Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors

Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors

Daniela Calvetti[1], **Monica Pragliola**[2], **Erkki Somersalo**[1], **Alexander Strang**[1]

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

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

E-mail: dxc57@case.edu,monica.pragliola2@unibo.it, ejs49@case.edu,ags61@case.edu

Abstract. Solving inverse problems with sparsity promoting regularizing penalties can be recast in the Bayesian framework as finding a maximum a posteriori (MAP) estimate with sparsity promoting priors. In the latter context, a computationally convenient choice of prior is the family of conditionally Gaussian hierarchical models for which the prior variances of the components of the unknown are independent and follow a hyperprior from a generalized gamma family. In this paper, we analyze the optimization problem behind the MAP estimation and identify hyperparameter combinations that lead to a globally or locally convex optimization problem. The MAP estimation problem is solved using a computationally efficient alternating iterative algorithm. Its properties in the context of the generalized gamma hypermodel and its connections with some known sparsity promoting penalty methods are analyzed. Computed examples elucidate the convergence and sparsity promoting properties of the algorithm.

Efficient algorithms promoting sparsity of the solution of inverse problems have been, and continue to be actively pursued. Sparsity-promoting regularization has been known for a long time in geophysics, statistics and signal processing, see, e.g., [29, 30, 15, 16, 27], and [4] for further references. Interest in the topic was revived a little over a decade ago with the recovery of compressed sensing, because it was shown that under certain conditions about the underlying signal and the observation model, ℓ^1-regularization could recover exactly a sparse solution [19, 5]. While the regularizing properties of Tikhonov regularization with ℓ^p penalty, for $1 \leq p \leq 2$, are quite well understood from the theoretical and computational point of view, the design of efficient methods for computing the corresponding regularized solution when $p < 1$ continues to pose significant challenges due to the non-convexity of the Tikhonov functional.

Recently many classical regularization questions have been recast in probabilistic terms and efficient numerical schemes for their solutions have been proposed in the literature. In the Bayesian framework, where all unknowns are modeled as random variables, the prior density encodes what is believed about the solution before taking the data into account. Therefore, the prior plays a role similar to that of the penalty term in classical regularization. The connection between the classical Tikhonov regularized solution and the Bayesian maximum a posteriori (MAP) estimate is well established, relating classes of penalty functionals and priors favoring similar types of solutions. The philosophically correct way to promote sparsity in the Bayesian framework is by designing suitable priors. Gaussian priors, while computationally very convenient when seeking MAP estimates, are not well suited to promote sparsity unless the support of the underlying array is known. One possible way to retain the computational convenience while promoting sparsity is via conditionally Gaussian priors, with diagonal covariance matrices with unknown positive diagonal entries. In line with the Bayesian paradigm, the unknown diagonal entries of the prior covariance are modeled as random variables, and their estimation becomes part of the problem. In order for the hierarchical prior model to promote sparsity, the variances of the components of the sought solution should be independent of each other, and all but few of them should be close to zero. Thus, the prior model for the variances, usually referred to as hyperprior, should have mean close to zero and favor few outliers, implying that a suitable choice must be a fat tailed, or leptokurtic distribution. An example of such hyperprior is the gamma distribution, and a computationally efficient globally convergent algorithm for the corresponding MAP estimate can be found, referred to as the iterated alternating sequential (IAS) algorithm. Recently it has been shown that the shape parameter of the gamma distribution controls the sparsity of the solution, and in an appropriate limit, the computed solution converges to the ℓ^1 regularized solution [14, 10]. Moreover, an approximation of the MAP estimate, referred to as the qMAP for quasi-MAP, can be computed very efficiently by incorporating a Krylov subspace least squares solver equipped with a suitable stopping rule in the IAS algorithm [13]. The latter is particularly efficient for the recovery of sparse solutions, because as the algorithm learns the effective support of the underlying signal, the number of matrix-vector products with the matrix and its transpose needed for the Krylov iterative solution approaches the cardinality of the support. Because of this, the algorithm implicitly performs an adaptive and automatic model reduction.

In [14, 10], the analysis of sparsity promoting hierarchical models was limited to hyperpriors from the family of gamma distributions, a choice that guarantees the convexity of the objective function in the MAP estimation. This work extends the analysis to the wider class of hierarchical models with hyperpriors from the family of generalized gamma distributions [8], some of which are shown to correspond to earlier studied regularizing penalty functions. In particular, we identify combinations of the parameter values of the generalized gamma distribution for which the MAP objective function is either globally or locally convex, and find conditions guaranteeing that the approximate solutions determined by the IAS algorithm stay in the convexity region. By analyzing the asymptotics of the penalty function, we

show that the priors are closely related to appropriately chosen ℓ^p-penalties, helping to better contextualize and interpret the sparsity promoting properties of IAS. We point out that the ℓ^p-penalties represent only a limit case of the MAP estimates for hierarchical priors, which encompass a genuinely richer class of priors. Therefore, rather than an alternative, the proposed algorithm of computing the iterative updates for generalized gamma hyperpriors offers a uniform computational alternative for dealing with the ℓ^p regularization. Furthermore, based on ideas arising from convex optimization, we propose an extension of the IAS algorithm that can account for bound constraints. Finally, we point out that a significant reduction in the IAS computational effort can be achieved, in particular for large scale underdetermined inverse problems, by solving a reduced problem via a priorconditioned Krylov subspace solver [6, 7, 12] equipped with an early stopping criterion. Computed examples show that the performance of the reduced model can be comparable, and even superior to that of the full model for many choices of the hyperprior parameters. In particular, they illustrate how the choice of hyperprior from the generalized gamma family affects the quality of the solution in terms of sparsity promotion and accuracy, elucidating how to choose the parameters of the hyperprior to reflect the strength of the a priori sparsity belief.

The paper is organized as follows: in Section 1, the Bayesian parametric model and the algorithm for computing the MAP estimate is introduced. Section 2 discusses the model scaling and sensitivity analysis. The computation of the updating of the hyperparameters for values when no closed form formula exist is discussed in Section 3. Moreover, the connection of the model to previously discussed models such as the ℓ^p penalized regularization is discussed. Chapter 4 contains the convexity analysis of the objective function with various model parameter values. In Section 5, the IAS algorithm is extended to cases with box constraints on parameters, and computed examples are presented in Section 6, followed by Discussion.

1. Hierarchical Bayesian Models

Consider the linear observation model with additive Gaussian noise,

$$b = \mathsf{A}x + e\,, \quad e \sim \mathcal{N}(0, \Sigma), \tag{1}$$

where $\mathsf{A} \in \mathbb{R}^{m\times n}$, $m \leq n$, $\Sigma \in \mathbb{R}^{m\times m}$ is a symmetric positive definite covariance matrix that we assume to be known, and $x \in \mathbb{R}^n$ is the unknown that we are interested in recovering. Without loss of generality, we may assume that $\Sigma = \mathsf{I}$, since by defining a symmetric factorization $\Sigma^{-1} = \mathsf{S}^\mathsf{T}\mathsf{S}$, we may whiten the noise through a redefinition of $\mathsf{A} \to \mathsf{S}\mathsf{A}$ and $b \to \mathsf{S}b$. Then the likelihood distribution is of the form

$$\pi(b \mid x) \propto \exp\Big(-\frac{1}{2}\|\mathsf{A}x - b\|^2\Big). \tag{2}$$

We assume that, possibly after a change of variables, the unknown is represented in a basis where the generative vector x is sparse. The a priori belief that x is sparse is

encoded by modeling its components as independent random variables following a zero mean *conditionally Gaussian* distribution, i.e.,

$$x_j \sim \mathcal{N}(0, \theta_j)\,, \quad \theta_j > 0\,, \quad 1 \le j \le n, \tag{3}$$

with unknown prior variances θ_j. According to the Bayesian paradigm, the unknown variances are also modeled as random variables, hence the expression for the conditional Gaussian prior must take into account the portion of the normalizing factor that depends on the variances,

$$\pi(x \,|\, \theta) \propto \frac{1}{\prod_{j=1}^n \sqrt{\theta_j}} \exp\Big(-\frac{1}{2}\|\mathsf{D}_\theta^{-1/2} x\|^2\Big), \quad \mathsf{D}_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n). \tag{4}$$

In this manner, the a priori believed sparsity of x can be formulated as a property of the variances of the components, with smaller variances promoting values closer to zero. In turn, the a priori beliefs about the variance are encoded in the *hyperprior* $\pi_{\mathrm{hyper}}(\theta)$. In the Bayesian setting, where all unknowns are modeled as random variables, solving (1) is tantamount to estimating x and θ, or more generally, to exploring their joint posterior distribution conditional on b. The joint prior distribution $\pi_{\mathrm{prior}}(x, \theta)$ is the product of the conditional prior and the hyperprior. It follows from Bayes' formula that the posterior distribution $\pi(x, \theta \mid b)$ is

$$\pi(x, \theta \mid b) \propto \pi_{\mathrm{prior}}(x, \theta)\, \pi(b \mid x) = \pi(x \mid \theta)\, \pi_{\mathrm{hyper}}(\theta)\, \pi(b \mid x). \tag{5}$$

To promote sparsity of the signal, we select the hyperprior from the parametric family of generalized gamma distributions,

$$\pi_{\mathrm{hyper}}(\theta) = \pi_{\mathrm{hyper}}(\theta \mid r, \beta, \vartheta) = \frac{|r|^n}{\Gamma(\beta)^n} \prod_{j=1}^n \frac{1}{\vartheta_j} \left(\frac{\theta_j}{\vartheta_j}\right)^{r\beta - 1} \exp\left(-\left(\frac{\theta_j}{\vartheta_j}\right)^r\right), \tag{6}$$

where $r \in \mathbb{R} \setminus \{0\}$, $\beta > 0$, $\vartheta_j > 0$; further restrictions on the parameters of the generalized gamma may be necessary to guarantee finite mean and variance. Observe that the generalized gamma hyperprior family could be generalized further by letting each component θ_j have its own hyperparameter r and β. This generalization is not considered here.

The Maximum A Posteriori (MAP) estimate of (1) is the maximizer of the posterior density, thus the minimizer of the negative logarithm of the posterior,

$$(x^*, \theta^*) = \mathrm{argmin}_{x,\theta}\{-\log \pi(x, \theta \mid b)\} = \mathrm{argmin}_{x,\theta}\{\mathcal{F}(x, \theta)\}. \tag{7}$$

With our choices of prior, hyperprior, and likelihood,

$$\begin{aligned} &\mathcal{F}(x, \theta) = \mathcal{F}(x, \theta \mid r, \beta, \vartheta) \\ &= \frac{1}{2}\|\mathsf{A}x - b\|^2 + \frac{1}{2}\|\mathsf{D}_\theta^{-1/2} x\|^2 - \Big(r\beta - \frac{3}{2}\Big) \sum_{j=1}^n \log \frac{\theta_j}{\vartheta_j} + \sum_{j=1}^n \left(\frac{\theta_j}{\vartheta_j}\right)^r \\ &= \frac{1}{2}\|\mathsf{A}x - b\|^2 + \mathcal{P}(x, \theta \mid r, \beta, \vartheta). \end{aligned} \tag{8}$$

In the following, we will refer to $\mathcal{P}(x, \theta \mid r, \beta, \vartheta)$ as the penalty term in the MAP objective function. Our aim in this work is to analyze hierarchical Bayesian models with generalized gamma hyperpriors for different choices of the hyperparameters. In particular, we are interested in shedding some light on

(i) how the sparsity of the MAP estimate depends on the hyperparameters;
(ii) how, for some choices of the hyperparameters, the MAP penalty term approaches classical penalty terms;
(iii) the dependency of the convexity - or lack thereof - of the MAP objective function on the hyperparameters;
(iv) the performance of the iterative alternating sequential minimization algorithm, reviewed in the next subsection, with various generalized gamma hyperpriors for the reconstruction of sparse signals from underdetermined noisy data.

