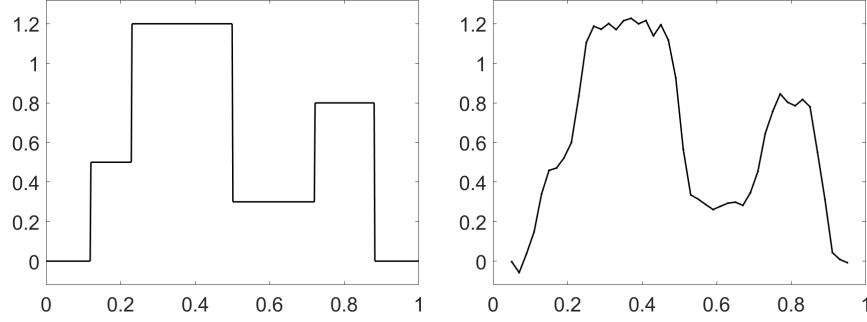


FIGURE 1. The generative model (left) and the blurred and noisy data vector  $b \in \mathbb{R}^{46}$  (right).

hence

$$x = Lz \quad \text{with} \quad L = B^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & & \ddots & \\ 1 & \dots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Our goal is to test the effectiveness of the outlined framework in recovering the most natural sparse representation of the given signal. Let  $C$  denote the discrete cosine transform matrix, providing an alternative and accurate way of representing the signal,

$$x = C^T y, \quad y = Cx,$$

which is, however, not sparse. To test whether the algorithm is able to identify the frame that allows a sparse representation, we consider the overcomplete dictionary,

$$W = [W_1, W_2] \in \mathbb{R}^{n \times 2n} \quad \text{with} \quad W_1 = L \in \mathbb{R}^{n \times n} \text{ and } W_2 = C^T \in \mathbb{R}^{n \times n},$$

and formulate the underlying linear inverse problem as

$$b = AW\alpha + \varepsilon = A[W_1, W_2] \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_m),$$

where  $A$  is the discrete blur operator.

In this example, the global hybrid IAS is run with parameters  $(r^{(1)}, \eta^{(1)}) = (1, 10^{-4})$ ,  $(r^{(2)}, \eta^{(2)}) = (1/2, 10^{-3})$ ; we recall that the sensitivity weights  $\vartheta^{(1)}$  are set automatically according to (10), while the vector  $\vartheta^{(2)}$  is fixed so that condition (11) is satisfied.

The signal reconstructed by the global hybrid IAS scheme is shown in Figure 2. The restored  $\alpha_1$  and  $\alpha_2$  and their contribution in the estimated signal are shown in Figure 3, together with the scaled variances corresponding to  $\alpha_1$  and  $\alpha_2$ , i.e.

$$\frac{\theta_j}{\vartheta_j^{(2)}}, \quad 1 \leq j \leq n \quad \text{and} \quad \frac{\theta_j}{\vartheta_j^{(2)}}, \quad n+1 \leq j \leq 2n.$$

Notice that the output variances are scaled by the sensitivities corresponding to the second hyperprior used to design the hybrid scheme.

Despite the relatively high level of degradation (blur and noise) and down-sampling in the observed data  $b$ , the algorithm has no problem detecting the basis that provides a more natural and sparse representation for the original signal. In

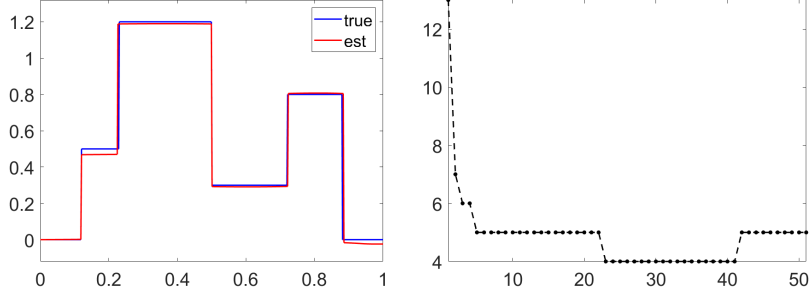


FIGURE 2. Reconstruction of the signal  $x$  (left) and the count of CGLS steps per outer iteration of the global hybrid IAS (right).

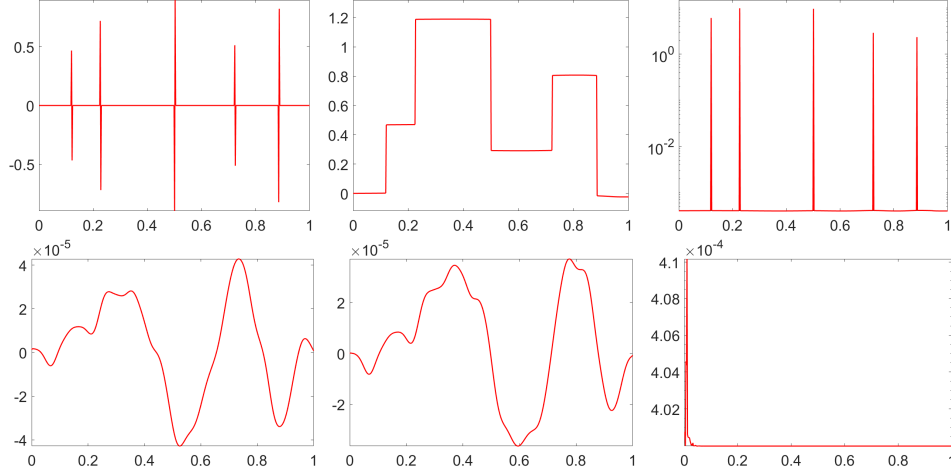


FIGURE 3. Vector  $\alpha_i$  (left panels), corresponding scaled variances (middle panels) and contribution of the signal  $W_i \alpha_i$  (right panels) for  $i = 1$ , i.e. representation in terms of increments, (top row) and  $i = 2$ , i.e. representation in terms of cosine transform, (bottom row).

fact, the coefficients  $\alpha_2$  are five to six orders of magnitude smaller than the non-vanishing components of  $\alpha_1$ . The degree of sparsity in the final representation is also reflected in the number of CGLS steps per outer iteration of the global hybrid IAS - see Figure 2 - which quickly settles around the cardinality of the support of  $\alpha$ .

