SIAM/ASA J. Uncertainty Quantification
Vol. 12, No. 2, pp. 524–548

Computationally Efficient Sampling Methods for Sparsity Promoting Hierarchical Bayesian Models*

D. Calvetti† and E. Somersalo†

Abstract. Bayesian hierarchical models have been demonstrated to provide efficient algorithms for finding sparse solutions to ill-posed inverse problems. The models comprise typically a conditionally Gaussian prior model for the unknown, augmented by a hyperprior model for the variances. A widely used choice for the hyperprior is a member of the family of generalized gamma distributions. Most of the work in the literature has concentrated on numerical approximation of the maximum a posteriori estimates, and less attention has been paid on sampling methods or other means for uncertainty quantification. Sampling from the hierarchical models is challenging mainly for two reasons: The hierarchical models are typically high dimensional, thus suffering from the curse of dimensionality, and the strong correlation between the unknown of interest and its variance can make sampling rather inefficient. This work addresses mainly the first one of these obstacles. By using a novel reparametrization, it is shown how the posterior distribution can be transformed into one dominated by a Gaussian white noise, allowing sampling by using the preconditioned Crank–Nicholson (pCN) scheme that has been shown to be efficient for sampling from distributions dominated by a Gaussian component. Furthermore, a novel idea for speeding up the pCN in a special case is developed, and the question of how strongly the hierarchical models are concentrated on sparse solutions is addressed in light of a computed example.

Key words. preconditioned Crank–Nicholson, Markov chain Monte Carlo, noncentered sampling

MSC codes. 65C05, 65C60

DOI. 10.1137/23M1564043

1. Introduction.

A common problem in computational inverse problems is the estimation of an unknown quantity that is a priori believed to be sparse, in the sense that it can be represented with only very few elements of a given basis or frame. In many cases, in particular when the solution is computed numerically, sparsity is replaced by compressibility, meaning that most of the coefficients in the representation are below a small threshold value. The concepts of sparsity and compressibility are particularly important in the framework of compressed sensing [17] and sparse dictionary learning [3]. From the very definition of sparsity and compressibility, it is clear that these characterizations are qualitative in nature, as "most of the coefficients" is to some extent arbitrary, depends on the dimensionality of the problem, and is open to subjective interpretations. Standard methods for finding sparse solutions

*Received by the editors April 5, 2023; accepted for publication (in revised form) February 7, 2024; published electronically June 7, 2024.
https://doi.org/10.1137/23M1564043
Funding: The work of the first author was partly supported by the NSF grant DMS 1951446, and that of the second author by the NSF grant DMS-2204618. The support of John Simon Guggenheim Memorial Foundation for the work of the second author is acknowledged.

†Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, Cleveland, OH 44106 USA (dxc57@case.edu, ejs49@case.edu).

include the introduction of sparsity promoting penalties, the most popular being the ℓ^1-penalty. Thus, if the data $b \in \mathbb{R}^m$ are related to the vector $x \in \mathbb{R}^n$ whose entries are the coefficients of the representation of the unknown in a given basis or frame by $b = f(x) + \text{noise}$, the standard ℓ^1-penalized least squares solution is a minimizer of the functional

$$F_\alpha(x) = \|b - f(x)\|^2 + \alpha\|x\|_1, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidian norm in $\mathbb{R}^m$, $\|\cdot\|_1$ the ℓ^1-norm in $\mathbb{R}^n$, and $\alpha > 0$ is a regularization parameter. The problem is referred to as basis pursuit [13, 12], or LASSO [30], depending on the context. For sparsity and wavelet techniques, we refer to [27]. The existence and uniqueness of such solutions depend on the properties of the forward map $f : \mathbb{R}^n \to \mathbb{R}^m$. An alternative, but closely related, approach to sparsity is rooted in the Bayesian analysis of inverse problems [11]. In the Bayesian framework, the prior belief of the sparsity of the signal is encoded into a prior that favors sparse solutions, and single point estimates such as the maximum a posteriori (MAP) or posterior mean estimates are generated in the hope that they have the desired sparsity properties [10, 7]. In this article, we restrict our attention to a particular family of Bayesian sparsity promoting priors, namely, hierarchical Gaussian priors augmented with a hyperprior from the family of generalized gamma distributions [8], reviewed in section 2. In that section, we also give a brief review of an algorithm to compute the MAP estimates, the iterative alternating sequential (IAS) algorithm, which has been shown to generate sparse, or compressible solutions at a relatively small computational cost. Under certain conditions, the IAS algorithm has been demonstrated to converge to a unique solution that approximates the ℓ^1-penalized solution [7]. The convergence properties are leveraged in a hybrid algorithm that combines different choices of generalized gamma priors [9]. For an alternative but related way of estimating the hyperparameters based on data, we refer to [5, 31].

A pertinent question concerning the IAS solution or other MAP estimates and, in general, algorithms searching for a single estimate, is how representative the sparse solution is. A common criticism of the MAP estimate is that it may capture poorly the posterior distribution and could be unstable with respect to perturbations in the data, in particular in the case when the posterior density is multimodal, whereas the posterior mean estimate, when calculated by means of Monte Carlo sampling from the posterior represents a more reliable alternative as a single point estimator. The question is closely related to the wider question of uncertainty quantification under sparsity constraints, often addressed by Markov chain Monte Carlo (MCMC) sampling [21, 26]. While the use of MCMC methods to explore posterior densities is rather standard, hierarchical sparsity promoting models are known to pose significant challenges to sampling methods. The problems are twofold: The problems are typically high dimensional, and the typically strong correlation between the unknown of primary interest and the hyperparameters leads to poor mixing and extremely slow convergence of the samplers. Remedies to the latter problem have been proposed in the literature, including appropriate changes of variables; see, e.g. [2, 4, 15, 18, 28, 29]. For contributions to quantify the uncertainties in the inverse problems based on hierarchical prior models, we refer to [2, 1, 32, 19].

In this article, we propose changes of variables specifically tailored for the hierarchical Bayesian models that the IAS algorithm is based on. The main goal of this study is to address

the problems arising from the high dimensionality of the problem. The change of variables that we propose, combined with the preconditioned Crank–Nicholson (pCN) sampling algorithm [14], leads to an easy to implement sampling algorithm that is fast and relatively easy to tune. The algorithm provides an efficient sampling of the posterior, in particular when the hypermodel is based on a gamma distribution or generalized gamma distributions near it. Hypermodels with strong nonconvexity remain a challenge, although for the inverse gamma hyperprior, we propose a particular radial parametrization generalizing the pCN sampler that improves the convergence significantly. Another novel contribution of this work is an automatic way to choose the hypermodel parameters in the hybrid IAS algorithm, thus completing the work in [9]. The computed examples using this sampler to explore posterior with generalized gamma hyperpriors seem to suggest a rather surprising result, namely, that while the MAP estimate is itself very definitely compressible, neither the Monte Carlo samples around it nor the corresponding posterior mean are necessarily sparse or compressible unless the hypermodel is chosen to be strongly nonconvex. Therefore, one can argue that the MAP estimates based on hierarchical Gaussian models are optimal when capturing the sparse nature of the unknown is important.