1.1. IAS Algorithm

The algorithm that we employ for the solution of the minimization problem (7) is the iterative alternating sequential scheme, whose properties and performance for some choices of hyperparameters have been analyzed in [11, 10, 13]. Given an initial θ^0, the IAS algorithm proceeds through a sequence of simple alternating updates of the form

$$x^{t+1} = \operatorname{argmin}\{\mathcal{F}(x, \theta^t)\}, \quad \theta^{t+1} = \operatorname{argmin}\{\mathcal{F}(x^{t+1}, \theta)\},$$

until a convergence criterion is met. In practice, two natural convergence criteria can be introduced: Either, the relative change of the objective function value is below a given threshold, or the relative change in the variable updates is below a threshold. In the computed examples, both criteria are used.

Among the appealing features of the IAS scheme, we mention the fact that both updating steps are particularly simple to implement, and that the algorithm has been shown to converge [11], with a convergence rate at least linear [10] for some classes of problems. We point out that the minimization can be pursued by standard optimization schemes such as the Newton method, however, the proposed algorithm is found to often lead to faster convergence. For completeness, we review the updating steps below.

1.1.1. Step 1: Updating x Due to the structure of the objective function (9), the updating of x given θ reduces to solving a quadratic optimization problem,

$$x^{t+1} = \operatorname{argmin}_x\{\|\mathsf{A}x - b\|_2^2 + \|\mathsf{D}_\theta^{-1/2} x\|_2^2\}, \quad \theta = \theta^t \tag{9}$$

or, equivalently, to finding the solution of the linear system

$$\begin{bmatrix} \mathsf{A} \\ \mathsf{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{10}$$

in the least squares sense. The latter is a well posed problem because $\theta \in \mathbb{R}^n_+$; if x is of large dimensions or A is not explicitly available, an iterative least squares solver may be the method of choice to solve (10). Due to the well-posedness of the problem, the iteration will continue until a sufficient reduction of the residual norm has been achieved. A computationally efficient way to compute an approximation of the MAP solution that is particularly appealing when the data vector is much lower dimensional than x, has been proposed in the cited articles on IAS and further analyzed in [12]. After the change of variable,

$$\mathsf{D}_\theta^{-1/2} x = w \sim \mathcal{N}(0, \mathsf{I}), \tag{11}$$

which is equivalent to whitening x via a Mahalanobis transformation, the linear system (10) becomes

$$\begin{bmatrix} \mathsf{A}_\theta \\ \mathsf{I} \end{bmatrix} w = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \mathsf{A}_\theta = \mathsf{A}\mathsf{D}_\theta^{1/2}. \tag{12}$$

It has been pointed out repeatedly in the literature that the Tikhonov regularized solution is close to the approximate solution obtained by solving the linear system

$$\mathsf{A}_\theta w = b \tag{13}$$

with an iterative linear solvers, equipped with a suitable early termination rule discussed below.

When the iterative solver selected is the Conjugate Gradient for Least Squares (CGLS) method [23], the kth iterate satisfies

$$w_k = \operatorname{argmin}\{\|b - \mathsf{A}_\theta w\| \mid w \in \mathscr{K}_k(\mathsf{A}_\theta^\mathsf{T} b, \mathsf{A}_\theta^\mathsf{T}\mathsf{A}_\theta)\}, \quad x_k = \mathsf{D}_\theta^{1/2} w_k, \tag{14}$$

where

$$\mathscr{K}_k(\mathsf{A}_\theta^\mathsf{T} b, \mathsf{A}_\theta^\mathsf{T}\mathsf{A}_\theta) = \operatorname{span}\{(\mathsf{A}_\theta^\mathsf{T}\mathsf{A}_\theta)^\ell \mathsf{A}_\theta^\mathsf{T} b \mid 0 \le \ell \le k-1\},$$

is the kth Krylov subspace associated with the vector $\mathsf{A}_\theta^\mathsf{T} b$ and the matrix $\mathsf{A}_\theta^\mathsf{T}\mathsf{A}_\theta$. The quantity $b - \mathsf{A}_\theta w_k$ whose norm is minimized is the discrepancy vector corresponding to w_k. In the traditional inverse problems literature, the Morozov discrepancy principle states that the iterations should be stopped right before the norm of the discrepancy falls below the noise level. Recalling that the standard deviation of the m-variate white noise is $\sqrt{m}$, the Morozov stopping criterion can be written as

$$\|b - \mathsf{A}_\theta w_k\| \le \sqrt{m}.$$

On the other hand, letting

$$G(w) = \left\| \begin{bmatrix} \mathsf{A}\mathsf{D}_\theta^{1/2} \\ \mathsf{I} \end{bmatrix} w - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 = \|b - \mathsf{A}\mathsf{D}_\theta^{1/2} w\|^2 + \|w\|^2, \tag{15}$$

denote the norm of the discrepancy of the original linear system (12), it follows that the least squares solution of the original problem minimizes the functional G. It is therefore natural to monitor the behavior of $G(x_k)$ as the iterations proceed, and continue iterating only as long as $G(w_k)$ keeps decreasing. While it is known that the norm of the discrepancy of (13) decreases and $\|w_k\|$ increases with the number of CGLS iterations, we do not know a priori how the increase/decrease rates are related to each other, so without further analysis, it is not clear if a minimum is reached before the maximum allowed number of iterations.

In light of these observations, we propose the following approximation to the squares solution of (12).

Definition 1.1. *The* reduced Krylov subspace *(RKS) solution for the problem (12) is w_{k_0} defined by (14), with k_0 chosen to be the first index k satisfying the criterion*

$$(\mathscr{C}): \ \|b - \mathsf{A}_\theta w_{k+1}\| \leq \sqrt{m}, \quad \textit{or} \quad G(w_{k+1}) > \tau G(w_k),$$

where $\tau - 1 = \epsilon > 0$ is a small safeguard parameter.

Finding an optimal stopping criterion for CGLS is not a simple task. In [13], an alternative, statistically motivated stopping rule based on the χ^2-error of discrepancy was suggested.

In the following, we refer to the IAS algorithm as *approximate IAS* when the minimization of (15) is replaced by the RKS solution, as opposed to the original *exact IAS*.

1.1.2. Step 2: Updating θ The update of the prior variance θ is based on the first order optimality condition. Since the parameters θ_j are mutually independent, the update can be carried out separately for each component. It follows from the form of the MAP objective function that the updated jth component of θ must satisfy the algebraic equation

$$\frac{\partial \mathcal{F}}{\partial \theta_j} = -\frac{1}{2}\frac{x_j^2}{\theta_j^2} - \left(r\beta - \frac{3}{2}\right)\frac{1}{\theta_j} + r\frac{\theta_j^{r-1}}{\vartheta_j^r} = 0, \quad x_j = x_j^{t+1}. \tag{16}$$

There are combinations of the hyperparameter values for which the solution is available in closed form, as will be discussed in detail later. We derive a computationally efficient form for the general case in the ensuing discussion.

The IAS algorithm has a similarity to a class of a reweighted least squares methods [22, 17], or fixed point iterative methods with lagged diffusivity [18], providing iterative algorithms to compute ℓ^1-regularized solutions to inverse problems. For similar alternating iterative methods in the Bayesian framework, we refer to [2, 3] applied to compressed sensing and imaging.

2. Scaling

The analyses of the IAS algorithm previously published were limited to some specific hyperpriors from the generalized gamma family. Before extending the analysis to the family

of generalized gamma hyperpriors, we reformulate the problem in non-dimensional form. To that end we introduce non-dimensional parameters z_j and ξ_j such that

$$x_j = \vartheta_j^{1/2} z_j, \quad \theta_j = \vartheta_j \xi_j,$$

and express the objective function in terms of these variables as

$$\Phi(z,\xi) = \frac{1}{2}\|\widehat{\mathsf{A}} z - b\|^2 + \frac{1}{2}\sum_{j=1}^n \frac{z_j^2}{\xi_j} - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log \xi_j + \sum_{j=1}^n \xi_j^r, \tag{17}$$

where $\widehat{\mathsf{A}}$ is a column-scaled version of A, that is,

$$\widehat{\mathsf{A}} = \left[\ \vartheta_1^{1/2} a^{(1)}, \cdots, \vartheta_n^{1/2} a^{(n)}\ \right] = \mathsf{A}\,\mathrm{diag}(\vartheta_1^{1/2}, \cdots, \vartheta_{\mathrm{n}}^{1/2}).$$

Column scaling the forward map is a common practice in some geophysics and biomedical imaging applications, where it has been motivated by sensitivity considerations. Define the sensitivity of the forward map $x \mapsto \mathsf{A}x$ with respect to the jth component x_j as

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

where e_j is the jth canonical unit vector in $\mathbb{R}^n$ and $a^{(j)}$ is the jth column of the matrix A. Then, weighting the component x_j by the corresponding sensitivity scalar can be seen as a way to avoid favoring solutions with support concentrated near the receiver locations. This observation can be used as a guidance for selecting the value of the hyperparameters ϑ_j; however, in the Bayesian framework this reasoning was considered problematic, as canonically, the selection of the prior should not depend on the observation model. Recently, however, a Bayesian justification for such choice of ϑ in the case where $r = 1$ has been given in [10, 14]. The following theorem generalizes the result to the case of general gamma hyperpriors. The result is formulated in terms of the signal-to-noise ratio (SNR) of the inverse problem (1),

$$\mathrm{SNR} = \frac{\mathsf{E}\{\|b\|^2\}}{\mathsf{E}\{\|\varepsilon\|^2\}}.$$

Theorem 2.1. *(a) Assuming that a support set $S \subset \{1, 2, \ldots, n\}$ is given, the SNR conditional to the unknown x being supported on S, denoted by SNR_S, is given by*

$$\mathrm{SNR}_S = \frac{\sum_{j\in S} \nu(r,\beta)\vartheta_j}{\mathrm{trace}(\Sigma)} + 1, \quad \nu(r,\beta) = \frac{\Gamma(\beta + 1/r)}{\Gamma(\beta)},$$

provided that $\beta > -1/r$.

(b) Let $p_k = \mathsf{P}\{\|x\|_0 = k\}$ denote the probability that the support of the signal has cardinality k, for $k = 1, 2, \ldots, n$. Then the exchangeability condition ($\mathscr{E}$),

$$(\mathscr{E}): \quad \mathrm{SNR}_S = \mathrm{SNR}_{S'} \text{ whenever } S \text{ and } S' \text{ are of the same cardinality,}$$

is satisfied if and only if ϑ_j is chosen as

$$\vartheta_j = \frac{C}{\|a^{(j)}\|^2}, \quad C = \frac{(\overline{\mathrm{SNR}} - 1)\mathrm{trace}(\Sigma)}{\nu(r,\beta)} \sum_{k=1}^n \frac{p_k}{k}.$$

Proof: The proof is a slight modification of that for the gamma hyperprior ($r = 1, \beta > 3/2$) in [10] to account for the fact that if θ_j follows the generalized gamma distribution, then

$$\mathsf{E}\{\theta_j\} = \nu(r, \beta)\vartheta_j.$$

□

An important corollary of the above theorem is that, under the stated assumptions, scaling the columns $a^{(j)}$ by $\vartheta_j^{1/2}$ is tantamount to making them all of the same norm $\sqrt{C}$. From the point of view of linear inverse problems, this scaling renders the data equally sensitive to each component of the unknown x. While this sensitivity scaling has been used for decades in geophysics and biomedical imaging applications to compensate for the tendency of regularized variational methods to favor solutions concentrated near the receivers or solutions parametrized by components with maximal sensitivity, to the best of our knowledge this is the first proper Bayesian interpretation of sensitivity scaling.

Furthermore, as already pointed out in [10], the theorem provides a Bayesian argument to choose the value of Tikhonov regularization parameter in linear inverse problems from an estimated SNR and a priori belief about the cardinality of the support. The effect of the sensitivity scaling has been demonstrated in [10] by computed examples.

3. Variance updating: a closer look

In this section, we analyze in detail the process of updating the variance vector given an updated estimate of the signal. After scaling the variables as described in the previous section to arrive at a non-dimensional formulation, the algebraic relation (16) for the non-dimensional variance ξ_j given the non-dimensional signal z_j becomes

$$-\frac{1}{2}z^2 - \eta\xi + r\xi^{r+1} = 0, \quad \eta = r\beta - \frac{3}{2}, \tag{18}$$

where we dropped the subscript j to simplify the notation. Since the expression depends only on the square of z, we restrict our discussion to the case where z assumes non-negative values, the negative values being covered by symmetry

The following result characterizes the variance as the solution of an initial value problem.