**Image denoising on a synthetic image.** In the second example, we consider the problem of denoising a blocky gray scale test image  $x \in \mathbb{R}^{n \times n}$ ,  $n = 200$ . The pixel values, which are between 0 and 1, are corrupted by scaled white Gaussian noise with standard deviation  $\sigma$  set to 10% of the maximum of the noiseless image, i.e.  $\sigma = 0.1$  - see Figure 4.

The test image presents sharp edges lying along the horizontal and vertical axes. Therefore,  $x$  admits a sparse representation both in the vertical and horizontal increment bases, the latter being slightly less sparse than the former. After representing the image in vector form  $x \in \mathbb{R}^{n^2}$  by stacking the pixel values columnwise,

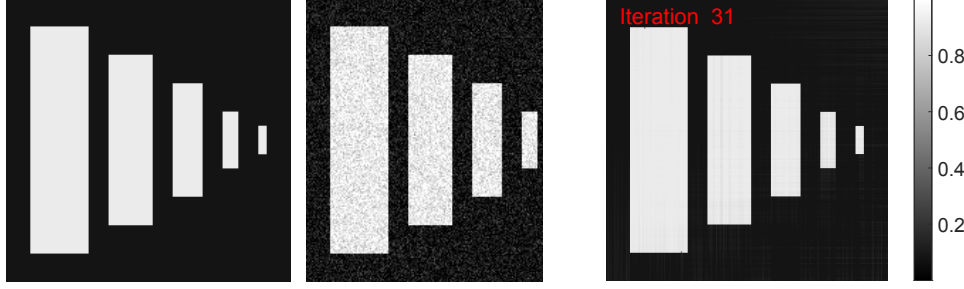


FIGURE 4. Original image (left), observed data (middle) and re-constructed image (right).

5

we introduce the redundant dictionary  $W = [W_1, W_2] \in \mathbb{R}^{n^2 \times 2n^2}$  with

$$(13) \quad W_1 = (I_n \otimes B)^{-1} \in \mathbb{R}^{n^2 \times n^2} \text{ and } W_2 = (B \otimes I_n)^{-1} \in \mathbb{R}^{n^2 \times n^2},$$

where  $B$  is defined as in (12), and  $\otimes$  stands for the Kronecker product. Homogeneous Dirichlet boundary conditions are assumed on the left and top edges of the image. We want to estimate the sparse vector  $\alpha = [\alpha_1, \alpha_2]^T$ , with  $\alpha_i \in \mathbb{R}^{n^2}$ ,  $i = 1, 2$ , from the data vector  $b \in \mathbb{R}^{n^2}$ , given the forward model

$$b = W\alpha + \varepsilon = [W_1, W_2] \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n^2}).$$

It is worth remarking here that we require  $\alpha$  to be not only sparse, but as sparse as possible.

The hyperparameters of the global hybrid IAS are set as  $(r^{(1)}, \eta^{(1)}) = (1, 10^{-3})$  and  $(r^{(1)}, \eta^{(1)}) = (1/2, 10^{-2})$ , while, as before,  $\vartheta^{(1)}, \vartheta^{(2)}$  are automatically fixed according to (10) and (11), respectively.

The restored image is shown in Figure 4, while the contribution of the vertical and horizontal increment bases together with the output scaled variances corresponding to vectors  $\alpha_1$  and  $\alpha_2$  are shown in Figure 5. We observe that the image is almost completely restored in terms of the basis vectors corresponding to increments in the vertical direction ( $\alpha_1$ ), whereas the entries of  $\alpha_2$ , corresponding to increments in the horizontal direction is negligible. The representation in terms of  $W_1$  is indeed sparser than that in terms of  $W_2$ , due to the shorter horizontal boundary of the white inclusion compared to the vertical boundary.

**Image denoising on a natural image.** We demonstrate the scalability of our approach with large-scale denoising problem. Consider the  $512 \times 512$  gray-scale natural image with pixel values between 0 and 1 shown in the top left panel of Figure 6. The observed data  $b$  is a version corrupted by added white Gaussian noise with standard deviation  $\sigma$  set to 5% of the maximum of the noiseless image, i.e.  $\sigma = 0.05$ , shown in the top middle panel of Figure 6.

We consider an overcomplete basis  $W = [W_1, W_2] \in \mathbb{R}^{n^2 \times 2n^2}$ , with  $W_1, W_2 \in \mathbb{R}^{n^2 \times n^2}$ ,  $n = 512$ , defined in (13); the vector describing the 2D signal in the selected dictionary can thus be written as  $\alpha = [\alpha_1, \alpha_2]^T$ , with  $\alpha_1, \alpha_2 \in \mathbb{R}^{n^2}$  representing the vertical and horizontal increments, respectively. Note that since this test image is not piecewise constant, but rather a mixture of jumps, smooth, and textured parts, we do not expect the vertical and horizontal increments to be *naturally* sparse. In