2. Hierarchical models and sparsity. Consider the inverse problem of estimating an unknown $x \in \mathbb{R}^n$ from noisy observations of a linear transformation of it,

$$b = f(x) + e,$$

where $f : \mathbb{R}^n \to \mathbb{R}^m$ is assumed to be a known function. In the Bayesian framework, all parameters not known exactly are modeled as random variables. In the rest of the paper, we denote random variables by capital letters, and their realizations by lowercase letters. The stochastic extension of the observation model is therefore

$$B = f(X) + E, \tag{2}$$

where it is assumed that X and E are mutually independent random variables. If the probability distribution of the noise E is given in terms of a probability density function π_E, the likelihood model for B is

$$\pi_{B|X}(b \,|\, x) = \pi_E(b - f(x)).$$

If we assume that E is a zero mean Gaussian noise with positive definite covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$, we obtain the likelihood model

$$B \,|\, X \sim \mathcal{N}(f(x), \Sigma)$$

or, in terms of the probability densities,

$$\pi_{B|X}(b \,|\, x) \propto \exp\left(-\frac{1}{2}(b - f(x))^{\mathsf{T}}\Sigma^{-1}(b - f(x))\right).$$

Furthermore, by a standard whitening argument, multiplying both b and f by a symmetric factor S of the precision matrix, $\Sigma^{-1} = \mathsf{S}^{\mathsf{T}}\mathsf{S}$, we may assume without loss of generality that $\Sigma = \mathsf{I}_m$, the $m \times m$ identity matrix, and the likelihood model simplifies to

$$\pi_{B|X}(b \,|\, x) \propto \exp\left(-\frac{1}{2}\|b - f(x)\|^2\right). \tag{3}$$

To express the sparsity or compressibility belief, we consider a conditionally Gaussian prior model,

$$X_j \,|\, \Theta_j \sim \mathcal{N}(0, \theta_j), \quad 1 \le j \le n \quad \text{(mutually independent)},$$

with probability density

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

Furthermore, we assume that the variances Θ_j are mutually independent and distributed according to the generalized gamma distribution,

$$\Theta_j \sim \mathrm{GenGamma}(r, \beta, \vartheta_j) \quad 1 \le j \le n \quad \text{(mutually independent)},$$

with densities

$$\Theta_j \sim \pi_{\Theta_j}(\theta_j \,|\, \vartheta_j, \beta, r) = \frac{|r|}{\Gamma(\beta)\vartheta_j}\left(\frac{\theta_j}{\vartheta_j}\right)^{r\beta-1} \exp\left(-\left(\frac{\theta_j}{\vartheta_j}\right)^r\right), \quad 1 \le j \le n,$$

where $r \neq 0$ and the shape parameter $\beta > 0$ is assumed to be the same for all j while the scale parameter ϑ_j may differ for every j. Taking into account the mutual independency of the variances, we can write the joint prior model in the form

$$\begin{aligned}\pi_{X,\Theta}(x,\theta) &= \pi_{X|\Theta}(x \,|\, \theta)\pi_\Theta(\theta \,|\, \vartheta, \beta, r) \\ &\propto \exp\left(-\frac{1}{2}\sum_{j=1}^n \frac{x_j^2}{\theta_j} - \sum_{j=1}^n \left(\frac{\theta_j}{\vartheta_j}\right)^r + \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log\frac{\theta_j}{\vartheta_j}\right).\end{aligned} \tag{4}$$

Next we proceed by *nondimensionalizing* the model. Introduce the nondimensional variables

$$\xi_j = \frac{x_j}{\sqrt{\vartheta_j}}, \quad \lambda_j = \frac{\theta_j}{\vartheta_j}, \tag{5}$$

and observe that after the change of variables, the components (ξ_j, λ_j) are a priori independent and identically distributed,

$$(\xi_j, \lambda_j) \sim \exp\left(-\frac{1}{2}\frac{\xi_j^2}{\lambda_j} - \lambda_j^r + \left(r\beta - \frac{3}{2}\right)\log\lambda_j\right),$$

and the posterior density can be written as

$$\pi_{\Xi,\Lambda|B}(\xi, \lambda \,|\, b) \propto \exp\left(-\frac{1}{2}\|b - f(\mathsf{D}_\vartheta^{1/2}\xi)\|^2 - \frac{1}{2}\sum_{j=1}^n \frac{\xi_j^2}{\lambda_j} - \sum_{j=1}^n \lambda_j^r + \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log\lambda_j\right), \tag{6}$$

where $\mathsf{D}_\vartheta \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the vector ϑ along its diagonal. Observe that in the case of a linear forward model, $f(x) = \mathsf{A}x$, where $\mathsf{A} \in \mathbb{R}^{m\times n}$, we have

$$f(\mathsf{D}_\vartheta^{1/2}\xi) = \mathsf{A}\mathsf{D}_\vartheta^{1/2}\xi, \tag{7}$$

where the transformation $\mathsf{A} \mapsto \mathsf{A}\mathsf{D}_\vartheta^{1/2}$ is a column scaling of the forward map. In [6, 7], it was demonstrated that this scaling, which is part of the prior, can be associated with the sensitivity of the unknowns x_j to the data. This notion will be further elaborated in the next section.

3. Exploring the posterior density.

In this section, after a brief review of the IAS algorithm for estimating the MAP estimate, we proceed to discuss the exploration of the posterior density using MCMC methods.

3.1. MAP estimate and the IAS algorithm.

The MAP estimate corresponding to the posterior density (6) is, by definition,

$$(\xi_{\mathrm{MAP}}, \lambda_{\mathrm{MAP}}) = \mathrm{argmax}\{\pi_{\Xi,\Lambda}(\xi, \lambda \,|\, b)\},$$

provided that such a maximizer exists. Equivalently, the MAP estimator, if it exists, is also a minimizer of the Gibbs energy:

$$\mathscr{E}(\xi, \lambda) = \frac{1}{2}\|b - f(\mathsf{D}_\vartheta^{1/2}\xi)\|^2 + \frac{1}{2}\sum_{j=1}^n \frac{\xi_j^2}{\lambda_j} + \sum_{j=1}^n \lambda_j^r - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log\lambda_j. \tag{8}$$

The IAS algorithm searches for the MAP estimate through the alternating steps as follows. Given an initial λ^0, set the counter at $t = 1$, and iterate the steps until a convergence criterion is met:

(a) Update ξ by defining

$$\xi^t = \mathrm{argmin}\{\mathscr{E}(\xi, \lambda^{t-1})\} = \mathrm{argmin}\left\{\frac{1}{2}\|b - f(\mathsf{D}_\vartheta^{1/2}\xi)\|^2 + \frac{1}{2}\sum_{j=1}^n \frac{\xi_j^2}{\lambda_j^{t-1}}\right\}.$$

(b) Update λ by defining

$$\lambda^t = \mathrm{argmin}\{\mathscr{E}(\xi^t, \lambda)\} = \mathrm{argmin}\left\{\frac{1}{2}\sum_{j=1}^n \frac{(\xi_j^t)^2}{\lambda_j} + \sum_{j=1}^n \lambda_j^r - \left(r\beta - \frac{3}{2}\right)\sum_{j=1}^n \log\lambda_j\right\}.$$

(c) Advance the counter, $t \to t+1$ and check for convergence.

Observe that if the forward model is linear, the first step is a standard least squares problem, and regardless of the linear model, the second step is a componentwise updating problem requiring the solution of the first order optimality condition

$$\frac{\partial}{\partial\lambda_j}\left(\frac{(\xi_j^t)^2}{\lambda_j} + \lambda_j^r - \left(r\beta - \frac{3}{2}\right)\log\lambda_j\right) = 0. \tag{9}$$

For certain values of r, (9) admits a closed form solution. When $r = 1$ and $\eta = \beta - 3/2 > 0$, corresponding to gamma hyperpriors, the updating formula for λ^t becomes

$$\lambda_j^t = \frac{1}{2}\left(\eta + \sqrt{\eta^2 + 2(\xi_j^t)^2}\right), \quad \eta = \beta - \frac{3}{2},$$

while, when $r = -1$, corresponding to inverse gamma hyperpriors, we have

$$\lambda_j^t = \frac{1}{\kappa}\left(\frac{(\xi_j^t)^2}{2} + 1\right), \quad \kappa = \beta + \frac{3}{2}. \tag{10}$$

For choices of r that do not admit closed form solutions, the values λ_j^t satisfy

$$\lambda_j^t = \varphi(|\xi_j^t|),$$

where the function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ solves the nonlinear initial value problem

$$\varphi'(t) = \frac{2t\varphi(t)}{2r^2\varphi(t)^{r+1} + t^2}, \quad \varphi(0) = \left(\beta - \frac{3}{2r}\right)^{1/r}, \tag{11}$$

obtained from implicit differentiation of (9). The evaluation of the function φ at the points $|\xi_j^t|$ can be done efficiently by sorting the values $|\xi_j^t|$ in ascending order and solving the initial value problem by any standard numerical integrator at those points. For details, we refer to [8].

The existence and uniqueness of the MAP estimate, as well as the convergence of the IAS algorithm are not obvious. The following result has been proved in [7] for the case when $r = 1$, $\beta > 3/2$, and A is linear.