Lemma 3.1. *If $r < 0$ and $\eta < -3/2$, or $r > 0$ and $\eta > 0$, formula (18) defines an implicit function*

$$\varphi(z) = \xi, \quad \varphi : \mathbb{R}_+ \to \mathbb{R}_+,$$

which is smooth and strictly increasing. Moreover, ξ is the solution of the initial value problem

$$\varphi'(z) = \frac{2z\varphi(z)}{2r^2\varphi(z)^{r+1} + z^2}, \quad \varphi(0) = \left(\frac{\eta}{r}\right)^{1/r}. \tag{19}$$

Proof: Starting from (18), we define the function

$$g(\xi) = \xi(r\xi^r - \eta), \quad 0 < \xi_0 = \left(\frac{\eta}{r}\right)^{1/r} \leq \xi < \infty,$$

which is differentiable, with $g'(\xi) = -\eta + r(r+1)\xi^r$. For $r \leq -1$ and $\eta < -3/2$, the derivative is always positive. For $-1 < r < 0$ and $\eta < -3/2$, the condition $\xi > \xi_0$ implies that $\xi^r < \eta/r$, and consequently, $g'(\xi) > r\eta > 0$, and for $r > 0$ and $\eta > 0$, we have $\xi^r > \eta/r$, and therefore $g'(\xi) > \eta r > 0$. Hence, the function $g(\xi)$ is strictly increasing for $\xi > \xi_0$. Furthermore,

$$g(\xi_0) = 0, \quad \lim_{\xi\to\infty} g(\xi) = \infty.$$

Therefore the equation

$$g(\xi) = \xi(r\xi^r - \eta) = \frac{1}{2}z^2$$

has a unique solution $\xi = \xi(z) \in [\xi_0, \infty)$ for every $z \geq 0$, hence the mapping from z to the solution (18) defines a strictly increasing function $\xi = \varphi(z)$. We have $\varphi(z) = g^{-1}(z^2/2)$ and by the implicit function theorem, the function g^{-1}, as an inverse of a differentiable function is also differentiable, therefore φ is differentiable. Substituting $\xi = \varphi(z)$ in (18) ,

$$-\frac{1}{2}z^2 - \eta\varphi(z) + r\varphi(z)^{r+1} = 0, \tag{20}$$

and differentiating with respect to z, we get

$$((r+1)r\varphi(z)^r - \eta)\varphi'(z) = z,$$

or, equivalently,

$$\Big(r^2\varphi(z)^r + \frac{1}{\varphi(z)}(r\varphi(z)^{r+1} - \eta\varphi(z))\Big)\varphi'(z) = \Big(r^2\varphi(z)^r + \frac{z^2}{2\varphi(z)}\Big)\varphi'(z) = z,$$

yielding the differential equation (19). □

The characterization of ξ in terms of the differential equation (19) can be used to compute effectively the values of the updates of the variances in the IAS algorithm, as we will show in the computed examples. Moreover, Lemma 3.1 also makes it possible to analyze the asymptotic behavior of the variance parameter when the corresponding value of z is either close to zero or very large.

Lemma 3.2. *The asymptotic behavior of φ when z is close to zero is*

$$\varphi(z) = \left(\frac{\eta}{r}\right)^{1/r} + \frac{1}{2\eta r}z^2 + \mathcal{O}(z^4). \tag{21}$$

whereas the asymptotics for $z > 0$ large is

$$\varphi(z) = \kappa z^{2/(r+1)}\left(1 + o(1)\right), \quad \kappa = \left(\frac{1}{2r}\right)^{1/(r+1)} \tag{22}$$

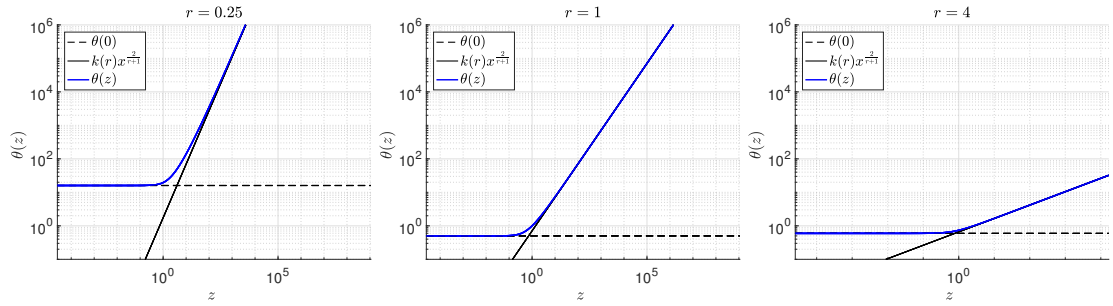

Figure 1: Logarithmic plots of the updating functions with different values of the parameters r, with $\eta = 0.5$ in each case. The asymptotics given by Lemma 3.2 as well as the initial values $\varphi(0) = (\eta/r)^{1/r}$ are indicated in this figure.

when $r > 0$, *and*

$$\varphi(z) = \kappa z^2 \left(1 + o(1)\right), \quad \kappa = \frac{1}{2|\eta|}$$

when $r < 0$.

Proof. The asymptotic behavior of φ for z near zero can be obtained from its Taylor expansion at $z = 0$. It follows from (19) that $\varphi(0) = \left(\frac{\eta}{r}\right)^{1/r}$, $\varphi'(0) = 0$, and differentiating (19) with respect to z yields

$$\varphi''(0) = \frac{1}{r^2 \varphi(0)^r} = \frac{1}{r\eta},$$

The asymptotic estimate follows from the observation that the third derivative of φ vanishes at $z = 0$.

To obtain the asymptotics of φ for large z and $r > 0$, observe that (20) implies

$$\lim_{z\to\infty} \varphi(z) \to \infty,$$

therefore, since $(1 + o(1))^{-1} = 1 + o(1)$, for large z,

$$\frac{1}{2}z^2 = \varphi(z)^{r+1}\left(r - \frac{\eta}{\varphi(z)^r}\right) = r\varphi(z)^{r+1}\left(1 + o(1)\right),$$

implying (22). Similarly, if $r < 0$, we write

$$\frac{1}{2}z^2 = \xi\left(|\eta| - \frac{|r|}{\xi^{|r|}}\right) = |\eta|\xi\left(1 + o(1)\right),$$

completing the proof. □

Figure 1 shows the updating functions in a logarithmic scale with selected values of the parameter r.

The asymptotic behavior of the updating function helps us to interpret the role of the model parameters r and β. For this interpretation, we need the following theorem establishing the equivalence of the IAS optimization of the objective function with respect to the pair (z, ξ) in $\mathbb{R}^{2n}$ and the optimization of the objective function along the manifold $\xi = \varphi(z)$,with the last equality to be understood as componentwise.

Lemma 3.3. *Let (z^*, ξ^*) be a local minimizer of the objective function $\Phi(x, \xi)$ given by (17). Then, the point z^* is a local minimizer of $\Psi(z) = \Phi(z, \varphi(z))$. Conversely, if z^* is a local minimizer of $\Psi(z)$, then $(z^*, \varphi(x^*))$ is a local minimizer of $\Phi(x, \xi)$.*

Proof: If (z^*, ξ^*) is a local minimizer of Φ, then it must satisfy

$$\frac{\partial \Phi}{\partial \xi_j}(z^*, \xi^*) = 0, \text{ implying that } \xi^* = \varphi(z^*).$$

Let $U = B_1 \times B_2 \in \mathbb{R}^{2n}$ be a neighborhood of (z^*, ξ^*) such that for any $(z, \xi) \in U$, $\Phi(z^*, \xi^*) \leq \Phi(z, \xi)$. Since φ is continuous, for each z in some neighborhood $B_1' \subset B_1$ of z^*, $\varphi(z) \in B_2$, therefore $\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z))$, that is, z^* is a local minimizer of Ψ.

Conversely, let z^* be a local minimizer of Ψ. Then, there is a neighborhood B of z^* such that for any $z \in B$, $\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z))$. However, for each z, $\theta = \varphi(z)$ is the unique minimizer of $\theta \mapsto \Phi(z, \theta)$, therefore

$$\Phi(z^*, \varphi(z^*)) \leq \Phi(z, \varphi(z)) \leq \Phi(z, \theta), \quad (z, \theta) \in B \times \mathbb{R}^+,$$

implying that $(z^*, \varphi(z^*))$ indeed is a local minimizer of Φ.□

It follows from the lemma that in order to understand the sparsity promoting properties of the various hyperpriors, one can consider the objective function $\Psi(z) = \Phi(z, \varphi(z))$, and in particular, the scaled penalty term

$$\Pi(z) = \frac{1}{2} \sum_{j=1}^{n} \frac{z_j^2}{\varphi(z_j)} - \eta \sum_{j=1}^{n} \log \varphi(z_j) + \sum_{j=1}^{n} \varphi(z_j)^r.$$

We will use this observation together with the asymptotic forms of the updating function to elucidate how the regularization properties of the penalty functions change with the hyperparameter values. Before addressing the general case, we consider some special choices of the parameter values.

3.1. Special generalized gamma hyperpriors

There are hyperparameter combinations for which the updating function is available in closed form. Some of these special cases have been used in numerical computations in earlier works [9, 8].

3.1.1. Gamma distribution and ℓ^1 prior The most thoroughly analyzed hyperprior in the context of the IAS algorithm is the gamma distribution, which is a generalized gamma with $r = 1$ and $\eta > 0$. With that choice of parameters, equation (16) simplifies to

$$-\frac{1}{2}z^2 - \eta\xi + \xi^2 = 0,$$

and can be readily solved for ξ, yielding

$$\xi = \varphi(z) = \frac{1}{2}\left(\eta + \sqrt{\eta^2 + 2z^2}\right).$$

As pointed out in [13, 10], substituting $\xi_j = \varphi(z_j)$ in the MAP penalty function and letting η go to zero yields

$$\begin{aligned}\Pi(z) = \sum_{j=1}^{n}\left\{\frac{1}{2}\frac{z_j^2}{\xi_j} - \eta\log\xi_j + \xi_j\right\} = \sum_{j=1}^{n}\Big\{&\frac{z_j^2}{\eta + \sqrt{\eta^2 + 2z_j^2}} \\ &- \eta\log\frac{1}{2}(\eta + \sqrt{\eta^2 + 2z_j^2}) + \frac{1}{2}(\eta + \sqrt{\eta^2 + 2z_j^2})\Big\} \\ &\to \sqrt{2}\sum_{j=1}^{n}|z_j|, \text{ as } \eta \to 0+,\end{aligned}$$

that is, in the limit, the penalty function approaches the ℓ^1-penalty. In [10], it was further shown that the unique solution of the IAS algorithm converges to the solution with the ℓ^1-penalty, thus recovering a compressible solution, if the data came from a sparse generative model. For further results, we refer to [10].

3.1.2. Inverse gamma distribution and Student prior The second special case is that of the inverse gamma hyperprior, corresponding to setting $r = -1$. In this case, equation (16) becomes

$$\frac{1}{2}z^2 - \eta\xi - 1 = 0,$$

and the update formula is

$$\xi = \varphi(z) = \frac{1}{2k}(z^2 + 2), \quad k = \beta + \frac{3}{2}.$$

As for the gamma hyperprior, substituting $\xi_j = \varphi(x_j)$ in the MAP penalty functional yields

$$\begin{aligned}\Pi(z) &= \sum_{j=1}^{n}\left\{\frac{1}{2}\frac{z_j^2}{\xi_j} + k\log\xi_j + \frac{1}{\xi_j}\right\} = \sum_{j=1}^{n}\left\{\frac{z_j^2 + 2}{2\xi_j} + k\log\xi_j\right\} \\ &= n(k - \log 2k) + \sum_{j=1}^{n}\log(z_j^2 + 2)^k,\end{aligned}$$

which corresponds to the prior model

$$\pi_{\text{prior}}(z) \propto \exp(-\Pi(z)) \propto \prod_{j=1}^{n}\frac{1}{(z_j^2 + 2)^k},$$

We observe that as $\beta \to 0+$, $k \to 3/2$, and the distribution of the individual components z_j approaches the Student distribution,

$$St(z \mid \nu) \propto \frac{1}{\left(1 + \frac{z^2}{\nu}\right)^{(\nu+1)/2}},$$

with parameter $\nu = 2$, a prominently fat tailed distribution that favors outliers, thus promoting sparsity.