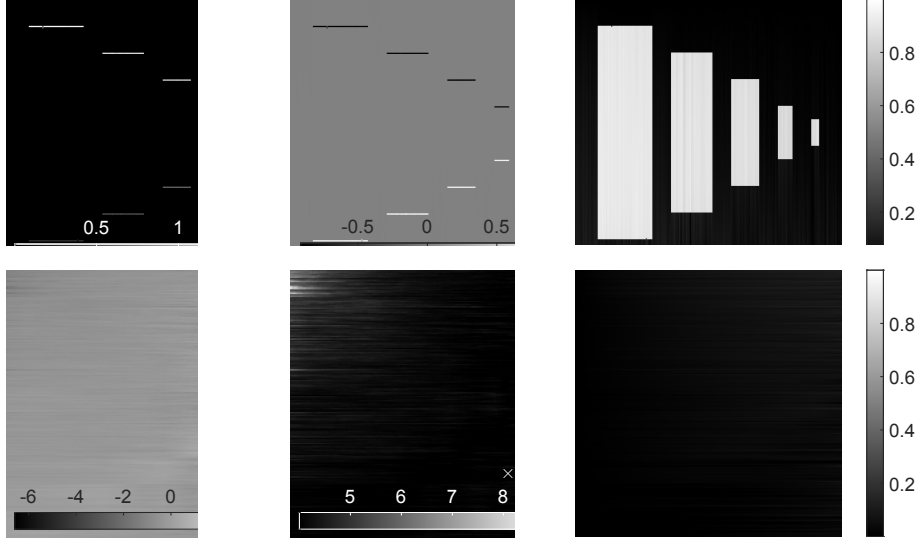


FIGURE 5. Vector  $\alpha_i$  (left panels), base-10 logarithmic plot of the corresponding scaled variances (middle panels) and vector  $W_i \alpha_i$  contributing to the final restoration (right panels) for  $i = 1$ , i.e. vertical increments representation, (top row),  $i = 2$ , i.e. horizontal increments representation, (bottom row).

fact, when representing the original image in the chosen dictionary, about 47% of the coefficients  $\alpha_1, \alpha_2$  in both horizontal and vertical directions are non-zero.

We run the global hybrid IAS with  $(r^{(1)}, \eta^{(1)}) = (1, 10^{-4})$ ,  $(r^{(2)}, \eta^{(2)}) = (1/2, 10^{-3})$ , and letting  $\vartheta^{(1)}, \vartheta^{(2)}$  be set automatically.

The restored image is shown in the top right panel of Figure 6, together with a close-up to facilitate the visual comparison. In Figure 7, we also show the output representation vectors  $\alpha_1, \alpha_2$ , with the corresponding scaled output variances and their contribution in the final restoration.

Besides the quality of the restored image, we are interested in highlighting the compressing capability of our approach. Consider first the coefficients  $\alpha = [\alpha_1, \alpha_2]^T$  of the original image, and choose a threshold value  $\beta > 0$  to be the smallest non-zero coefficient,

$$\beta = \min\{|\alpha_j| \mid |\alpha_j| > 0\}.$$

After computing the restored coefficients using the sparsity promoting hybrid IAS, setting to zero those whose absolute value is below a threshold value, we find that only 3.92% of those for the vertical components, and 4.07% of those for the horizontal components are nonvanishing. Thus, the representation of the image in this basis was compressed by a factor more than ten from the original image without a significant deterioration in the image quality. Figure 8 shows the value distributions of the original and restored coefficients, plotted as histograms in logarithmic scale: clearly the coefficient values are significantly compressed towards zero in both directions. It is worth remarking here that the Bayesian target in sparsity promoting problems has to be understood as compressibility, as the entries of  $\alpha$  cannot vanish by construction of the algorithm, but they can be made arbitrary small by suitable parameter selection.

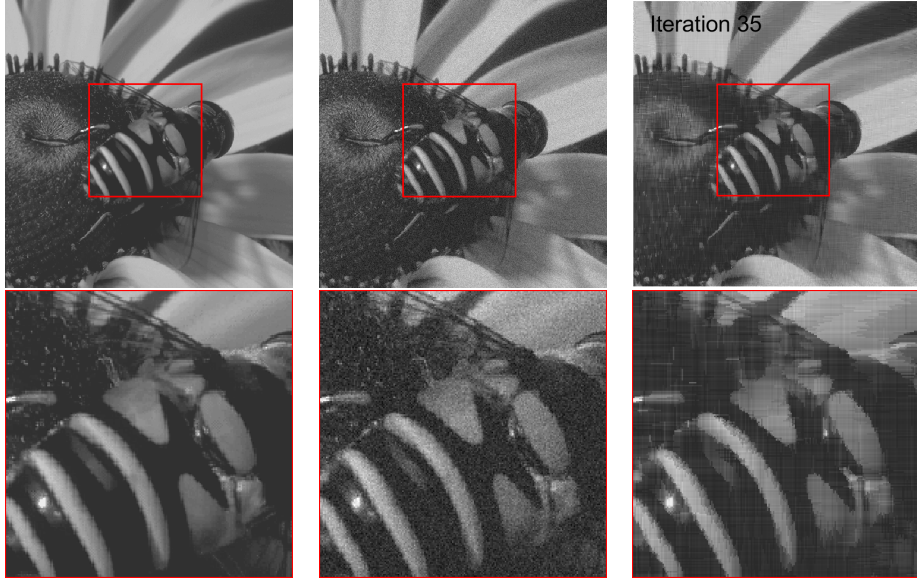


FIGURE 6. *Top row*: original  $512 \times 512$  test image (left), observed data (middle) and denoised image (right). *Bottom row*: respective close-up(s).