Theorem 3.1. *Assume that the forward map is linear, $r = 1$, and $\eta = \beta - 3/2 > 0$. Then the Gibbs energy functional (8) has a unique minimizer (ξ^*, λ^*), and the IAS algorithm converges to that minimizer. Moreover, it holds that*

$$\lambda_j^* = \frac{1}{2}\left(\eta + \sqrt{\eta^2 + 2(\xi_j^*)^2}\right) = f_j(\xi_j^*),$$

and ξ_j^* is the unique minimizer of the functional

$$\widehat{\mathscr{E}}(\xi) = \mathscr{E}(\xi, f(\xi)).$$

Moreover,

$$\lim_{\eta \to 0+} \widehat{\mathcal{E}}(\xi) = \frac{1}{2}\|b - \mathsf{A}\mathsf{D}_\vartheta^{1/2}\xi\|^2 + \sqrt{2}\sum_{j=1}^n |\xi_j|, \tag{12}$$

and the IAS solution converges to the minimizer of the above right-hand side.

The above theorem underlines the role of the parameter η in promoting sparsity of the solution. For other hyperparameter values, in particular for $r < 1$, the uniqueness of the minimizer may not be guaranteed even in the linear case and, in fact, in some cases the algorithm is known to converge to local minima. The sparsity promoting nature of the prior

can often be understood by restricting the Gibbs energy functional to the manifold defined by the minimization condition (b) for λ. For instance, in the case of the inverse gamma model $r=-1$, defining $g_j(\xi_j)$ through the formula (10), we obtain

$$\mathscr{E}(\xi, g(\xi)) = \frac{1}{2}\|b - f(\mathsf{D}_\vartheta^{1/2}\xi)\|^2 + \sum_{j=1}^n \log\left(1+\frac{\xi_j^2}{2}\right)^\kappa + \text{constant}, \quad \kappa = \beta + \frac{3}{2}.$$

The minimizer of the Gibbs functional above is also the MAP estimate corresponding to a prior of the form

$$\pi_\Xi(\xi) \propto \prod_{j=1}^n \frac{1}{(1+\xi_j^2/2)^\kappa}, \tag{13}$$

which is strongly sparsity promoting. Observe that, setting $\beta > 0$ and $\kappa > 3/2$ to guarantee a finite expectation, at the limit as $\beta \to 0+$, $\kappa \to \frac{3}{2}+$, thus (13) converges to the Student distribution with $\nu = 2$; see [8] for details.

To clarify the role of the hyperparameter ϑ, consider first the regularization scheme (1), which can be interpreted as the MAP estimate corresponding to a Laplace prior,

$$\pi_X(x) \propto \exp\left(-\alpha\|x\|_1\right).$$

The above prior assumes a priori that the components are independent and equally distributed, which is a particular case of a more general assumption of exchangeability. Recall that random variables are called exchangeable if their joint probability density is invariant under permutations. While exchangeability is a good design principle for noninformative priors not favoring any particular component, this is not the case in many applications such as in geophysics [22, 23] and medical imaging [24]. If the data are more sensitive to some components than others, exchangeable priors favor solutions in which the data are primarily explained by the components with highest sensitivity. In subsurface imaging in geophysics and biomedicine, this means that the algorithm would bias the solution to mostly involve sources near the receivers, thus ignoring possible deep sources. A popular fix is to replace the ℓ^1-penalty term by a weighted version of it,

$$\sum_{j=1}^n |x_j| \to \sum_{j=1}^n w_j|x_j|, \tag{14}$$

with the weights w_j to be proportional to the sensitivity of the data with respect to the corresponding components. In the case of a linear forward model, this is tantamount to setting the jth weight,

$$w_j \propto s_j = \|\partial_{x_j}\mathsf{A}x\| = \|\mathsf{A}e_j\| = \|a^{(j)}\|, \tag{15}$$

equal to the norm of the jth column of the forward map. Weighting is a common procedure in optimization [16] for balancing the optimization problem. The scaling (15), however, not only violates the exchangeability condition, but it is also problematic in the Bayesian context,

because the prior is now informed by the measurement configuration. In particular, in the applications above, this means that the prior favors deep sources, which may not be a defendable prior belief. A tenable Bayesian argument that justifies the weighting was presented in [6, 7] along the reasoning presented below.

Given an observation model (2), the signal-to-noise ratio (SNR) is defined as

$$\mathrm{SNR} = \frac{\mathrm{E}(\|B\|^2)}{\mathrm{E}(\|E\|^2)} = \frac{\text{signal power}}{\text{noise power}}.$$

Furthermore, assume that the variable X is supported in a set $S \subset \{1, 2, \ldots, n\}$, that is, $X_j = 0$ for $j \notin S$. When restricting the support of X to S, we denote the corresponding SNR by SNR_S.

In the cited articles, the following concept, weaker than the exchangeability, was introduced.

Definition 3.2. *The random variable satisfies the SNR-exchangeability condition with respect to the observation model if*

$$SNR_S = SNR_{S'} \text{ for all subsets } S, S' \in \{1, 2, \ldots, n\} \text{ with the same cardinality.}$$

Exchangeability implies the weaker SNR-exchangebility, while the converse is not true. The following theorem, however, shows that while the SNR-exchangeability does not imply exchangeability, it does, at least in the case of a linear forward model, guarantee that each component of x has an equal chance to explain the data.

Theorem 3.3. *Assume that the random variable X satisfies the SNR-exchangeability condition with respect to the linear observation model (2) with (7), and the prior is the hierarchical prior (4). Then the scaling parameters must satisfy*

$$\vartheta_j = \frac{C}{\|a^{(j)}\|^2}, \quad C = C(r, \beta),$$

where $a^{(j)} \in \mathbb{R}^n$ is the jth column of the matrix A, and the constant $C(r, \beta)$ depends on the SNR.

In [8], an explicit formula for C is given. Observe that substituting $\xi_j = x_j/\sqrt{\vartheta_j}$ on the right-hand side of (12), the penalty term assumes the sensitivity-weighted form (14)–(15) as advocated in the literature, while the scaled version corresponds to a priori satisfying the exchangeability.

Finally, we address the question of nonconvexity of the functional when $r < 1$. In [9] a hybrid IAS algorithm was proposed and investigated. The idea behind the hybrid model is to first run the IAS algorithm by using the parameter value $r = 1$ that is guaranteed to converge in the linear case, then switching to a greedier scheme by choosing a generalized gamma hyperprior with $r < 1$. When the switching occurs, the two hypermodels are matched using the following two criteria:

1. Whenever $x_j = 0$, we require that the baseline values for θ_j coincide. This way, the a priori variance of the background outside the support of x is consistently defined independently of the model.
2. The marginal expected values for θ_j are equal using both models.

Let us denote by $(r_1, \beta_1, \varphi_1)$ and $(r_2, \beta_2, \varphi_2)$ the hyperparameter values for two models, where in practice $r_1 = 1$. From the initial condition in (11), we conclude that the compatibility condition implies that

$$\vartheta_1 \left(\beta_1 - \frac{3}{2r_1}\right)^{1/r_1} = \vartheta_2 \left(\beta_2 - \frac{3}{2r_2}\right)^{1/r_2}. \tag{16}$$

Recalling the expectation of a generalized gamma distribution, the second condition can be written as

$$\vartheta_1 \frac{\Gamma(\beta_1 + \frac{1}{r_1})}{\Gamma(\beta_1)} = \vartheta_2 \frac{\Gamma(\beta_2 + \frac{1}{r_2})}{\Gamma(\beta_2)}. \tag{17}$$

The assumed finiteness of the expectation poses restrictions to possible parameter values. In the section on computed examples, these conditions are discussed in detail for several special cases. We point out that the second condition was not considered in the cited paper, and it is introduced here for the first time to make the hyperparameter selection automatic.

In the following, we shall use the IAS algorithm, and the hybrid scheme in particular, to find an appropriate initial point for the MCMC sampling.