3.1.3. Generalized gamma and ℓ^p prior The third special case that we consider here is that where $r\beta = 3/2$, for which the update formula becomes

$$-\frac{1}{2}z^2 + r\xi^{r+1} = 0,$$

or

$$\xi = \varphi(z) = \frac{|z|^{2/(r+1)}}{(2r)^{1/(r+1)}}.$$

Substituting $\xi_j = \varphi(z_j)$ in the MAP penalty functional we get

$$\begin{aligned}\Pi(z) &= \sum_{j=1}^{n} \left\{ \frac{1}{2}\frac{z_j^2}{\xi_j} + \xi_j^r \right\} = \sum_{j=1}^{n} \left\{ \frac{(2r)^{1/(r+1)}}{2} |z_j|^{2-2/(r+1)} + \frac{1}{(2r)^{1/(r+1)}} |z_j|^{2r/(r+1)} \right\} \\ &= C_r \sum_{j=1}^{n} |z_j|^{2r/(r+1)}, \quad C_r = \frac{r+1}{(2r)^{r/(r+1)}}.\end{aligned}$$

and letting

$$p = \frac{2r}{r+1}, \quad 0 < p < 2,$$

yields

$$\Pi(z) = C_r \sum_{j=1}^{n} |z_j|^p, \quad 0 < p < 2$$

The ℓ^p-penalties for $0 < p < 1$ are known for their sparsity promoting properties, and have been analyzed extensively in the literature. However, since are non-convex, they pose challenges when it comes to computing the corresponding regularized solution.

3.2. General case: asymptotics

Consider now the penalty functional $\Pi(z) = \sum_{j=1} \Pi_j(z_j)$ in the general case with $r > 0$. From Lemma 3.2 we see that if $|z_j|$ is large, the penalty function of the jth component can be

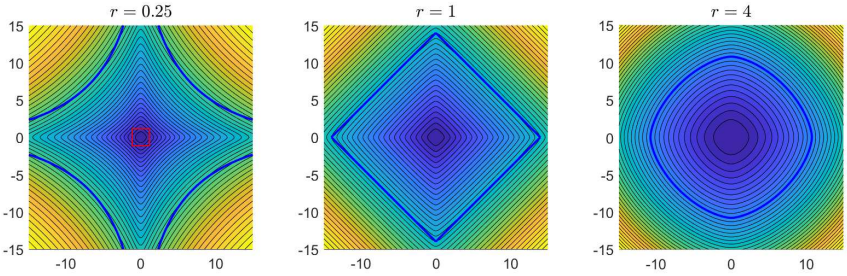

Figure 2: Level set plots of the reduced penalty function $\Pi(z)$ in two dimensions with different values of r, corresponding asymptotically to ℓ^p-penalties with $p = 2/5$ (left), $p = 1$ (center) and $p = 8/5$ (right). The corresponding ℓ^p-sphere is superimposed with dark blue. The boundary of the convexity region (see Section 4) for $r = 1/4$ is marked by the red square.

written as

$$\begin{aligned}\Pi_j(z_j) &= \frac{1}{2}\frac{z_j^2}{\varphi(z_j)} + \varphi(z_j)^r - \eta\log\varphi(z_j)\\ &= \frac{1}{2\kappa}|z_j|^{2-2/(r+1)}(1+o(1)) + \kappa^r|z_j|^{2r/(r+1)}(1+o(1))\\ &\qquad\qquad - \frac{2\eta}{r+1}\log(|z_j|(1+o(1)))\\ &\propto |z_j|^p(1+o(1)), \quad p = \frac{2r}{r+1}.\end{aligned}$$

Similarly, for small values of $|z_j|$, (21) yields the asymptotic estimate

$$\begin{aligned}\Pi_j(z_j) &= \frac{1}{2}\frac{z_j^2}{a+bz_j^2+\mathcal{O}(z_j^4)} + (a+bz_j^2+\mathcal{O}(z_j^4))^r - \eta\log(a+bz_j^2+\mathcal{O}(z_j^4))\\ &= C_1 + C_2 z_j^2 + \mathcal{O}(z_j^4))\end{aligned}$$

with $a = (\eta/r)^{1/r}$, $b = 1/(2\eta r)$, and C_1 and C_2 are some scalars. Therefore, for large $|z_j|$, the penalty behaves like an ℓ^p-penalty with $p = 2r/(r+1) \in (0,2)$, while for small $|z_j|$, the penalty is essentially Gaussian.

Figure 2 shows the level curves of the function $\Pi(z)$ in two dimensions for some parameter choices. From these plots it is clear that for large values of $\|z\|$, the level sets look like the ℓ^p spheres, while for small values, the level curves become increasingly circular as predicted by the asymptotic formulas.

4. Convexity

Our first goal in this section is to find out for which choices of the parameters (r,β) the objective function Φ given by (17) is globally convex for all $(z,\xi) \in \mathbb{R}^n \times \mathbb{R}^n_+$, or, alternatively, convex in a specified subset. The following theorem summarizes the results.

Theorem 4.1. *Let $\beta > 0$ and $r \neq 0$, and let $\Phi(z,\xi) = \Phi(z,\xi \mid r,\beta)$ be the objective function (17) for the dimensionless formulation of the problem.*

(a) If $r \geq 1$ and $\eta = r\beta - 3/2 > 0$, the function $\Phi(z,\xi)$ is convex everywhere.

(b) If $0 < r < 1$ and $\eta = r\beta - 3/2 > 0$, or, if $r < 0$ and $\beta > 0$, the function $\Phi(z,\xi)$ is convex provided that

$$\xi_j < \overline{\xi} = \left(\frac{\eta}{r|r-1|}\right)^{1/r}.$$

Proof: Recall that the positive definiteness of the Hessian is a sufficient condition for the convexity of the underlying functional. Consider the block partitioning of the Hessian of Φ,

$$\mathsf{H} = \mathsf{H}(z,\xi) = \begin{bmatrix} \nabla_z\nabla_z\Phi(z,\xi) & \nabla_z\nabla_\xi\Phi(z,\xi) \\ \nabla_\xi\nabla_z\Phi(z,\xi) & \nabla_\xi\nabla_\xi\Phi(z,\xi) \end{bmatrix},$$

where,

$$\begin{aligned}
\nabla_z\nabla_z\Phi(z,\xi) &= \mathsf{D}_\xi^{-1} + \widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}}, \\
\nabla_z\nabla_\xi\Phi(z,\xi) &= \nabla_\xi\nabla_z\Phi(x,\theta) = \operatorname{diag}\Big(-\frac{z_j}{\xi_j^2}\Big), \\
\nabla_\xi\nabla_\xi\Phi(z,\xi) &= \operatorname{diag}\Big(\frac{z_j^2}{\xi_j^3} + r(r-1)\xi_j^{r-2} + \eta\frac{1}{\xi_j^2}\Big),
\end{aligned}$$

For any vector $q = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{2n}$, we have

$$\begin{aligned}
q^T\mathsf{H}q &= \|\widehat{\mathsf{A}}u\|^2 + \sum_{j=1}^n \frac{u_j^2}{\xi_j} + \sum_{j=1}^n \Big(\frac{z_j^2}{\xi_j^3}v_j^2 + r(r-1)\xi_j^{r-2}v_j^2 + \eta\frac{v_j^2}{\xi_j^2}\Big) - 2\sum_{j=1}^n \frac{z_j}{\xi_j^2}u_jv_j \\
&= \|\widehat{\mathsf{A}}u\|^2 + \sum_{j=1}^n \frac{1}{\xi_j}\Big(u_j - \frac{z_j}{\xi_j}v_j\Big)^2 + \sum_{j=1}^n \phi_j(\xi_j \mid r,\beta)v_j^2, \qquad (23)
\end{aligned}$$

where

$$\phi_j(\xi_j \mid r,\beta) = r(r-1)\xi_j^{r-2} + \eta\frac{1}{\xi_j^2}. \qquad (24)$$

Note that the first two terms in (23) are always non-negative, so the positivity of the quadratic form defined by the Hessian it is guaranteed if $\phi_j(\xi_j \mid r,\beta) > 0$ for all j, $1 \leq j \leq n$. The proof for the different cases follows by enforcing this condition. □

Figure 3 shows the regions in the r,β plane corresponding to hyperparameter choices leading to convex or conditional convex MAP objective functions. Observe that the ℓ^p-penalty corresponds to the boundary $\beta = 3/(2r)$, with $p = 2r/(r+1)$. In particular, for $p \leq 1$, the generalized gamma family provides nearby penalty functionals that yield at least a locally convex objective function. Similarly, for $r = -1$, the non-convex Student distribution penalty corresponds to the limit $\beta \to 0+$, while for $\beta > 0$, the convexity radius is positive.

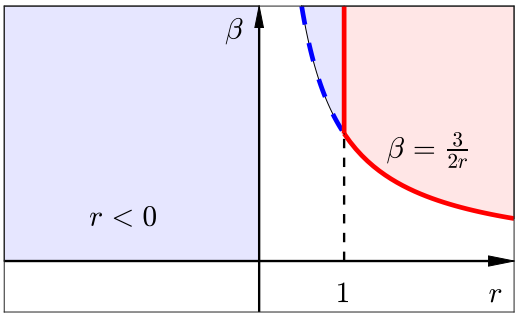

Figure 3: Convexity regions in the (r, β) plane. The red shadowing denotes parameter choices for which the MAP objective function is convex everywhere, and the blue shadowing parameter choices for which the MAP objective function is only locally convex. The curve $\beta = 3/(2r)$ marks the parameter pairs for which the hierarchical model is an ℓ^p-penalty priors, with $p = 2r/(r+1)$, which are convex if $p \geq 1$ or, equivalently, $r \geq 1$. The vertical line $r = 1$ corresponds to the family of gamma hyperpriors.

We define the *convexity radius* $\rho = \rho(r, \beta) \geq 0$, by

$$\rho = \varphi^{-1}(\overline{\xi}),$$

that is, for $\|z\|_\infty < \rho$, we have $\|\xi\|_\infty < \overline{\xi}$ guaranteeing the convexity. If the objective function is globally convex as in the case (a) of the Theorem 4.1, we set $\rho = \infty$.

4.1. Stable convexity

Consider the IAS algorithm for computing the MAP estimate, and denote the current iterate by (z^t, ξ^t). The update of z requires the solution of the minimization problem

$$z^{t+1} = \operatorname{argmin} \left\{ \frac{1}{2} \|\widehat{\mathsf{A}} z - b\|^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{z_j^2}{\xi_j^t} \right\}.$$

We say that $\Phi(z, \xi)$ is *stably convex* if there is a $T > 0$ such that, for $t > T$,

$$\|z^t\|_\infty < \rho \Rightarrow \|z^{t+1}\|_\infty < \rho,$$

in other words, stable convexity is tantamount to guaranteeing that once the IAS iterates (z^t, ξ^t) enter the convexity basin they do not leave it, thus keeping the optimization problem convex.

To find a sufficient condition for stable convexity, we need an estimate of the ℓ^∞-norm of the least squares solution of the system

$$\begin{bmatrix} \widehat{\mathsf{A}} \\ \mathsf{D}_\xi^{-1/2} \end{bmatrix} z = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \|\xi\|_\infty < \overline{\xi}.$$

In the following, we assume that the columns $\widehat{a}^{(j)}$ of $\widehat{\mathsf{A}}$ have been scaled according to the sensitivity and satisfy $\|\widehat{a}^{(j)}\| = C^{1/2}$. The following lemma provides an estimate for the size of the components of the updated z.