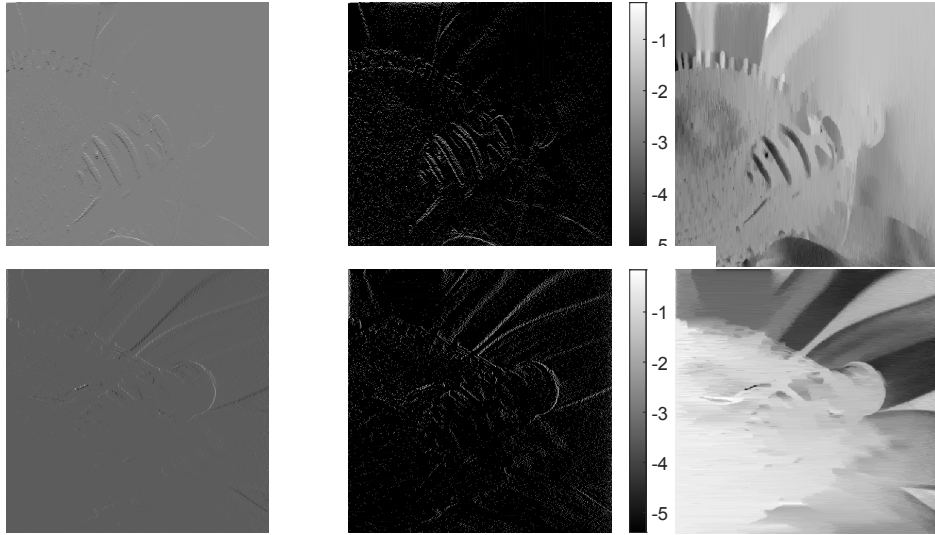


FIGURE 7. Vector  $\alpha_i$  (left panels), base-10 logarithmic plot of the corresponding scaled variances (middle panels), and vector  $W_i \alpha_i$  (right panels) for  $i = 1$ , i.e. vertical increments representation, (top row) and  $i = 2$ , horizontal increments representation, (bottom row).

**Image restoration.** In the fourth example, we consider the restoration problem of the  $n \times n$  generative image in Figure 9, with  $n = 100$ , with values in  $[0, 1]$ .

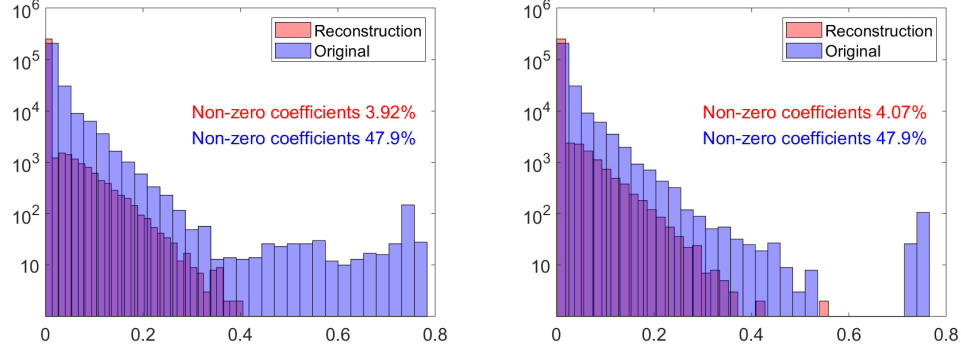


FIGURE 8. Value distributions of the coefficients corresponding to vertical (left) and horizontal (right) increments in logarithmic scale. The blue distributions correspond to the original image which does not allow sparse representation in the basis, which is reflected in the high percentage of non-vanishing coefficients, while the red distribution is the sparse denoised reconstruction, in which the percentage of coefficients above the negligible threshold value is reduced by an order of magnitude.

The image has been corrupted by Gaussian blur of width  $w = 0.006$  and additive scaled white Gaussian noise with standard deviation  $\sigma$  set to 1% of the maximum of the noiseless signal, i.e.  $\sigma = 0.01$  - see Figure 9. The test image presents three distinctive features, namely point-wise stars, the blocky moon and the smooth cloud. After re-arranging the original  $x$  in a vectorized form by stacking its entries in columnwise order, we hypothesize that a suitable dictionary for the problem of interest is  $W = [W_1, W_2, W_3, W_4] \in \mathbb{R}^{n^2 \times 4n^2}$ , with

$$W_1 = I_{n^2}, W_2 = (I_n \otimes B)^{-1} \in \mathbb{R}^{n^2 \times n^2}, W_3 = (B \otimes I_n)^{-1} \in \mathbb{R}^{n^2 \times n^2} \text{ and } W_4 = C^T,$$

where  $B$  is as in (12) and  $C$  is the 2D cosine transform matrix. The problem is to estimate the sparse vector  $\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]^T$ , with  $\alpha_i \in \mathbb{R}^{n^2}$ , for  $i = 1, 2, 3, 4$ , from the data vector  $b \in \mathbb{R}^{n^2}$ , given the forward model

$$b = AW\alpha + \varepsilon = [W_1, W_2, W_3, W_4] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n^2}),$$

with  $A$  representing the discrete blur operator.

The global hybrid IAS is run with hyperparameters  $(r^{(1)}, \eta^{(1)}) = (1, 10^{-4})$ ,  $(r^{(2)}, \eta^{(2)}) = (1/2, 10^{-4})$  and  $\vartheta^{(1)}, \vartheta^{(2)}$  automatically fixed as in the previous examples.

The image restored via the global hybrid IAS algorithm is shown in Figure 9, while Figure 10 shows the reconstructions of the representation vectors  $\alpha_i$ , the corresponding variances scaled by the sensitivities, and the contribution of the vectors  $W_i \alpha_i$  in the final restoration, for  $i = 1, 2, 3, 4$ .