3.2. Sampling with a Gaussian prior: pCN. In preparation of the reparametrization of the hypermodels, we recall some known results concerning random draws from Gaussian distributions. Let $\mathsf{C} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Assuming that a symmetric factorization $\mathsf{C} = \mathsf{L}^{\mathsf{T}}\mathsf{L}$, such as the Cholesky factorization of the matrix, is available, independent random draws from the normal distribution $\mathcal{N}(x \,|\, 0, \mathsf{C})$ can be generated through the formula

$$X = \mathsf{L}^{\mathsf{T}} W, \quad W \sim \mathcal{N}(0, \mathsf{I}_n),$$

where I_n is the $n \times n$ identity matrix. The independent sampling, while generating a sequence X^j of independent draws from the distribution, does not provide any way to control the step size $\|X^j - X^{j-1}\|$. Step size control is fundamental when the Gaussian distribution is used as a proposal distribution for exploring posterior distributions. A way to enable step size control is to consider the sequence

$$X^j = \sqrt{1-h^2}\, X^{j-1} + h\mathsf{L}^{\mathsf{T}} W, \quad W \sim \mathcal{N}(0, \mathsf{I}_n), \quad X^0 \sim \mathcal{N}(0, \mathsf{C}),$$

where $0 < h < 1$. It is a straightforward matter to check by induction that X^j is a Gaussian zero mean random variable, with covariance equal to C; therefore, the produced sequence is distributed according to $\mathcal{N}(0, \mathsf{C})$. The parameter h controls the step size in the sense that

$$\begin{aligned} \mathrm{E}\|X^j - X^{j-1}\|^2 &= \mathrm{trace}(\mathrm{E}(X^j - X^{j-1})(X^j - X^{j-1})^{\mathsf{T}}) \\ &= 2(1 - \sqrt{1-h^2})\,\mathrm{trace}(\mathsf{C}) \approx h^2\,\mathrm{trace}(\mathsf{C}) \end{aligned}$$

for h small. The disadvantage, compared to independent sampling, corresponding to $h = 1$, is that consecutive samples are correlated, since

$$\mathrm{E}(X^{j-1}(X^j)^{\mathsf{T}}) = \sqrt{1-h^2}\,\mathsf{C}.$$

These results are well known in the literature. In [25], this observation was used to modify Gaussian mixtures so as to avoid artificial diffusion in the mixture model. In [14], this observation is used to define a Metropolis–Hastings type algorithm that is efficient for high dimensional inverse problems, known as the pCN algorithm, which is a key tool in this paper. In the cited article, the authors consider distributions of the type

$$\pi(X) \propto e^{-\Phi(x)} \mathcal{N}(x \mid 0, \mathsf{C}).$$

By defining a proposal y drawn from a nonsymmetric proposal $\mathcal{N}(y \mid \sqrt{1-h^2}\, x^{j-1}, h^2\mathsf{C})$, it is easy to check that the Metropolis–Hastings acceptance ratio reduces to

$$\alpha(x^{j-1}, y) = e^{\Phi(x^{j-1}) - \Phi(y)},$$

that is, the potentially high dimensional Gaussian distribution does not appear in the acceptance ratio, thus avoiding the problems that in the high dimension limit lead to practically automatic rejection, in line with the Cameron–Martin theorem.

3.3. Reparametrization. We are now ready to introduce a particular reparametrization that makes it possible to take advantage of the pCN algorithm for MCMC sampling hierarchical models. Consider the posterior density (6) and let $\mathbb{R}^{2n} = \otimes_{j=1}^{N} \mathbb{R}^2$. In each subspace $\mathbb{R}^2$ introduce the new parameters (v_j, τ_j) such that

$$v_j^2 = \frac{\xi_j^2}{\lambda_j}, \quad \tau_j^2 = 2\lambda_j^r, \quad \lambda_j > 0.$$

In the transformation of the old variables (ξ_j, λ_j) in terms of the new ones (v_j, τ_j), we choose the signs so that

$$\lambda_j = \left(\frac{\tau_j^2}{2}\right)^{1/r}, \quad \xi_j = v_j\sqrt{\lambda_j} = v_j \left(\frac{\tau_j^2}{2}\right)^{1/2r} = \frac{v_j |\tau_j|^{1/r}}{2^{1/2r}}.$$

The reparametrization in the probability density requires the determinant of the Jacobian in each subspace $\mathbb{R}^2$:

$$J(v_j, \tau_j) = \det\left(\begin{bmatrix} \frac{\partial \lambda_j}{\partial \tau_j} & \frac{\partial \lambda_j}{\partial v_j} \\ \frac{\partial \xi_j}{\partial \tau_j} & \frac{\partial \xi_j}{\partial v_j} \end{bmatrix}\right) = \det\left(\begin{bmatrix} \frac{2}{r2^{1/r}} \frac{|\tau_j|^{2/r}}{\tau_j} & 0 \\ \frac{1}{r2^{1/2r}} \frac{v_j |\tau_j|^{1/r}}{\tau_j} & \frac{|\tau_j|^{1/r}}{2^{1/2r}} \end{bmatrix}\right) = \frac{2^{1-3r/2}}{r} \frac{|\tau_j|^{3/r}}{\tau_j}.$$

The posterior density, written in terms of the new variables (v, τ) then becomes

$$\begin{aligned} \pi_{V,T}(v, \tau \mid b) &= \prod_{j=1}^{n} |J(v_j, \tau_j)| \pi_{\Xi,\Lambda}(\xi(v,\tau), \lambda(v,\tau) \mid b) \\ &\propto \exp\Bigg(-\frac{1}{2} \left\| b - f\left(\frac{1}{2^{1/2r}} \mathsf{D}_\vartheta^{1/2}(v|\tau|^{1/r})\right)\right\|^2 + \left(r\beta - \frac{3}{2}\right) \sum_{j=1}^{n} \log |\tau_j|^{2/r} \\ &\quad + \sum_{j=1}^{n} \log |\tau_j|^{3/r-1} - \frac{1}{2}\|v\|^2 - \frac{1}{2}\|\tau\|^2 \Bigg) \\ &\propto e^{-\Phi(v,\tau)} \mathcal{N}(v, \tau \mid 0, \mathsf{I}_{2n}), \end{aligned}$$

where the products are understood to be componentwise, i.e.,

$$(v|\tau|^{1/r})_j = v_j|\tau_j|^{1/r}, \quad 1 \le j \le n,$$

and the potential function Φ is defined as

$$\Phi(v,\tau) = \frac{1}{2}\left\| b - f\left(\frac{1}{2^{1/2r}}\mathsf{D}_\vartheta^{1/2}(v|\tau|^{1/r})\right)\right\|^2 - (2\beta-1)\sum_{j=1}^n \log|\tau_j|.$$

In the following, we will consider four special cases: When $r=1$, the hyperprior is the gamma distribution, and we have

$$r=1: \quad \Phi(v,\tau) = \frac{1}{2}\left\| b - f\left(\frac{1}{\sqrt{2}}\mathsf{D}_\vartheta^{1/2}(v|\tau|)\right)\right\|^2 - (2\beta-1)\sum_{j=1}^n \log|\tau_j|,$$

while if $r=-1$, the hyperprior is the inverse gamma distribution, and

$$r=-1: \quad \Phi(v,\tau) = \frac{1}{2}\left\| b - f\left(\sqrt{2}\mathsf{D}_\vartheta^{1/2}\frac{v}{|\tau|}\right)\right\|^2 - (2\beta-1)\sum_{j=1}^n \log|\tau_j|.$$

In the case $r=1/2$, we have

$$r=\frac{1}{2}: \quad \Phi(v,\tau) = \frac{1}{2}\left\| b - f\left(\frac{1}{2}\mathsf{D}_\vartheta^{1/2}(v\tau^2)\right)\right\|^2 - (2\beta-1)\sum_{j=1}^n \log|\tau_j|,$$

and for $r=-1/2$, we have

$$r=-\frac{1}{2}: \quad \Phi(v,\tau) = \frac{1}{2}\left\| b - f\left(2\mathsf{D}_\vartheta^{1/2}\left(\frac{v}{\tau^2}\right)\right)\right\|^2 - (2\beta-1)\sum_{j=1}^n \log|\tau_j|.$$

In the following section, we use the proposed reparametrization to explore the posterior densities corresponding to these four choices of r.

4. Computed examples. In this section, we investigate numerically the effectiveness of the reparametrization of the problem in combination with the pCN algorithm, with special emphasis on the role of the parameter r. We start by discussing the model problems used in testing the proposed sampler.