Lemma 4.2. *Assume that $\xi_j < \overline{\xi}$. Then the entries of the solution z of the normal equations,*

$$(\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}} + \mathsf{D}_{\xi}^{-1})z = \widehat{\mathsf{A}}^{\mathsf{T}}b$$

satisfy

$$|z_j| \leq \frac{C\xi_j}{1 + C\xi_j}(hC\overline{\xi} + 1)\|\widehat{\mathsf{A}}^{\mathsf{T}}b\|_2, \quad h = \max_j \{ \sum_{i \neq j} |\cos \angle(\widehat{a}^{(i)}, \widehat{a}^{(j)})| \}.$$

Proof: In terms of the quadratic forms associated with the symmetric positive definite matrices, we have that

$$\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}} + \mathsf{D}_{\xi}^{-1} \geq \mathsf{D}_{\xi}^{-1},$$

from which it follows that

$$(\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}} + \mathsf{D}_{\xi}^{-1})^{-1} \leq \mathsf{D}_{\xi},$$

establishing the following inequality for the induced ℓ^2-norms,

$$\|(\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}} + \mathsf{D}_{\xi}^{-1})^{-1}\|_2 \leq \|\mathsf{D}_{\xi}\|_2 \leq \overline{\xi},$$

and further, the estimate

$$\|z\|_{\infty} \leq \|z\|_2 \leq \overline{\xi}\|\widehat{\mathsf{A}}^{\mathsf{T}}b\|_2. \tag{25}$$

Next we express $\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}}$ as the sum of the two matrices $C\mathsf{I}$ and R containing, respectively, its diagonal and off-diagonal entries,

$$\widehat{\mathsf{A}}^{\mathsf{T}}\widehat{\mathsf{A}} = C\mathsf{I} + \mathsf{R}, \quad \mathsf{R}_{ij} = \begin{cases} C\cos\angle(\widehat{a}^{(i)}, \widehat{a}^{(j)}), & i \neq j, \\ 0, & i = j. \end{cases}$$

A substitution of this expression in the normal equations gives

$$\mathrm{diag}(C + 1/\xi_j)z + \mathsf{R}z = \widehat{\mathsf{A}}^{\mathsf{T}}b,$$

yielding the following upper bounds for the components of the solution,

$$|z_j| \leq \frac{\xi_j}{C\xi_j + 1}\left(|(\mathsf{R}z)_j| + |(\widehat{\mathsf{A}}^{\mathsf{T}}b)_j|\right).$$

Furthermore, since

$$|(\mathsf{R}z)_j| \leq \|\mathsf{R}\|_{\infty}\|z\|_{\infty} = \max_k |\sum_{i \neq k} \mathsf{R}_{ik}|\|z\|_{\infty} = Ch\|z\|_{\infty},$$

replacing $\|z\|_\infty$ with its upper bound from (25) and observing that $\|(\widehat{\mathsf{A}}^\mathsf{T} b)_j| \leq \|\widehat{\mathsf{A}}^\mathsf{T} b\|_2$, we have

$$|z_j| \leq \frac{\xi_j}{C\xi_j + 1}\left(Ch\|z\|_\infty + \|(\widehat{\mathsf{A}}^\mathsf{T} b)\|_2\right) \leq \frac{C\xi_j}{1 + C\xi_j}(hC\overline{\xi} + 1)\|\widehat{\mathsf{A}}^\mathsf{T} b\|_2$$

thus completing the proof. □

While not definitive, the previous lemma points to some of the factors that contribute to the stable convexity. First, we observe that if $|z_j^t| \ll 1$, choosing the shape parameter $\eta > 0$ small implies that $\xi_j^t = \varphi(|z_j^t|) \ll 1$. The above lemma suggests that in the IAS iterations, small entries remain small, and therefore one can hope that they remain below the convexity bound. On the other hand, if the columns of the matrix $\widehat{\mathsf{A}}$ are almost orthogonal, we have $h \ll 1$, we have an upper bound close to the norm $\|\widehat{\mathsf{A}}^\mathsf{T} b\|$ for the entries $|z_j|$. In such case, choosing the parameters (r, β) so that $\rho = \tau\|\widehat{\mathsf{A}}^\mathsf{T} b\|_2$ for some safeguard factor $\tau > 1$ guarantees stable convexity of the objective function. The quantity h is closely related to the mutual coherence of the matrix [21], and the Welch bounds for frames, widely studied in frame theory and signal processing literature [31].

Figure 4 shows graphically the convexity radius as a function of the parameters r and η, as well as the evolution of the level curves in two dimensions of the reduced objective function together with the convexity spheres. Eight selected zoomed-in tiles of the panel are shown in Figure 5. For $r > 0$, the ℓ^p-spheres corresponding to the large norm asymptotics are also plotted to underline the similarity between them and the contour lines.

Remark 4.3. *In general, one may not have an a priori guarantee that the components of the unknown are bounded by a constant smaller than the convexity radius. However, if we know that a priori, $|x_j| < M$ for some $M > 0$, we may choose the parameters (r, η) so that $\rho \geq M$, guaranteeing global convexity and thus the existence of a unique minimizer. However, such parameter adjustment is a non-trivial optimization problem that is not addressed here. A natural question that arises then is, how the IAS algorithm should be modified for a case in which a box constraint is part of the prior. This question is addressed in the next section.*

5. IAS with bound constraints

Consider the constrained optimization problem:

$$\text{minimize } \Phi(z, \xi) \text{ subject to the constraints } 0 \leq z \leq H,$$

for some $H > 0$. The minimizer corresponds to the MAP estimate under the belief that the components of the solution are nonnegative and not larger than H. More general box contraints can be treated in a similar way.

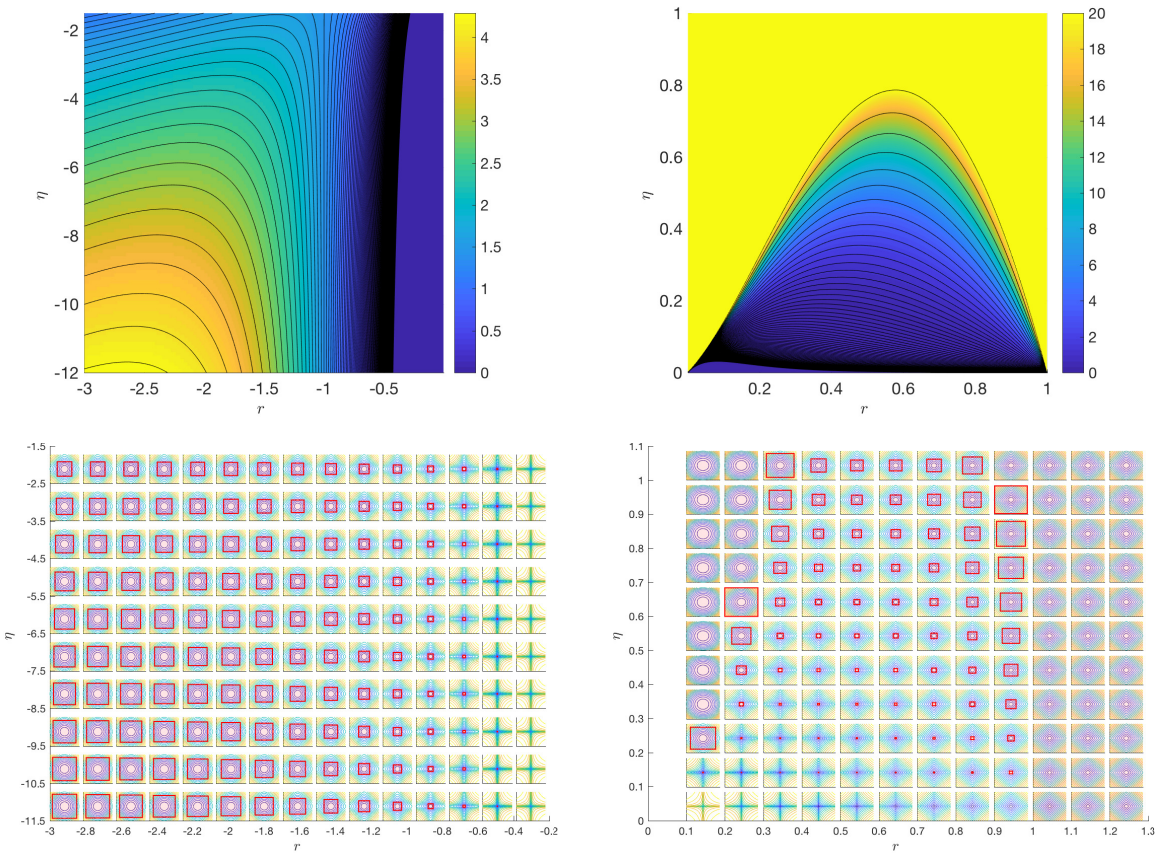

Figure 4: The radius of the convexity region as a function of r and η for generalized gamma hyperpriors with $-3 \leq r < 0$ (top left) and $0 < r \leq 1$ (top right). The two panels on the bottom row show, for different choices of (r, η) in the generalized gamma family, the level curves of the corresponding functionals. In each tile, a red curve, if present, marks the boundary of the region inside which the functional is convex. The absence of a red curve indicates that for that choice of (r, η) the functional is always convex.

We begin by introducing the penalty function

$$G(z) = \begin{cases} 0, & \text{when } 0 < z \leq H, \\ \infty & \text{otherwise,} \end{cases}$$

and write the posterior density with the bound constraints as

$$\pi(z, \xi \mid b) \propto \exp\left(-\Phi(z, \xi) - G(z)\right) = \exp\left(-\Phi_G(z, \xi)\right).$$

Following the ideas in [28, 20], consider the Moreau-Yoshida envelope of the objective function,

$$\Phi_G^\lambda(z, \xi) = \Phi(z, \xi) + G^\lambda(z),$$

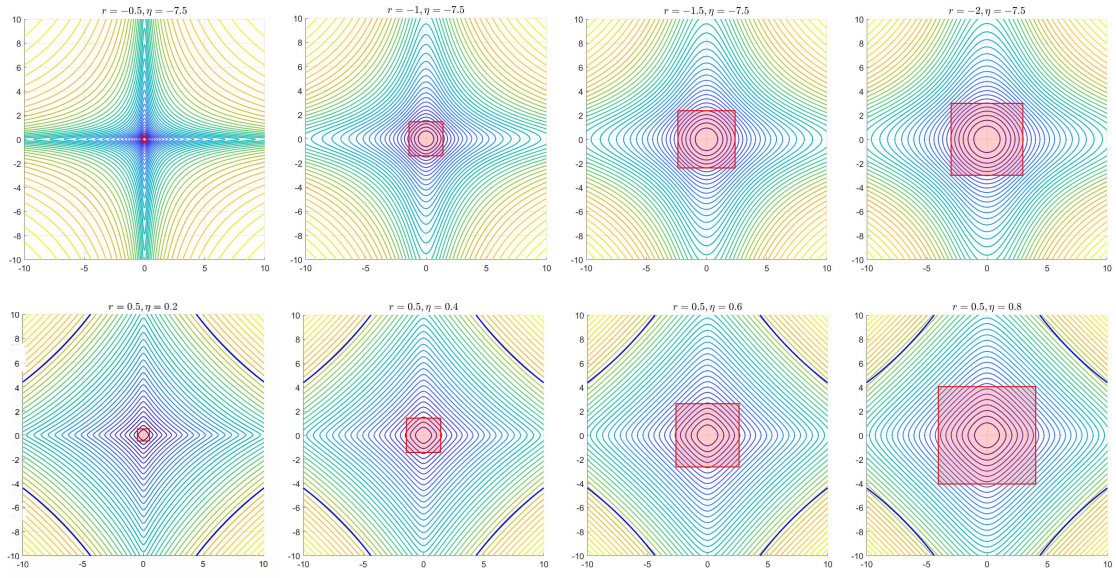

Figure 5: Eight of the panels in figure 4 corresponding to negative but varying r keeping η fixed (upper row), and positive fixed r with varying η (lower row). The convexity region is shaded in red. For positive values of r, the plot of the ℓ^p-ball corresponding to the asymptotic behavior is also plotted to underline the similarity with the contour lines.

where

$$G^{\lambda}(z) = \min_{u\in\mathbb{R}^n}\left\{G(u) + \frac{1}{2\lambda}\|z-u\|^2\right\},$$

with $\lambda > 0$ an auxiliary parameter. Its can be shown [26] that the Moreau-Yoshida envelope is differentiable with respect to z, and its gradient is of the form

$$\nabla_z \Phi_G^{\lambda}(z,\xi) = \nabla_z \Phi(z,\xi) + \frac{1}{\lambda}(z - \mathrm{prox}_G^{\lambda}(z)),$$

where the proximal operator is defined as

$$\mathrm{prox}_G^{\lambda}(z) = \mathrm{argmin}_{u\in\mathbb{R}^n}\left\{G(u) + \frac{1}{2\lambda}\|z-u\|^2\right\} = \begin{cases} z, & \text{if } G(z) = 0 \\ \mathsf{P}z, & \text{if } G(z) = \infty \end{cases},$$

and P is the orthogonal projector onto the feasible set $[0,H]^n$. Since the derivatives of the objective function with respect to the parameters ξ_j are unaffected by the inclusion of the bounds, a natural extension of the IAS algorithm for bound constrained problems can be obtained by modifying the solution of the least squares minimization problem as follows:

Given the current ξ^t:

(a) Find $z = z^*$ by solving $\nabla_z \Phi(z,\xi^t) = 0$ in the least squares sense,

(b) Define $z^{t+1} = \mathrm{prox}_G^{\lambda}(z^*)$ as the projection of z^* onto the feasible set.