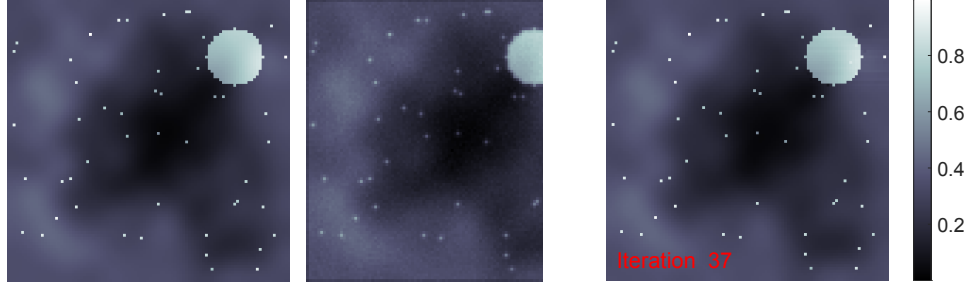


FIGURE 9. Original image (left), observed data (middle) and re-constructed image (right).

We point out that, as in the previous example, the representation vectors in both the vertical and horizontal increment bases are sparse. Nonetheless, the hybrid hypermodel selects the one with fewer non-zero entries.

**Dictionary learning.** The final example, coming from machine learning, is concerned with the sparse identification of hand-written digits based on a dictionary of annotated data. Consider the MNIST data set of hand-written digits  $0, 1, \dots, 9$  digitized as  $16 \times 16$  black-and-white images. Denoting by  $w^{(j)} \in \mathbb{R}^{(16)^2}$ ,  $1 \leq j \leq N$  the vectorized image vectors of  $N = 1707$  handwritten digits constituting the atoms of the dictionary, and by  $c_j \in \{0, 1, \dots, 9\}$  the corresponding annotations, we form the dictionary matrix

$$W = [w^{(1)} \quad \dots \quad w^{(N)}] \in \mathbb{R}^{n \times N}, \quad n = 256, N = 1705.$$

To identify an handwritten digit  $b$  drawn from an independent set of handwritten digits, we seek to represent it in a sparse manner in terms of the given dictionary,

$$b = W\alpha + \varepsilon,$$

where  $\alpha \in \mathbb{R}^N$  is a sparse vector, and  $\varepsilon$  represents the discrepancy between the data and its representation. The idea is represented schematically in Figure 11. We point out that in the dictionary consisting of all handwritten digits, the digits with same annotation can each be thought of representing a sub-dictionary, and as the proposed algorithm seeks the most economic representation, it is natural that the representation corresponds to picking the representing atoms from the sub-dictionary with greatest affinity with the digit that represents the data.

In this example, we run the global hybrid IAS algorithm using the parameters  $(r^{(1)}, \beta^{(1)}, \vartheta^{(1)}) = (1, 3/2 + 10^{-4}, 10^{-5})$ , where all components of the vector  $\vartheta^{(1)}$  are assumed equal, as sensitivity is not an issue in this example, and  $(r^{(2)}, \beta^{(2)}) = (-1, 1)$ , with the value  $\vartheta^{(2)}$  determined from the compatibility condition (11). Furthermore, since the digit images are non-negative, after each update step of the pair  $(\alpha, \theta)$ , we project the image to the positive cone. A theoretical justification of the projection step was given in [9]. We switch from the first to the second model in the hybrid IAS scheme when either the relative change in  $\theta$  with respect to the  $\ell_2$ -norm falls below  $10^{-3}$  or 80 iterations have been completed.

Figures 12, 13 and 14 show the results with different choices of the standard deviation of the likelihood. Observe that here, the noise term  $\varepsilon$  represents the discrepancy between the data  $b$  and its representation in terms of the dictionary, and can be chosen according to how much fidelity is required. Choosing  $\sigma$  large