4.1. Model problem. We consider here a linear inverse problem in the form of a one-dimensional deconvolution. Let $g : [0,1] \to \mathbb{R}$ be the function to be estimated from noisy observations of the convolution of the signal with a Gaussian kernel,

$$b_j = \int_0^1 a(t_j - s)g(s)ds + \varepsilon_j, \quad a(t) = A\exp\left(-\frac{1}{2w^2}t^2\right), \quad 0 < t_1 < \cdots < t_m < 1, \tag{18}$$

of width $w = 0.02$ and amplitude $A = 6.2$. We discretize the convolution integral using a piecewise constant approximation with $n = 128$ intervals, and assume that the observation

points t_j coincide with every sixth discretization node s_k, that is, $t_j = s_{1+6(j-1)}$, $1 \le j \le m = 22$. This yields the approximation

$$b = \mathsf{A}z + \varepsilon, \quad \mathsf{A} \in \mathbb{R}^{m\times n}, \quad z_j = g(s_j).$$

To generate the data, we assume that the generative model g is a piecewise constant function. To avoid the inverse crime of using the same model for data generation and for the solution of the inverse problem, the data are generated by using a fine discretization mesh with 1000 discretization intervals and, subsequently, Gaussian noise, ε from $\mathcal{N}(0, \sigma^2 \mathsf{I}_m)$ with $\sigma = 0.03$ added to them. The generative model and the noisy data are shown in the top left panel of Figure 1.

To find a sparse representation of the discrete signal z, let $\mathsf{L} \in \mathbb{R}^{n\times n}$ be the finite difference matrix,

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

and express z in terms of $x \in \mathbb{R}^n$ as $\mathsf{L}z = x$. Implicitly, we assume here a boundary condition $z_0 = g(0) = 0$. Therefore, we can write the forward model as

$$b = \mathsf{A}\mathsf{L}^{-1}x + \varepsilon,$$

and after scaling the data and the forward map by $1/\sigma$,

$$\widehat{b} = \frac{1}{\sigma} b, \quad \widehat{\mathsf{A}} = \frac{1}{\sigma}\mathsf{A}\mathsf{L}^{-1},$$

thus whitening the noise, we arrive at the expression of the problem in standard form,

$$\widehat{b} = \widehat{\mathsf{A}}x + e, \quad e \sim \mathcal{N}(0, \mathsf{I}_m). \tag{19}$$

We point out that the vanishing boundary value of the unknown z at the endpoint $t = 0$ makes the data nonuniformly sensitive to the components of the vector x. The sensitivity of the data to the components of the vector x is not addressed here: We refer to [7] for the discussion of the topic.

4.2. Hypermodels and MAP estimates. We test the sampling algorithm with four different generalized gamma hypermodels, corresponding to $r_1 = 1$, $r_2 = 1/2$, $r_3 = -1/2$, and $r_4 = -1$. We denote the corresponding hyperparameter values by (β_j, ϑ_j), $1 \le j \le 4$. Here, for simplicity, we choose all components of the vectors ϑ_j equal, so this parameter can be treated as a scalar.

In order to initialize the MCMC algorithm without the need of a long burn-in run, we first compute a MAP estimate for each case using the hybrid IAS algorithm:

1. Phase I: Run the IAS algorithm with values $(r_1, \beta_1, \vartheta_1)$ until convergence criterion is met.

2. Phase II: If $j > 1$, continue the IAS iteration with values $(r_j, \beta_j, \vartheta_j)$ until convergence criterion is met.

The compatibility conditions (16)–(17) give an automatic way to set the hyperparameters for $j > 1$. Recalling that by Theorem 3.1 for $r = 1$, a value β_1 close to $3/2$ promotes sparsity, we write $\beta_1 = 3/2 + \eta$, where $\eta > 0$ is small. We also set the value ϑ_1, allowing automatic selection of the hyperparameters for $j > 1$. The compatibility conditions with $r_1 = 1$ yield

$$\vartheta_j \left(\beta_j - \frac{3}{2r_j}\right)^{1/r_j} = \vartheta_1 \eta, \quad \vartheta_j \frac{\Gamma(\beta_j + \frac{1}{r_j})}{\Gamma(\beta_j)} = \vartheta_1 \left(\eta + \frac{3}{2}\right), \quad j = 2, 3, 4.$$

Straightforward calculations based on the properties of the gamma function lead to the following formulas for the parameters:

Case $r_2 = 1/2$: We have

$$\beta_2 = \frac{6m + 1 + \sqrt{48m + 1}}{2(m - 1)}, \quad \text{where } m = 1 + \frac{3}{2\eta},$$

and

$$\vartheta_2 = \vartheta_1 \frac{\eta}{(\beta_2 - 3)^2}.$$

Case $r_3 = -1/2$: We have

$$\beta_3 = \frac{6 + 3m \pm \sqrt{m^2 + 80m}}{2(m - 1)}, \quad \text{where } m = 1 + \frac{3}{2\eta},$$

and

$$\vartheta_3 = \vartheta_1 \eta (\beta_2 + 3)^2.$$

Case $r_3 = -1$: We have

$$\beta_4 = 1 + \frac{5}{3}\eta$$

and

$$\vartheta_4 = \vartheta_1 \eta \left(\beta_4 + \frac{3}{2}\right).$$

The numerical values used in the computations are give in Table 1.

In Phase II, the IAS iterations start with the final θ of Phase I. In both phases, the IAS iterations stop as soon as

$$\frac{\|\theta^{t-1} - \theta^t\|}{\|\theta^{t-1}\|} < 0.005.$$

The MAP estimates computed by the IAS algorithm are shown in Figure 1, where the number of iterations needed for satisfying the stopping criterion is indicated. Observe that the MAP estimate with $r = 1$ is the starting point for all the hybrid models with $r < 1$.

Table 1

Hyperparameter values used in the computations. The values ϑ for the hybrid models are determined by requiring that at $x_j = 0$, the values θ_j given by the IAS algorithm are independent of the hypermodel, and that the marginal expectations for θ_j are independent of the model.

	$r=1$	$r=1/2$	$r=-1/2$	$r=-1$
β	1.501	3.0918	2.0165	1.0017
ϑ	5×10^{-2}	5.9323×10^{-3}	1.2583×10^{-3}	1.2308×10^{-4}

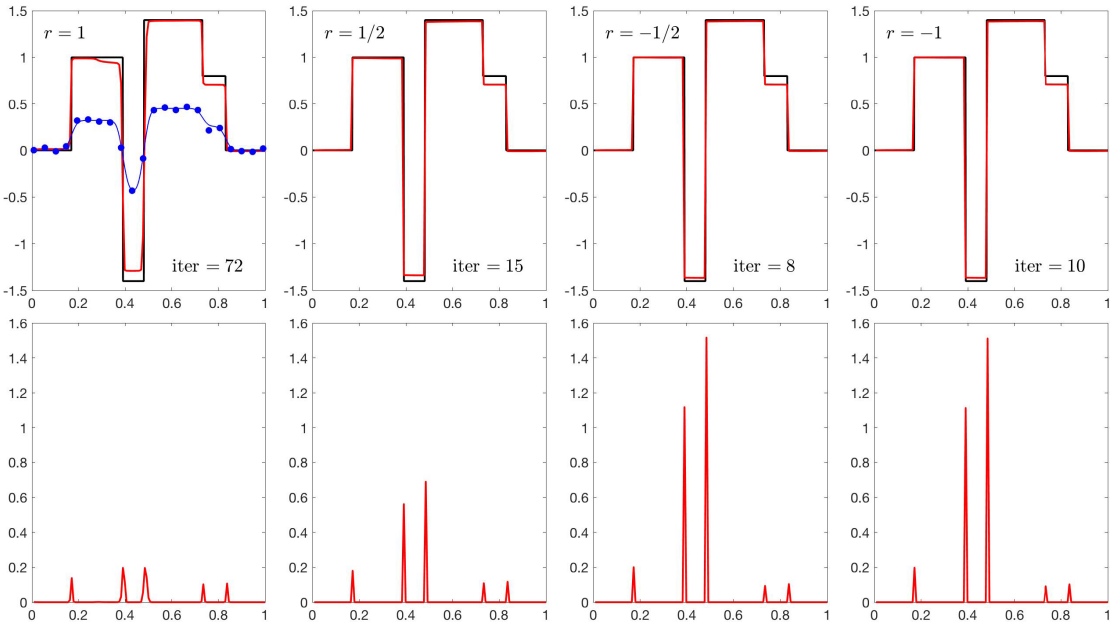

Figure 1. *The MAP estimates computed by the hybrid IAS algorithm for z (upper row) in red, the black curves corresponding to the generative model, and for θ (lower row). The noisy data are shown as blue dots in the top left panel, the solid blue line indicating the noiseless convolution data. The left panels correspond to the gamma hyperprior $r = 1$, or Phase I, and the result is the starting point for the Phase II iterations for $r < 1$. The number of iterations for $r < 1$ refers to the iteration rounds of Phase II. The parameter values are given in Table 1.*

4.3. Sampling. We begin by applying the sampling algorithm with the gamma hyperprior ($r = 1$) for the linear model (19), initiating the sample from the IAS-based MAP estimate. After some preliminary tests, the step size control parameter is set to $h = 0.05$, yielding consistently an acceptance rate close to 6.3%. Decreasing the step size increases the acceptance rate, e.g., $h = 0.02$ yields an acceptance rate of 33%. The choice of the step size will be justified momentarily.