Observe that unlike when exploring the posterior, for the computation of the MAP estimate it is not necessary to specify the auxiliary parameter λ, as the proximal operator is a projection regardless of the value of λ. It was proved in [20] that, as $\lambda \to 0+$, the posterior distribution defined in terms of the Moreau-Yoshida envelope converges in the sense of total variation towards the posterior distribution augmented by the positivity constraint.

6. Computed examples

In this section, we present computed examples that illustrate how the choice of the hyperprior from the generalized gamma family affects the sparsity promotion of the computed MAP solution. Moreover, we address some computational aspects when using a hypermodel that does not allow a closed form updating formula for the parameter ξ.

6.1. Example 1

The first computed example is a one-dimensional deconvolution problem with and Airy convolution kernel. The generative model is a piecewise constant signal $f : [0, 1] \to \mathbb{R}$, $f(0) = 0$, and the data consist of discrete noisy observations,

$$b_j = \int_0^1 A(s_j - t)f(t)dt + \varepsilon_j, \quad 1 \le j \le m, \quad A(t) = \left(\frac{J_1(\kappa|t|)}{\kappa|t|}\right)^2,$$

where J_1 is the Bessel function of the first kind and κ is a scaling controlling the width of the kernel. We set $\kappa = 40$, yielding a kernel with $\mathrm{FWHM} = 0.08$. We discretize the integral as

$$\int_0^1 A(s_j - t)f(t)dt \approx \sum_{j=1}^{n} w_k A(s_j - t_k)f(t_k), \quad 1 \le k \le n,$$

where $t_k = (k-1)/(n-1)$ and the w_k's are the trapezoidal quadrature weights. To generate the data, we use a dense discretization with $n = n_{\mathrm{dense}} = 1253$, while the forward model used for solving the inverse problem assumes $n = 500$. The observation points are given by $s_j = (4 + j)/100$, $1 \le j \le m = 91$, and the noise added is assumed to be scaled white noise, with standard deviation σ set to 1% of the noiseless generated signal. We denote $x_j = f(t_j)$. Figure 6 shows the generative signal and the data.

To compute the update of the hyperparameter ξ given the current vector z, we first sort the values of z so that $0 \le |z_{j_1}| \le \ldots \le |z_{j_n}|$, and subsequently solve numerically the differential equation (19) at these values. Observe that this solution is fast since the propagation needs not to be restarted from zero, but rather we only need to propagate from $|z_{j_\ell}|$ to $|z_{j_{\ell+1}}|$ to get the next value. The integration was done using the RK45 solver of Matlab.

While the generative signal, a piecewise constant function, is not sparse, it admits a sparse representation in terms of its increments $z_j = x_j - x_{j-1}$ over the interval of definition. If we

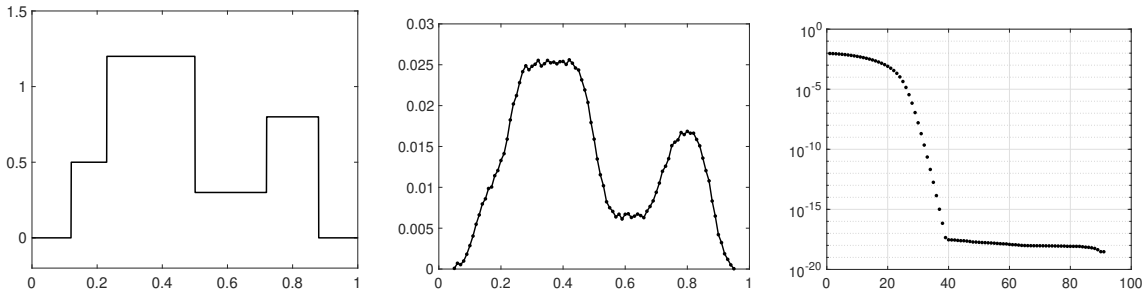

Figure 6: Left: The generative model. Center: the blurred and noisy data vector $b \in \mathbb{R}^{91}$. Right: Singular values of the discrete blurring kernel $\mathsf{A} \in \mathbb{R}^{91\times 500}$ used for the solution of the inverse problem. Since only the first 30 singular values are significantly different from zero, the matrix is numerically singular.

assume that $x_0 = 0$, then

$$z = \mathsf{D}\,x\,, \quad \mathsf{D} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{26}$$

hence

$$x = \mathsf{C}\,z \quad \text{with} \quad \mathsf{C} = \mathsf{D}^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & & \ddots & \\ 1 & \dots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{27}$$

Therefore our inverse problem is to estimate the vector z, assumed to be sparse, from the data vector b, given the forward model

$$b = \mathsf{AC}z + e, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathsf{I}), \quad \mathsf{A}_{jk} = w_k A(s_j - t_k). \tag{28}$$

To illustrate how the sparsity of the MAP estimate determined by the IAS algorithm is affected by the choice of hyperprior in the generalized gamma family, we show the results with the hyperpriors corresponding to $r = 3$, $r = 1$ and $r = 0.5$, see Figure 7. The results clearly demonstrates that with decreasing r, the sparsifying properties are strengthened. Observe that the dramatic decrease of the CGLS iterations, compared to the numerical rank of the matrix, makes the approximate IAS very attractive for large problems.

6.2. Example 2

In the second example, we consider the problem of estimating a nearly black two-dimensional object. The generating model is an impulse image, defined as a distribution on $\Omega =$

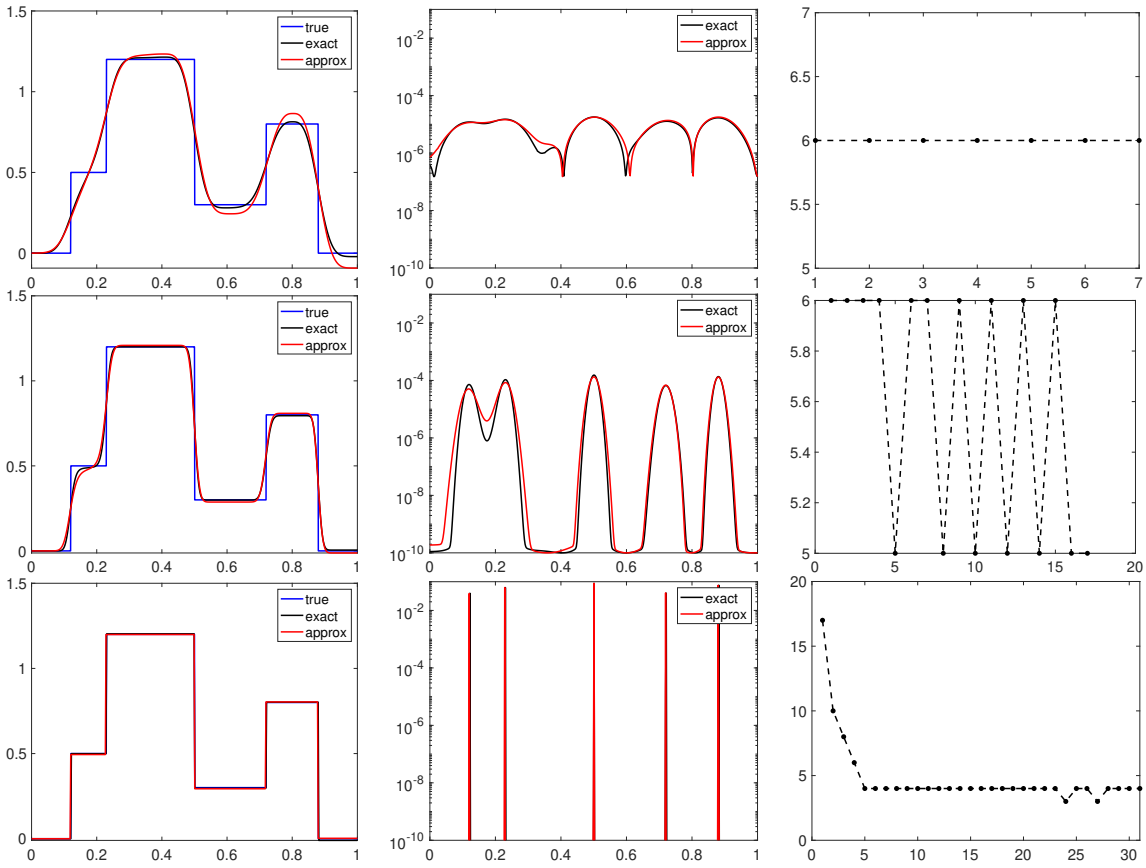

Figure 7: Reconstructions of the signal x (left), the hyperparameter θ (center) and the count of CGLS iterations per each IAS update when the approximate method is employed. In the top row, the parameter values are $r = 3$ and $\eta = 10^{-5}$, in the middle row, $r = 1$ and $\eta = 10^{-5}$, and in the bottom row $r = 1/2$ and $\eta = 10^{-5}$. The results with both the exact and approximate IAS are shown.

$[0, 1] \times [0, 1]$,

$$d\mu(p) = \sum_{k=1}^{J} a_k \delta(p - p_k) dp, \quad p_k \sim \text{Uniform}(\Omega), \quad a_k \sim \text{Uniform}([1.5, 2]),$$

and we assume that the distribution is observed with a Gaussian convolution kernel,

$$A(p, p') = \frac{1}{2\pi w^2} e^{-\|p-p'\|^2/2w^2}, \quad w = 0.01,$$

the discrete an noisy data being given at observation points $q_j \in \Omega$ by

$$b_j = \int_\Omega A(q_j, p') d\mu(p') + \varepsilon_j = \sum_{k=1}^{K} a_k A(q_j, p_k) + \varepsilon_j.$$

To solve the inverse problem, we divide the image Ω in $n = 128 \times 128 = 16\,384$ pixels,

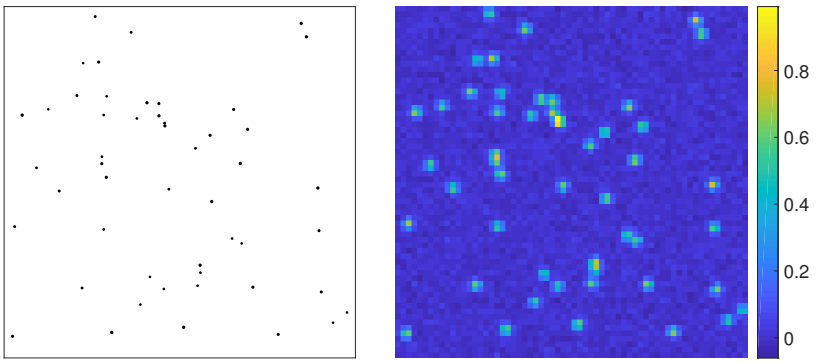

Figure 8: Left panel: The generative model, an impulse image of 50 point sources with variable amplitude. Right panel: The 64×64 blurred and noisy observation, degraded by Gaussian blur and additive while Gaussian noise, scaled so as to achieve SNR ≈ 25, corresponding to a standard deviation of about 1.8% of the maximum noiseless signal.

denoted by Ω_ℓ, and discretize the kernel, approximating

$$\int_\Omega A(q_j, p) d\mu(p) \approx \sum_{\ell=1}^{n} \underbrace{|\Omega_\ell| A(q_j, q'_\ell)}_{=\mathsf{A}_{j\ell}} x_\ell, \quad x_\ell = \frac{1}{|\Omega_\ell|} \int_{\Omega_\ell} d\mu(p),$$

where q'_ℓ denotes the center point of the pixel Ω_ℓ and $|\Omega_\ell|$ is its area. In this example, we assume that the number of observation points is $m = 64 \times 64 = 4\,096$, hence the forward model is defined by a matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$. The noiseless signal is then corrupted by scaled white noise with standard deviation approximately 1.8% of the maximum noiseless signal. In this case, since the signal itself is sparse, no change of variable is needed. Figure 8 shows the positions of the point masses in the true impulse image, as well as the noisy blurred image with kernel width $w = 0.01$.