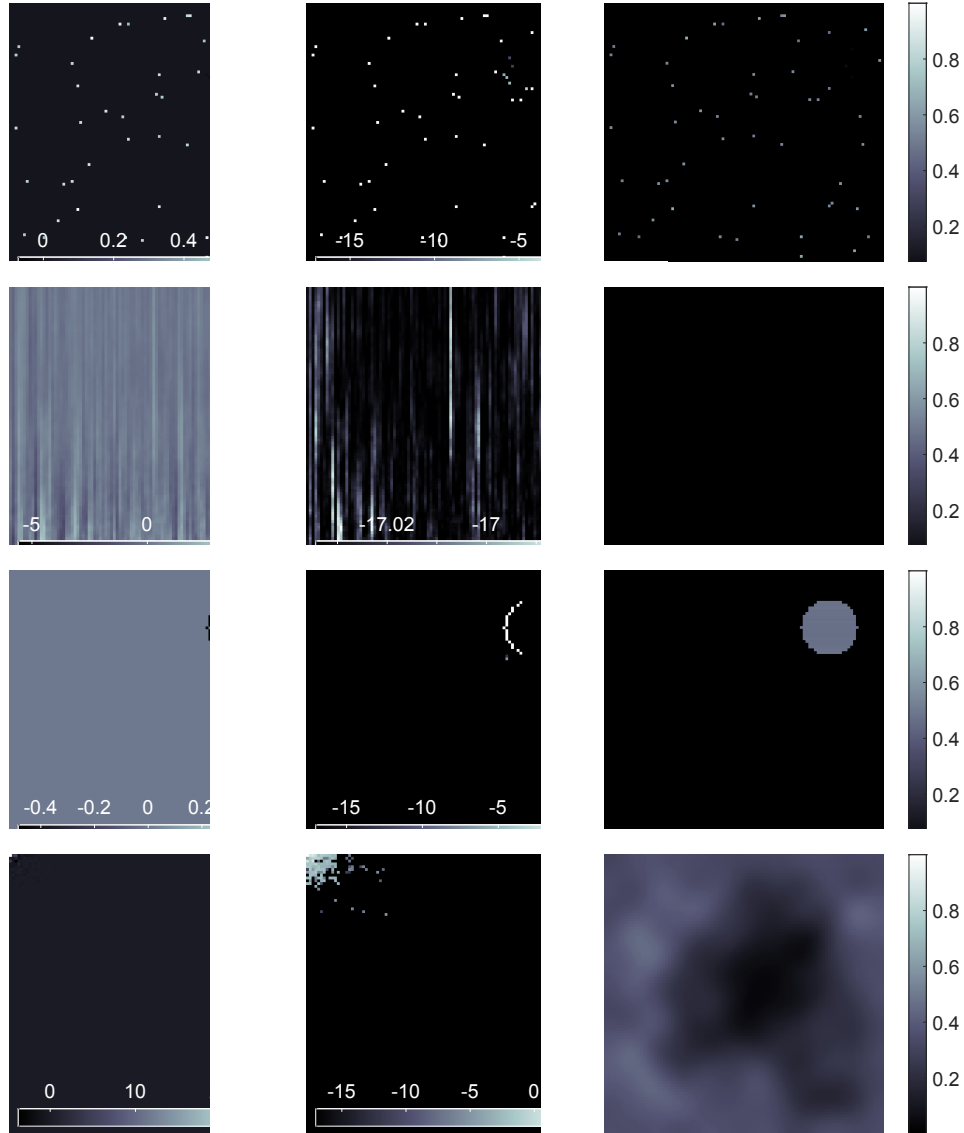


FIGURE 10. Representation vector  $\alpha_i$  (left panels), base-10 logarithmic plot of the corresponding scaled variances (middle panels) and vector  $W_i\alpha_i$  contributing to the final restoration (right panels) for  $i = 1$  (first row),  $i = 2$  (second row),  $i = 3$  (third row) and  $i = 4$  (fourth row).

allows a very sparse representation, as the required quality of the approximation is low, however, poor approximation easily leads to a mis-labeling of the digit. On the other hand, decreasing  $\sigma$  forces the approximation to be better, and more atoms are required. The labeling can be done using the majority vote principle. In the computed example, Figures 12 and 13, the labeling with majority vote is correct in each case, while in Figure 14 with sparser representation, mislabelings occur.



$$\begin{array}{|c|} \hline 3 \\ \hline \end{array} \approx \alpha_1 \times \begin{array}{|c|} \hline 7 \\ \hline \end{array} + \alpha_2 \times \begin{array}{|c|} \hline 3 \\ \hline \end{array} + \cdots + \alpha_N \times \begin{array}{|c|} \hline 4 \\ \hline \end{array}$$

FIGURE 11. A schematic representation of the dictionary learning example. The digit on the right is the non-annotated image  $b$ , which is approximated in terms of the annotated atoms  $w_i$  on the right. The coefficients  $\alpha_i$  can then be used to identify the digit.

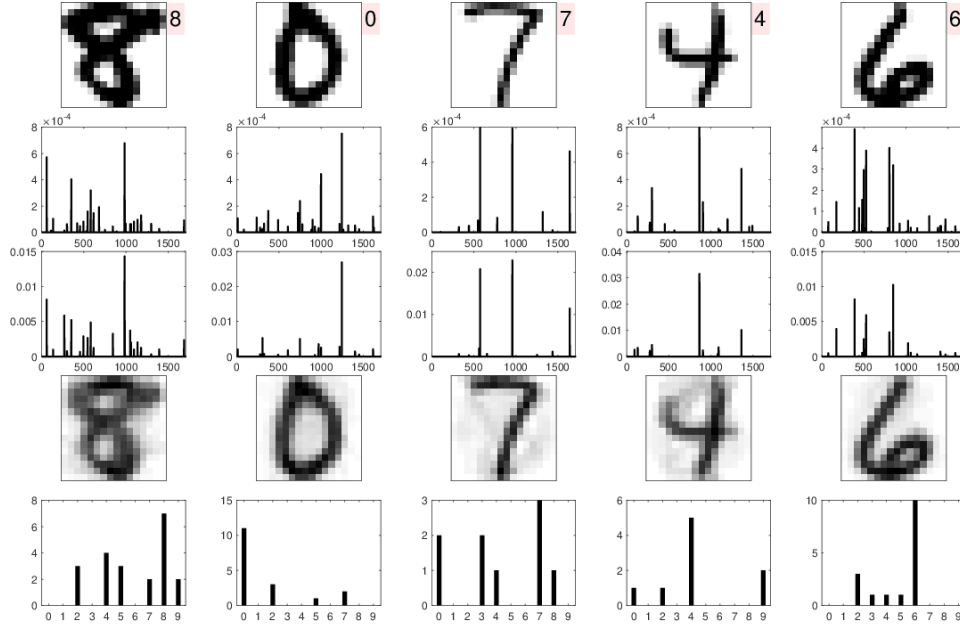


FIGURE 12. Dictionary learning results. The first row shows the test images of the digits to be classified by the dictionary learning algorithm (vector  $b$ ), the true annotation indicated in the figure, the second row the vectors  $\theta$  after the IAS iteration with the first hyperprior, and the third row after the iteration with the second hyperprior. The fourth row represents the synthesis  $W\alpha$  approximating the original digit, and finally, the fifth row gives the histogram of the annotations of the atoms corresponding to coefficients above a threshold  $\tau = 0.01$ . The annotation is done by majority vote, choosing the largest of the bins. In this example, the standard deviation of the noise representing the mismatch was  $\sigma = 0.01$ .