The relatively small step size h implies that the draws in the sample are correlated, so to improve the sample quality, we retain only a subsample: In our test, we choose the computed sample size to be 10 000 000, and to decrease the dependency of the sample point, we keep only every 1000th point, reducing the effective sample size to $N = 10\,000$. The run time in a standard laptop is only 135 s, as the proposal density is pure white noise, and only one matrix-vector product for deciding on the acceptance is required.

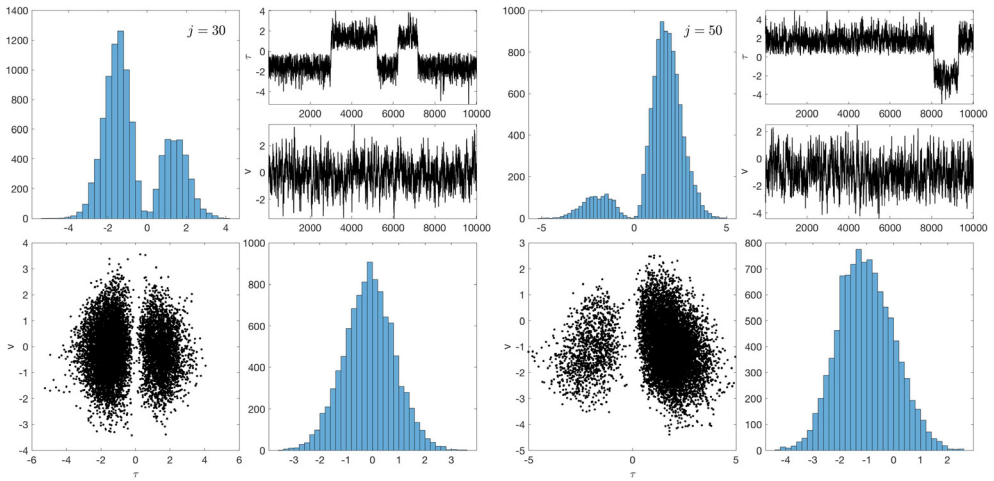

Figure 2. *Samples of two selected pairs (τ_j, v_j) shown as scatter plots and histograms with hypermodel $r = 1$. On the right, $j = 30$, corresponding to a position where the generative function is constant, and $j = 50$ corresponding to a local minimum of the MAP estimate.*

To analyze the mixing properties of the sampler, we select two indices, $j_1 = 30$ and $j_2 = 50$, corresponding to values $t_{j_1} \approx 0.23$ and $t_{j_2} \approx 0.39$, the former corresponding to a position around which the generative function is constant, and the latter near a jump where the MAP estimate has a local minimum. Figure 2 shows the scatter matrices of the pair (τ_j, v_j) for $j = j_1$ and $j = j_2$, respectively, as well as the time traces of the sample.

We observe that the distribution of τ_j is bimodal, reflecting the fact that the physical scaled parameter λ_j depends on τ_j^2 which is insensitive to the sign of τ_j. Furthermore, at $j = 30$, the sample is centered near the coordinate origin, indicating that the sampler recognizes the point as belonging to the flat background. At $j = 50$, the bimodal nature is visible, however, the values of v_j are predominantly negative, indicating a presence of a negative jump, since

$$x_j = \vartheta^{1/2} \frac{1}{\sqrt{2}} v_j |\tau_j|, \quad \theta_j = \frac{1}{2} \vartheta \tau_j^2.$$

However, the histogram of v_j does not exclude the value $v_j = 0$ corresponding to a background value $x_j = 0$, indicating uncertainty in identifying the jump unambiguously. Interestingly, if the step size h is decreased to increase the acceptance rate, the sampler fails to identify the bimodal nature of the distribution and samples only from the mode $\tau_j > 0$. This is not a serious issue, as the bimodality is simply a result of coordinate representation; however, we chose here to select the step size so that the feature becomes visible.

Figure 3 shows the autocorrelation functions of the retained sample of the components (x_j, θ_j), $j = 30$ and $j = 50$,

$$C(x_j)(\ell) = \frac{1}{\|x_j\|} \sum_{k=1+\ell}^{N-\ell} (x_j^k - \overline{x}_j)(x_j^{k-\ell} - \overline{x}_j), \quad \ell = 0, 1, 2, \ldots,$$

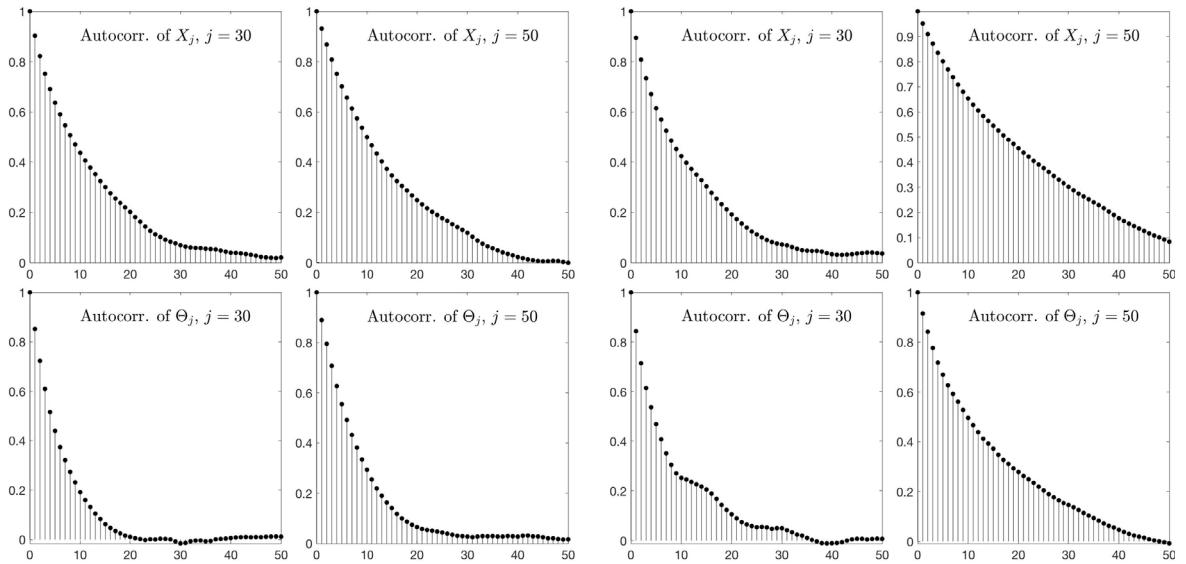

Figure 3. *Autocorrelation functions of the sample corresponding to the gamma hyperprior (two left columns) and to the hypermodel with $r = 1/2$ (two right columns). The top row shows the autocorrelation functions of the variables X_j corresponding to a background value $j = 30$ and to a jump value $j = 50$, and the bottom row the autocorrelation function for the corresponding varibles Θ_j.*

where

$$\overline{x}_j = \frac{1}{N}\sum_{k=1}^{N} x_j^k, \quad \|x_j\| = \left(\sum_{k=1}^{N}\left(x_j^k\right)^2\right)^{1/2}.$$

The autocorrelation plots give a sense of the independence of the subsampled draws. There seems to be no significant difference between a background variable and one corresponding to a jump.