We consider three hyperpriors from the generalized gamma family, corresponding to $r = 1$, $r = 0.5$, and $r = -1$. In this example we do not assume non-negativity, hence no projection is performed. To promote sparsity in the first two cases se set $\eta = 10^{-5}$, while in the third case, where $r = -1$ and η does not have the same role as for positive values of r, we set $\beta = 3$. We scale the hyperparameters by a constant value, setting $\vartheta_j = \vartheta_0 = \text{constant}$, and to make the results comparable, we select the parameter ϑ_0 so that the lower bound for the scaling parameters θ_j are equal,

$$\vartheta_0 \varphi(0) = \vartheta_0 \left(\frac{\eta}{r}\right)^{1/r} = 10^{-9}.$$

In this example, we consider only the approximate IAS algorithm.

The final reconstructions, shown in Figure 9 are almost identical, and the number of iterations are comparable. The number of the CGLS inner iterations per outer iteration in each case is low, no more than 15. To see a difference in the performance for the three parameter choices, we show how the reconstruction of the hyperparameter θ proceeds. Figure 11 shows

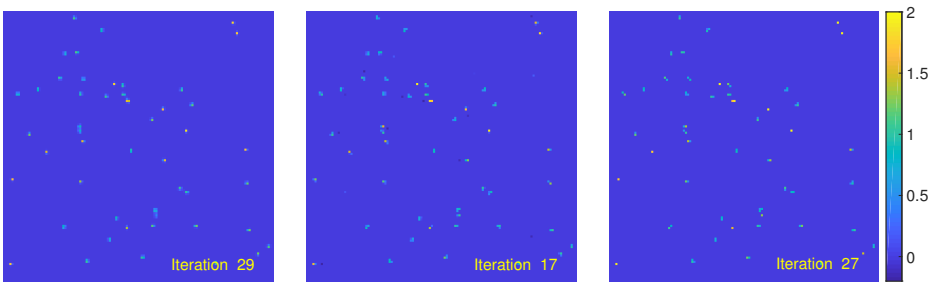

Figure 9: Reconstructions of the impulse image from blurred noisy observation. The reconstructed image is of size 128×128, and the hyperparameter values are, from left to right: $(r, \eta) = (1, 10^{-5})$, $(r, \eta) = (1/2, 10^{-5})$, and $(r, \beta) = (-1, 3)$. The images are in the same scale.

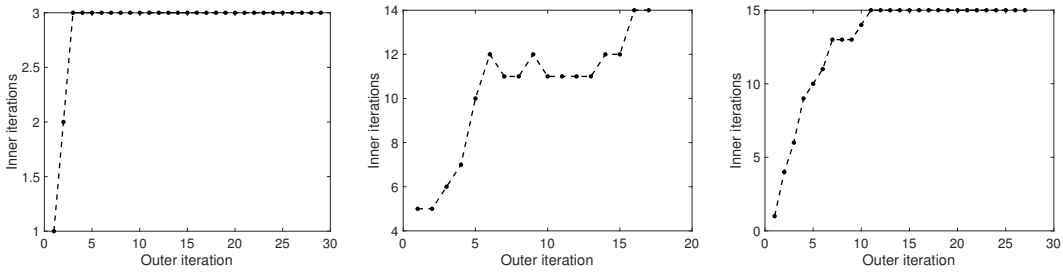

Figure 10: The number of CGLS iterations in each outer iteration. The hyperparameter values correspond to those in Figure 9.

logarithmic plots of the vector θ, rendered as a pixel image, after 2,4,8, and 16 iterations. The first observation is that even if the lower bound for the parameter vector θ was set equal, each choice of hyperprior leads to a different scale of values of θ: The value $r = 0.5$ yields the lowest values, while in the case $r = -1$ the interval is shifted to considerably higher values. Interestingly, however, the ratio between the largest and smallest value is in the same range. This observation is important, as the ratio informs us about the relative weights of each column of A in the scaling $\mathsf{A} \to \mathsf{A}\mathsf{D}_\theta^{1/2}$. The column scaling performs an effective model reduction, identifying the relevant columns of A and suppressing irrelevant ones. Each hyperparameter selection in the end identifies the same relevant columns, however the plots in Figure 11 show that the choice $r = 1$ is the most conservative, while when $r = -1$ the suppression of irrelevant columns happens sooner. Therefore one can argue that the parameter choices that correspond to less convex case pursue more greedily the support, however, the lack of convexity also makes it possible that the support corresponds to a local, rather than global minimum.

6.3. Example 3

The third example demonstrates the inclusion of bound constraints in the IAS algorithm. We consider a limited angle tomography inverse problem under the assumption that the generative

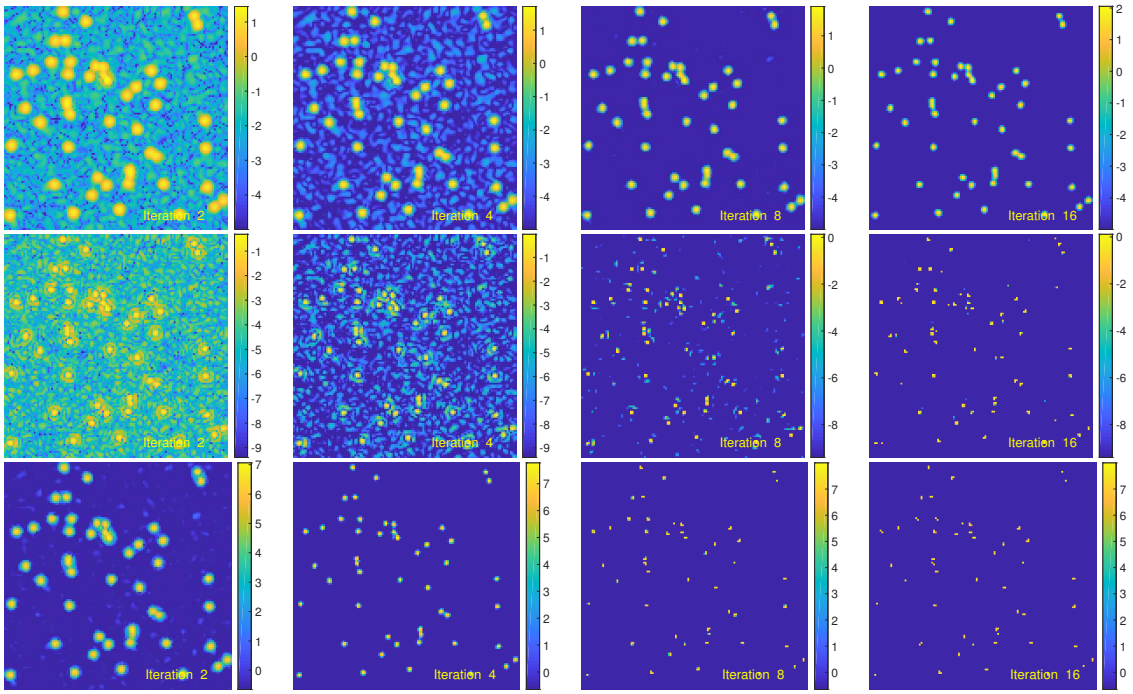

Figure 11: Logarithmic plot, top to bottom, of the estimate of θ for the three hyperpriors corresponding to Figure 9 at the end of the outer iteration 2, 4, 8 and 16 (left to right). Observe that the focusing effect corresponding to effective model reduction is fastest with $r = -1$ (bottom row) and slowest with $r = 1$ (top row).

density, shown in Figure 12, is piecewise constant, and the data consist of parallel beam sinogram data corresponding to illumination angles from -40° to 40° around the vertical illumination direction. More precisely, given a density function $\rho \geq 0$, the noiseless data can be written as

$$b_{jk} = \int_0^\infty \rho(p_j + tw_k)dt, \quad p_j = (s_j, 0), \quad w_k = (\sin\phi_k, \cos\phi_k), \tag{29}$$

where the values s_j, $1 \leq j \leq 100$ are uniformly distributed over the interval $[-0.4, 1.4]$, while the angles ϕ_k, $1 \leq k \leq 80$, cover uniformly the angle $[-40^\circ, 40^\circ]$, this constituting data of dimension $m = 8\,000$. We generate the noiseless data by formula (29), by finding the lengths of the line segments intersecting the two inclusions shown in Figure 12, and corrupt it with additive scaled white noise with standard deviation σ equal to 1% of the maximum of the noiseless data. To solve the inverse problem, we divide the image area $\Omega = [0, 1] \times [0, 0.5]$ into $n = n_x \times n_y = 200 \times 100 = 20\,000$ square pixels, denoted by Ω_i, approximating the conductivity by a piecewise constant density,

$$\int_0^\infty \rho(p_j + tw_k)dt \approx \sum_{i=1}^{n} |\ell_{jk} \cap \Omega_i| x_i,$$

where ℓ_{jk} is the line parametrized by intercept s_j and angle ϕ_k, $|\ell_{jk} \cap \Omega_i|$ is the length of its

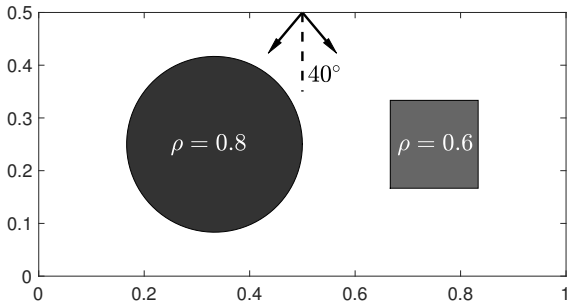

Figure 12: The underlying generative model with two inclusions in a transparent background $\rho = 0$. The sinogram data are computed by computing the lengths of the intersections of the lines with the two inclusions.

intersection with the pixel Ω_i, and x_i is the ith pixel value of the density. By enumerating the rays, the above formula defines the linear model with a sparse matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$.

The true target is not sparse in the pixel basis, but as in [9], we seek to represent it in terms of vertical and horizontal increments. Assuming that the density ρ is represented as an image, and the pixels are enumerated column-wise, the vertical and horizontal increments, v and h, of the image can be computed as

$$v = \underbrace{(\mathsf{I}_{n_x} \otimes \mathsf{D}_{n_y})}_{=\mathscr{D}_v} x, \quad h = \underbrace{(\mathsf{D}_{n_x} \otimes \mathsf{I}_{n_y})}_{=\mathscr{D}_h} x,$$

where $D_{n_x} \in \mathbb{R}^{n_x \times n_x}$ and $D_{n_y} \in \mathbb{R}^{n_y \times n_y}$ are the first order finite difference matrices (26) of the respective sizes, I_{n_y} and I_{n_x} are the identity matrices, and $\otimes$ is the Kronecker product.

As in [9], we write the conditional prior for the pair $(v, h) \in \mathbb{R}^{2n}$ as

$$\pi(v, h \mid \theta) = \frac{1}{(2\pi)^n \theta_1 \cdots \theta_n} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \frac{v_j^2 + h_j^2}{\theta_j} \right),$$

that is, rather than treating the vertical and horizontal increments as independent, the prior is written for the length of the gradient. Observe that unlike in Example 1, in this case the mapping between x and $x = (v, h)$ is not bijective. Moreover, the prior can be interpreted to promote group sparsity, being thus slightly more general than the plain sparsity promoting prior discussed in the paper, however, we apply the alternating iteration with projection to the feasible set as described above. We refer to the cited article for details about how the likelihood density is interpreted.