**7. Conclusions and future work.** The hierarchical Bayesian framework combined with Krylov subspace iterative solvers for large linear systems is well suited for the design of computationally efficient methods to solve large scale ill posed inverse problems with sparsity constraints. From the point of view of computational efficiency, replacing a whitened Tikhonov-type penalty by a Krylov subspace iteration equipped with early stopping is of crucial importance. The estimate found by

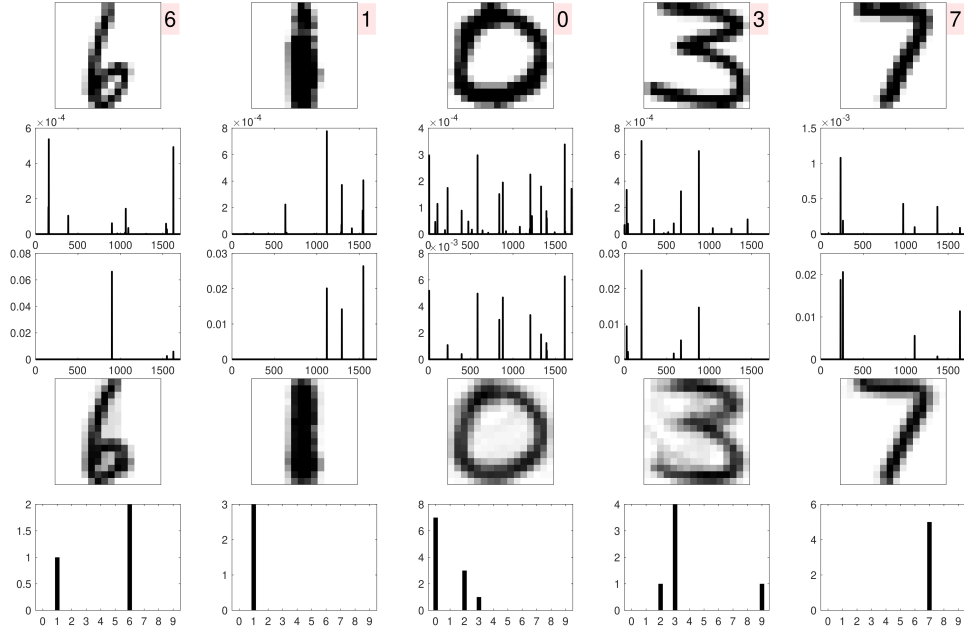


FIGURE 13. The rows are as in Figure 12. In this example, the standard deviation of the noise representing the mismatch was  $\sigma = 0.05$ . Observe that the approximation becomes sparser.

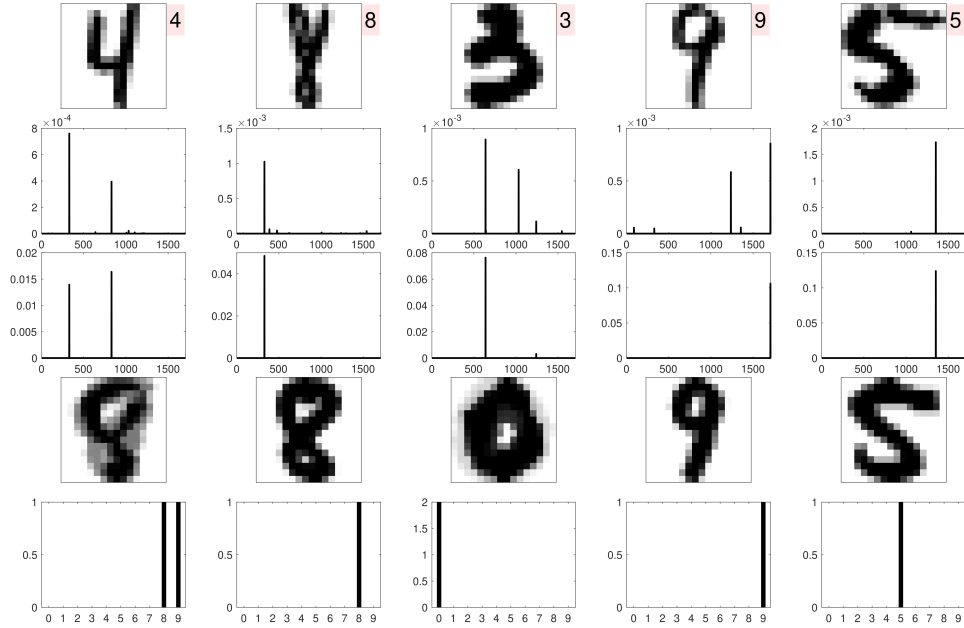


FIGURE 14. The rows are as in Figure 12. In this example, the standard deviation of the noise representing the mismatch was  $\sigma = 0.1$ . The increased sparsity here is traded with an increased number of misclassifications, such as in the first and the third columns.

this process is not guaranteed to coincide with the MAP estimate, and it was referred to as quasi-MAP estimate in [5]. The statistical analysis of the early stopping regularization strategy based on Krylov subspace methods is not straightforward, since the estimate depends non-linearly on the data, and it can be seen as an approximate Bayesian computing (ABC) strategy. For further discussion, see, e.g. [7]. Here we have shown that the framework can be naturally adapted for dealing with overcomplete systems, consisting of, e.g., combined frames or bases. The approach has significant potential when it may not be known a priori which frame is best suited for representing the unknown, leaving it up to the algorithm to find the most parsimonious representation. In order to avoid that one frame is favored over another, however, it is important that the data are equally sensitive to components in every frame. Fortunately, the sensitivity analysis developed by the authors in [6, 10, 9], provides naturally such scaling. The proposed sensitivity weights are rooted in the very natural Bayesian principle of exchangeability, stating that no set of non-zero components with a given cardinality should be favored over any other. In light of this principle, the scaling guarantees the same explanation power for every sub-frame, so the one leading to most sparse solution is automatically selected. This feature may turn out particularly useful in machine learning, with applications such as MRI fingerprinting (see, e.g., [16]). In [10], a connection between the proposed IAS algorithm and the compressed sensing literature [11] was considered, suggesting that when the forward model guarantees perfect sparse recovery, the IAS algorithm effectively finds a good approximation of it. It is reasonable to believe that the results can be extended to overcomplete dictionaries, for which similar recovery results are known [12].