Before analyzing the sample further, we run similar pCN sampling using the other hypermodels. It turns out that as r decreases, the sampling becomes more challenging. We start with $r = 1/2$. Using the same step size $h = 0.05$ as in the case $r = 1$ the acceptance rate falls to 4.8%, so we decrease the step size slighltly to $h = 0.03$, yielding an acceptance rate of 16.1%. Generating a sample of size 10 000 000, retaining again every 1000th sample point, takes 131 s on a standard laptop. In this case, even with a larger step size, the sampler is unable to detect the coordinate bimodality and samples only from the mode where $\tau_j > 0$, as shown by the scatter plots in Figure 4.

The level of independence of the draws can be inferred from the autocorrelation functions shown in Figure 3. We observe that the correlation level is not significantly different from that corresponding to the gamma hypermodel.

Consider now the hypermodels with negative r, yielding a highly nonlinear Gibbs energy functional with a strong sparsity promotion in the MAP estimation problems. It turns out that these models pose a challenge for sampling as well. We start with $r = -1/2$. Using the same step size as in the case $r = 1/2$, $h = 0.03$ the acceptance rate is as low as 0.12%, so it is

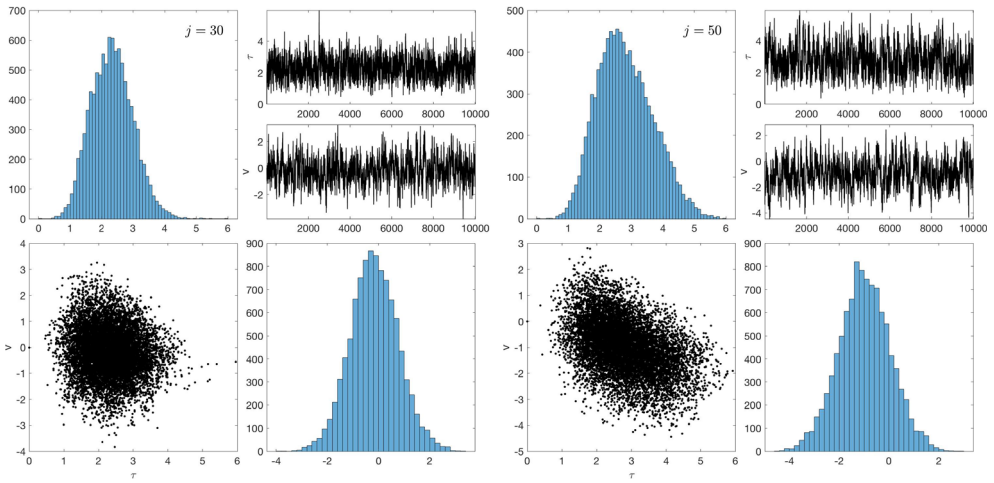

Figure 4. *Samples of two selected pairs (τ_j, v_j) shown as scatter plots and histograms, hypermodel corresponding to $r = 1/2$. On the right, $j = 30$, corresponding to a position where the generative function is constant, and $j = 50$ corresponding to a local minimum of the MAP estimate.*

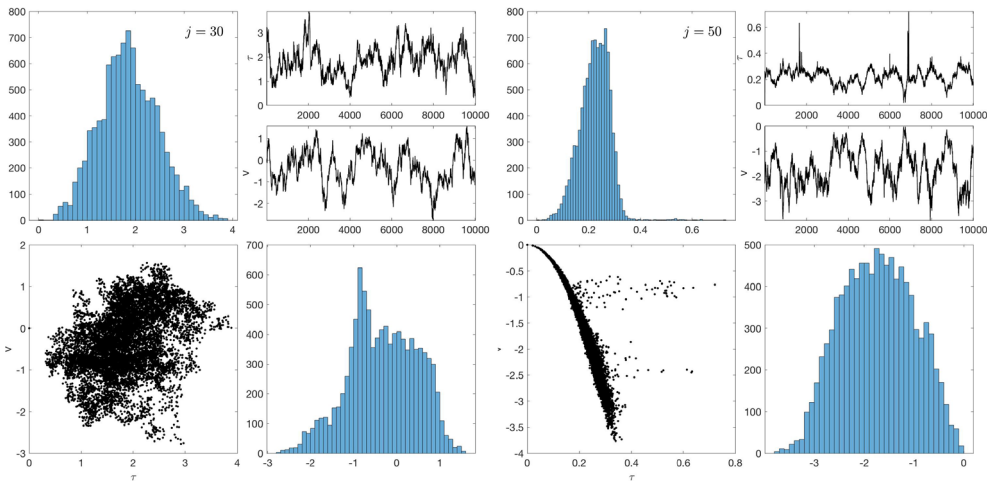

Figure 5. *Samples of two selected pairs (τ_j, v_j) shown as scatter plots and histograms with hypermodel $r = -1/2$. On the right, $j = 30$, corresponding to a position where the generative function is constant, and $j = 50$ corresponding to a local minimum of the MAP estimate.*

reasonable to decrease the step size. After extensive testing, we found that $h = 0.008$ yields a reasonable acceptance rate of 6%, and $h = 0.005$ gives a 12% acceptance. Since the former figure is close to the values in the previous experiments, we set $h = 0.008$. To make the results comparable to the previous ones, we keep the same sample size of 10 000 000, retaining every 1000th sample point. The computing times are close to the ones reported above, on the order of 110 s.

Figure 5 shows the scatter plots and histograms of the parameters (τ_j, v_j). We observe that at $j = 50$, the points are concentrated near a parabolic curve, which is to be expected, as the variables are related to each other through the formula

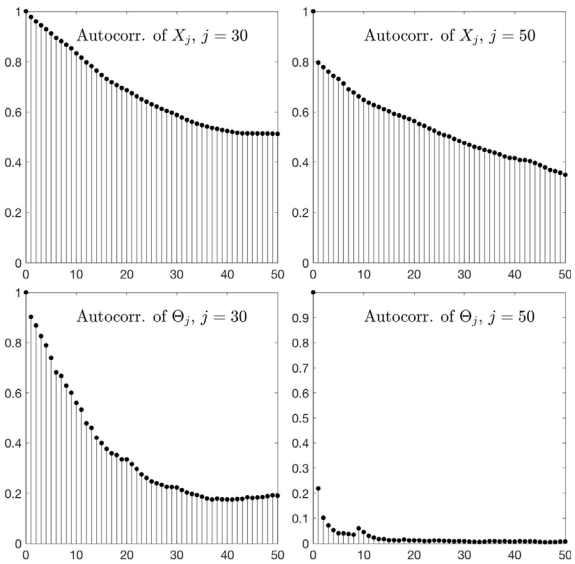

Figure 6. *Autocorrelation functions of the sample corresponding to the hyperprior with $r = -1/2$. The top row shows the autocorrelation functions of the variables X_j, $j = 30$, corresponding to a background value and to a jump value, $j = 50$; the bottom row shows the autocorrelation functions for the corresponding variables Θ_j.*

$$\xi_j = 2\frac{v_j}{\tau_j^2},$$

and at $j = 50$, we expect that the likelihood favors combinations of (τ_j, v_j) yielding a negative value for ξ_j.

The sample histories of τ_j and v_j do not indicate a good mixing, however, from the practical point of view, the important question is how well the sample mixes the values (x_j, θ_j). To explore this question, we plot again the autocorrelation functions of the selected components in Figure 6. Interestingly, while the autocorrelation functions of the x-variables are decreasing rather slowly, the θ-variable autocorrelations indicate rather good mixing.

Finally, we consider the inverse gamma hyperprior, $r = -1$. Numerical tests indicate that the proposed sampler struggles to find a reasonably well mixed sample. To demonstrate this, we select first the step size to be $h = 0.02$, which leads to an acceptance rate as low as 0.001%, that is, one proposal of every 100 000 is accepted on average. We then generate a sample of size 10^8, keeping only every 10 000th point, i.e., we have a sample of size 10 000 in which approximately 90% of the points are repeated values corresponding to rejections. The computing time of this sample is less than 18 minutes. The (τ_j, v_j) scatter plots for the two selected values of j are shown in Figure 7.