We run the approximate IAS algorithm with the bound constraints, assuming that $0 \leq x_j \leq H$, where the upper bound in this case is chosen high enough so that in practice the projection is performed only on the positive cone. We test the algorithm using two hypermodels, corresponding to parameter values $(r, \eta) = (1, 10^{-4})$ and $(r, \eta) = (0.5, 10^{-4})$. The stopping criterion for the IAS algorithm is $\|\theta^{t+1} - \theta^t\|_\infty / \|\theta^t\|_\infty < \tau$, with the threshold value $\tau = 10^{-2}$.

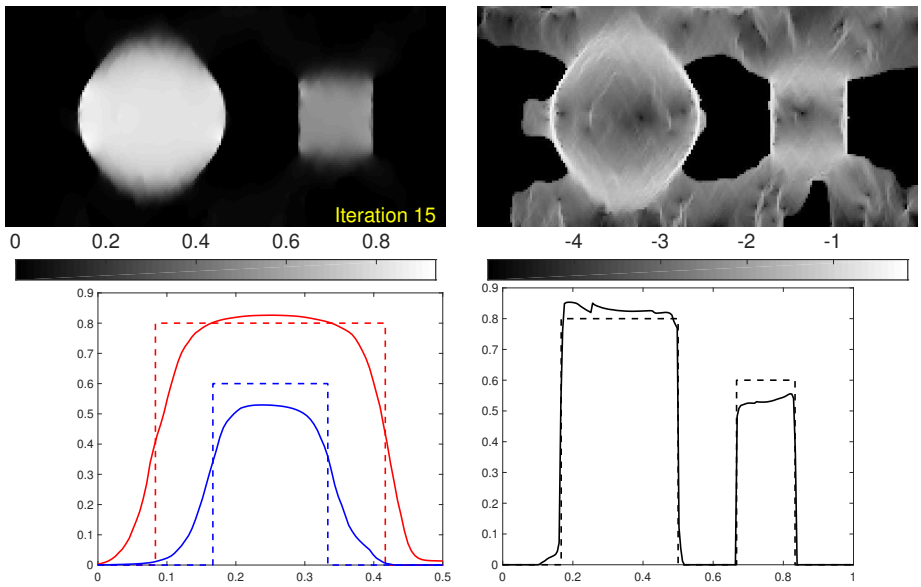

Figure 13: The limited angle tomography reconstruction with the hyperparameter values $(r, \eta) = (1, 10^{-5})$. The iterations converge in 15 iterations to the solution shown in the top row, left, and the corresponding scaled variance ξ, rendered as a pixel image, is shown in a logarithmic plot on the right. The vertical intersection through the two inclusion are shown in the bottom row, left, and the horizontal intersection on the right. We observe that the hyperprior is not focalizing enough to suggest sharp edges at the horizontal boundaries where the data do not support the jump.

Figure 13 shows the final reconstructions. The results confirm that the hyperprior $r = 1/2$ produces a sharper edge of the target objects on the horizontal edge where the data are not able to distinguish between a sharp and diffuse edge, while the hyperprior $r = 1$ is less committal. In this example, the approximate IAS algorithm with projection converges rather fast for $r = 1$ (15 iterations), while the convergence in the case $r = 0.5$ requires more iterations (153). However, as in the previous example, even after a few iterations, the latter hypermodel has already identified the support of the discontinuities; the intermediate results are not shown here. As in the previous example, the large dynamical range of ξ, around 9 orders of magnitude, provide a very efficient column reduction of the matrix AD_θ.

7. Conclusions

The present article discusses conditionally Gaussian hypermodels and the IAS algorithm, extending the previous analysis to a larger class of hyperpriors, and investigates the asymptotic behavior of the resulting priors, as well as the convexity of the optimization problem for finding the MAP estimate. The algorithm is modified to include simple bound constraints, and an emphasis is given to an approximate method of solving the imbedded least squares problem using Krylov subspace methods. The computed examples confirm that the further

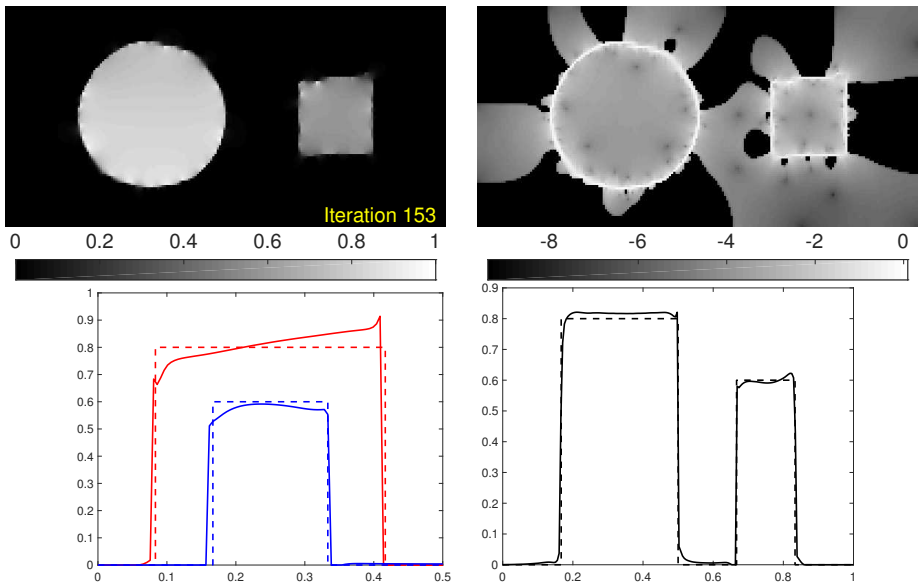

Figure 14: Reconstructions from the same data as in Figure 13 with parameter values $(r, \eta) = (0.5, 10^{-5})$. The number of iterations before convergence is higher than in the previous case, however, the contrast in ξ is orders of magnitudes larger than in the previous case. As in Example 2, even after few iterations, the algorithm is already identifying well the support of the discontinuities (not shown). Also, as the hyperprior is more sparsifying, the algorithm suggests sharp jumps across the horizontal boundaries of the inclusions.

away from the convexity conditions we are, the more pronounced the identification of the support of the sparse variable is, with the caveat that in lack of convexity, there is no guarantee of a single global minimum of the energy functional. Thus, the algorithm may get trapped in a local minimum. The question of how to effectively take advantage of the convexity properties requires further tools and will be a topic of a forthcoming work. The examples also show that the proposed methods for sparse recovery can be an efficient model reduction scheme. The latter property will be further analyzed in a forthcoming article.

The discussion in this article was limited to finite dimensional discretized inverse problems. While sparsity in infinite dimensions is not a well-defined concept, a significant body of work exists, with the interpretation that a prior is sparsity promoting if the finite dimensional truncation in a given basis is sparsity promoting, giving rise, e.g., to Besov space priors with wavelet bases, see, e.g., [24, 15, 25, 1]. The interpretation and extension of the hierarchical model discussed in this paper to infinite dimensional spaces remain an open problem.

Acknowledgements

The work of DC and AS was partly supported by NSF grant DMS-1522334, and of ES by NSF grant DMS-1714617. Part of this work was done while the authors DC, MP and ES were visiting Institut Henri Poincaré during the worshop *The Mathematics of Imaging* in March-

April 2019. The hospitality and support of IHP is gratefully acknowledged.

References

[1] Agapiou S, Burger M, Dashti M and Helin T (2018) Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. Inverse Problems **34**: p.045002.

[2] Aguerrebere C, Almansa A, Delon J, Gousseau Y and Musé P (2017) A Bayesian hyperprior approach for joint image denoising and interpolation, with an application to HDR imaging. IEEE Trans Comput Imaging **3**: 633-646.

[3] Babacan SD, Mancera L, Molina R and Katsaggelos AK (2009) Non-convex priors in Bayesian compressed sensing. In 2009 17th European Signal Processing Conference (August, pp. 110-114). IEEE.

[4] Bangti J, Mass P and Scherzer O (2017) Sparsity regularization in inverse problems. Inverse Problems **33**: 060301

[5] Candes EJ, Romberg JK and Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Comm Pure Appl Math **59**: 1207-1223.

[6] Calvetti D (2007) Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. J Comp Appl Math **198**: 378-395.

[7] Calvetti D and Somersalo E (2005) Priorconditioners for linear systems. Inverse Problems **21**: 1397.

[8] Calvetti D, Hakula H, Pursiainen S and Somersalo E (2009) Conditionally Gaussian hypermodels for cerebral source localization. SIAM J Imaging Sci **2** 879–909.

[9] Calvetti D and Somersalo E (2008) Hypermodels in the Bayesian imaging framework. Inverse Problems **24** 034013.

[10] Calvetti D, Somersalo E and Strang A (2019) Hierarchical Bayesian models and sparsity: ℓ_2-magic. Inverse Problems **35**: 035003 (26pp).

[11] 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.

[12] Calvetti D, Pitolli F, Somersalo E and Vantaggi B (2018) Bayes meets Krylov: Statistically inspired preconditioners for CGLS. SIAM Review, **60**: 429-461.

[13] Calvetti D, Pitolli F, Prezioso J, Somersalo E and Vantaggi B (2017) Priorconditioned CGLS-based quasi-MAP estimate, statistical stopping rule, and ranking of priors. SIAM J Sci Computing **39**: S477-S500.

[14] Calvetti, D, Pascarella A, Pitolli F, Somersalo E. and Vantaggi B (2019) Brain Activity Mapping from MEG Data via a Hierarchical Bayesian Algorithm with Automatic Depth Weighting. Brain Topography **32**: 363-393.

[15] Daubechies I, Defriese M and De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraints. Comm Pure Appl Math **57**:1413 - 57.

[16] Daubechies I, Defriese M and De Mol C (2016) Sparsity-enforcing regularization and ISTA revisited. Inverse Problems **32**: 104004.

[17] Daubechies I, DeVore R, Fornasier M. and Güntürk CS (2010) Iteratively reweighted least squares minimization for sparse recovery. Comm Pure Appl Math **63**: 1-38.

[18] Dobson DC and Vogel CR (1997) Convergence of an iterative method for total variation denoising. SIAM J Num Anal **34**: 1779-1791.

[19] D Donoho (2006) For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Commun. Pure and Appl. math. **59** 797 - 829.

[20] Durmus A, Moulines E and Pereyra M (2018). Efficient bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. SIAM J Imag Sci **11**: 473-506.

[21] Donoho DL and Elad M (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via L1 minimization. Proc Natl Acad Sci **100** : 21972202.

[22] Gorodnitsky IF and Rao BD (1997) Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. IEEE Trans Signal Proc **45**: 600-616.

[23] Hestenes MR and Stiefel E (1952) Methods of conjugate gradients for solving linear systems (Vol. 49, pp. 409–436). Washington, DC: National Bureau of Standards.

[24] Johnstone IM (1994) Minimax Bayes, asymptotic minimax and sparse wavelet priors. In: *Statistical Decision Theory and Related Topics V* (pp. 303-326). Springer, New York, NY.

[25] Kolehmainen V, Lassas M, Niinimäki K and Siltanen S (2012) Sparsity-promoting Bayesian inversion. Inverse Problems **28**: 025005.

[26] Moreau J-J (1962) Fonctions convexes duales et points proximaux dans un espace Hilbertien. C R Acad Sci Paris Sér A Math **255**:2897-2899.

[27] Natarajan BK (1995) Sparse approximate solutions to linear systems. SIAM J Comput **24**: 227-234.

[28] Pereyra M (2016). Proximal Markov chain Monte Carlo algorithms. Statistics and Computing **26**: 745-760.

[29] Taylor HL, Banks SC and McCoy JF (1979) Deconvolution with the ℓ^1 norm. Geophysics **44** 39-52.

[30] Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc (Ser B) **58**: 267-288.

[31] Welch LR (1974) Lower bounds on the maximum cross correlation of signals. IEEE Trans Inform Theory **20**: 397399.