The methodology developed in this paper has been tested only with Gaussian noise, leading to a quadratic fidelity term in the optimization problem. The IAS framework has been shown to work well with other noise models, e.g., Poisson distributed noise in connection with low dose X-ray tomography and PET, see [1]. The applicability of the approach to non Gaussian noise may be very important for its use for dictionary learning problems where the data consist, e.g., of word counts. The extension of the method to large scale problems, different noise models and nonlinear forward models is the next step and will be addressed in separate future contributions.

## REFERENCES

- [1] J. M. Bardsley, D. Calvetti and E. Somersalo [Hierarchical regularization for edge-preserving reconstruction of PET images](#), *Inverse Problems*, **26** (2010), 035010.
- [2] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, [Distributed optimization and statistical learning via the alternating direction method of multipliers](#), *Foundations and Trends in Machine Learning*, **3** (2011), 1–122.
- [3] A. M. Bruckstein, D. L. Donoho and M. Elad, [From sparse solutions of systems of equations to sparse modeling of signals and images](#), *SIAM Review*, **51** (2009), 34–81.
- [4] D. Calvetti, H. Hakula, S. Pursiainen and E. Somersalo, [Conditionally Gaussian hypermodels for cerebral source localization](#), *SIAM Journal on Imaging Sciences*, **2** (2009), 879–909.
- [5] D. Calvetti, F. Pitolli, J. Prezioso, E. Somersalo and B. Vantaggi, [Priorconditioned CGLS-based quasi-MAP estimate, statistical stopping rule, and ranking of priors](#), *SIAM Journal of Scientific Computing*, **39** (2017), S477–S500.
- [6] D. Calvetti, A. Pascarella, F. Pitolli, E. Somersalo and B. Vantaggi, [Brain activity mapping from MEG data via a hierarchical Bayesian algorithm with automatic depth weighting](#), *Brain Topography*, **32** (2019), 363–393.
- [7] D. Calvetti, F. Pitolli, E. Somersalo and B. Vantaggi, [Bayes meets Krylov: Statistically inspired preconditioners for CGLS](#), *SIAM Review*, **60** (2018), 429–461.

- [8] D. Calvetti, M. Pragliola and E. Somersalo, [Sparsity promoting hybrid solvers for hierarchical Bayesian inverse problems](#), *SIAM Journal on Scientific Computing* **42** (2020), A3761–A3784.
- [9] D. Calvetti, M. Pragliola, E. Somersalo and A. Strang, [Sparse reconstructions from few noisy data: Analysis of hierarchical Bayesian models with generalized gamma hyperpriors](#), *Inverse Problems*, **36** (2020), 025010.
- [10] D. Calvetti, E. Somersalo and A. Strang, [Hierarchical Bayesian models and sparsity:  \$\ell\_2\$ -magic](#), *Inverse Problems*, **35** (2019), 035003.
- [11] E. J. Candes, J. Romberg and T. Tao, [Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information](#), *IEEE Transactions on Information Theory*, **52** (2006), 489–509.
- [12] E. J. Candes, Y. C. Eldar, D. Needell and P. Randall, [Compressed sensing with coherent and redundant dictionaries](#), *Applied and Computational Harmonic Analysis*, **31** (2011), 59–73.
- [13] A. Chambolle, M. Holler and T. Pock, [A convex variational model for learning convolutional image atoms from incomplete data](#), *Journal of Mathematical Imaging and Vision*, **62** (2020), 417–444.
- [14] G. Chen and D. Needell, Compressed sensing and dictionary learning, *Finite Frame Theory, Proceedings of Symposia in Applied Mathematics*, Vol. 73, Providence, RI, (2016), 201–241.
- [15] S. S. Chen, D. L. Donoho, and M. A. Saunders, [Atomic decomposition by basis pursuit](#), *SIAM Journal on Scientific Computing*, **20**(1) (1998), 33–61.
- [16] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J. L. Sunshine, J. L. Duerk and M. A. Griswold, [Magnetic resonance fingerprinting](#), *Nature*, **495** (2013), 187–192.
- [17] S. G. Mallat and Z. Zhang, [Matching pursuits with time-frequency dictionaries](#), *IEEE Transactions on Signal Processing*, **41** (1993), 3397–3415.
- [18] R. Rubinstein, A. M. Bruckstein and M. Elad, [Dictionaries for sparse representation modeling](#), *Proceedings of the IEEE*, **98** (2010), 1045–1057.
- [19] J. Starck, J. Fadili and F. J. Murtagh, [The undecimated wavelet decomposition and its reconstruction](#), *IEEE Transactions on Image Processing*, **16** (2007), 297–309.
- [20] J. L. Starck, M. Elad and D. Donoho, [Redundant multiscale transforms and their application for morphological component separation](#), *Advances in Imaging and Electron Physics*, **132** (2004), 287–348.
- [21] A. F. Vidal, V. De Bortoli, M. Pereyra and A. Durmus, [Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical Bayesian approach Part I: Methodology and experiments](#), *SIAM Journal on Imaging Sciences*, **13** (2020), 1945–1989.

Received September 2020; revised April 2021.

E-mail address: [monica.pragliola2@unibo.it](mailto:monica.pragliola2@unibo.it)

E-mail address: [dxc57@case.edu](mailto:dxc57@case.edu)

E-mail address: [ejs49@case.edu](mailto:ejs49@case.edu)