The time traces of the samples reveal that the sample is of low quality, and reliable conclusions could hardly be based on this sample. Numerical experiments show that decreasing the step size does not improve significantly the sample quality. The plots show, however, some features that may help in designing a better sampler. First, the scatter plot for $j = 50$ shows again that the coordinate transformation leads to a bimodal distribution, which is hard to reproduce with a smaller step size. Second, we observe that the sampler proposes points

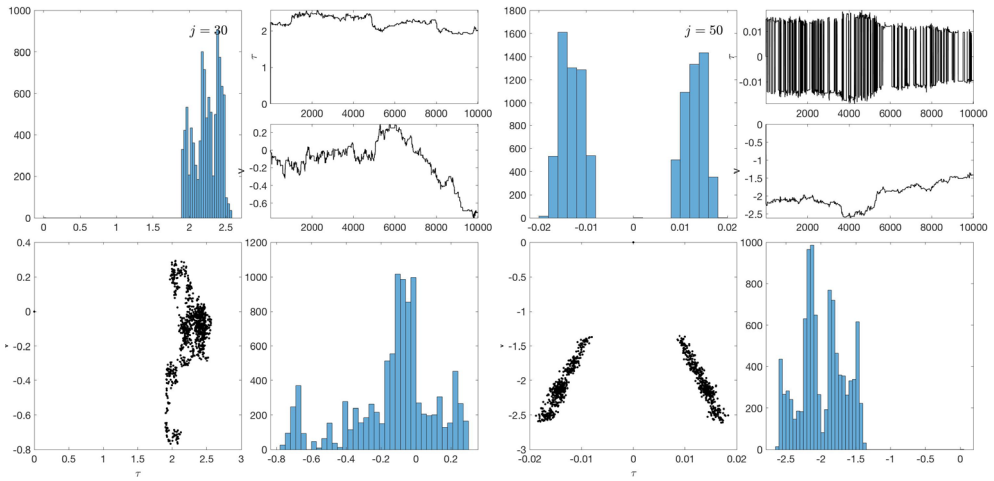

Figure 7. *Samples of two selected pairs (τ_j, v_j) shown as scatter plots and histograms with hypermodel $r = -1$. On the right, $j = 30$, corresponding to a position where the generative function is constant, and $j = 50$ corresponding to a local minimum of the MAP estimate.*

that lie along lines passing through the origin. Recalling that for $r = -1$, the coordinate transformation implies that

$$\xi_j = \sqrt{2}\frac{v_j}{|\tau_j|},$$

the absolute value of the slope of the line determines the proposed value for ξ_j. It turns out that for $j = 50$, the sample history of ξ_{50} is quite satisfactory, with a relatively short correlation length (not shown here.)

To improve the mixing in the case $r = -1$, we propose a modification of the pCN algorithm based on a reparametrization of the pairs (τ_j, v_j). Consider the problem of generating a standard normal distribution in the plane $\mathbb{R}^2$. Let $X^{j-1} \sim \mathcal{N}(0, \mathsf{I}_2)$, and denote $R^{j-1} = \|X^{j-1}\|$. We have

$$R^{j-1} \sim \text{Rayleigh}(1).$$

Let $W = (W_1, W_2) \sim \mathcal{N}(0, \mathsf{I}_2)$, and assume that W is independent of X^{j-1}. For any $k > 0$, we define

$$R^j = \left((1-k^2)\left(R^{j-1}\right)^2 + 2k\sqrt{1-k^2}R^{j-1}W_1 + k^2\|W\|^2\right)^{1/2}. \tag{20}$$

We claim that

$$R^j \sim \text{Rayleigh}(1).$$

This is a direct consequence of the fact that

$$X^j = \sqrt{1-k^2}X^{j-1} + kW \sim \mathcal{N}(0, \mathsf{I}_2).$$

Without loss of generality, we may assume that the coordinates are chosen so that $X_2^{j-1} = 0$. Then

$$R^j = \|X^j\|$$

follows a Rayleigh distribution.

Let Φ^{j-1} be the phase angle of X^{j-1}:

$$\Phi^{j-1} = \operatorname{atan}(X_2^{j-1}, X_1^{j-1}) \sim \text{Uniform}(\mathbb{R}/2\pi).$$

For an arbitrary $h > 0$, define

$$\Phi^j = \Phi^{j-1} + h\Omega, \quad \Omega \sim \mathcal{N}(0,1). \tag{21}$$

Then, $\Phi^j \sim \text{Uniform}(\mathbb{R}/2\pi)$. Based on these observations, we introduce the following two-phase proposal:

Given $h > 0$, $k > 0$, and the current point $x^{j-1} = (\tau^{j-1}, v^{j-1}) \in \mathbb{R}^2$,

1. *set $r^{j-1} = \|x^{j-1}\|$ and*

$$r^j = \left((1-k^2)\left(r^{j-1}\right)^2 + 2k\sqrt{1-k^2}r^{j-1}w_1 + k^2\|w\|^2\right)^{1/2}, \quad w \sim \mathcal{N}(0, \mathsf{I}_2);$$

2. *set $\varphi^{j-1} = \operatorname{atan}(v^{j-1}, \tau^{j-1})$ and*

$$\varphi^j = \varphi^{j-1} + h\omega, \quad \omega \sim \mathcal{N}(0,1);$$

3. *define*

$$x^j = (\tau^j, v^j) = (r^j \cos\varphi^j, r^j \sin\varphi^j).$$

In this manner, the new variable X^j with realization x^j follows the same Gaussian distribution as X^{j-1} as in the standard pCN proposal, but we control separately the step size in the radial and in the angular directions. The two free variables k and h add complexity to the tuning process, but may lead to a chain with better mixing properties.

To demonstrate the viability of the proposed algorithm, we run the sampler with parameter values $h = 0.001$ and $k = 0.05$, yielding an acceptance rate of 1.5%. We compute a sample of size 5 000 000, retaining every 500th realization. The computing time is slightly longer than with plain pCN, requiring 91 s on a standard laptop. The scatterplots corresponding to this sample are shown in Figure 8.

To assess the quality of the sample, we compute again the autocorrelation functions for the selected variables; see Figure 9. We observe that while the autocorrelation indicates poor quality of the sample of X_{30}, the autocorrelation of Θ_{30} decreases relatively rapidly. The conclusion therefore is that if the sample-based estimate of Θ_{30} is small, we may claim with high certainty that X_{30} is small, too.

To summarize the results of the samplers, in Figure 10 we plot the estimated posterior means and 90% credible envelopes of the variables z, x, and θ corresponding to the four hypermodels.

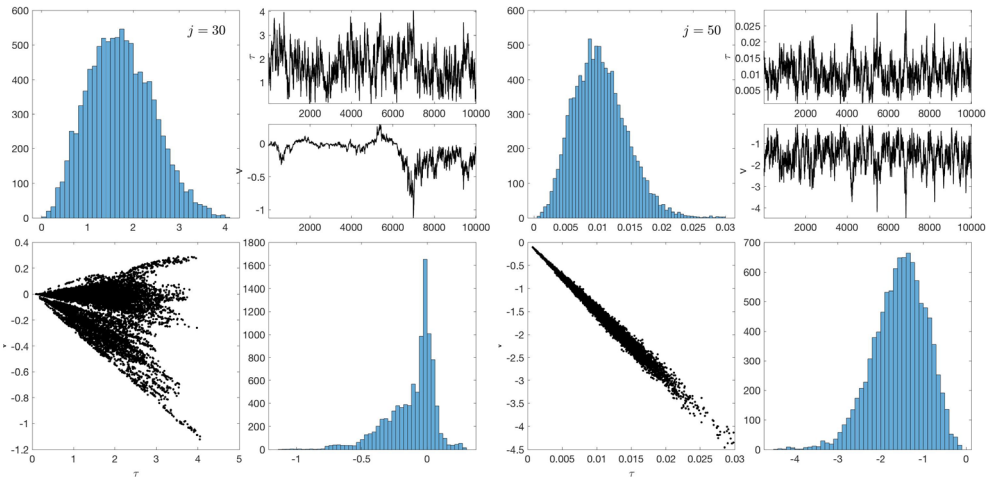

Figure 8. *Samples of two selected pairs (τ_j, v_j) shown as scatter plots and histograms with hypermodel $r = -1$ using the modified pCN algorithm. On the right, $j = 30$, corresponding to a position where the generative function is constant, and $j = 50$ corresponding to a local minimum of the MAP estimate.*

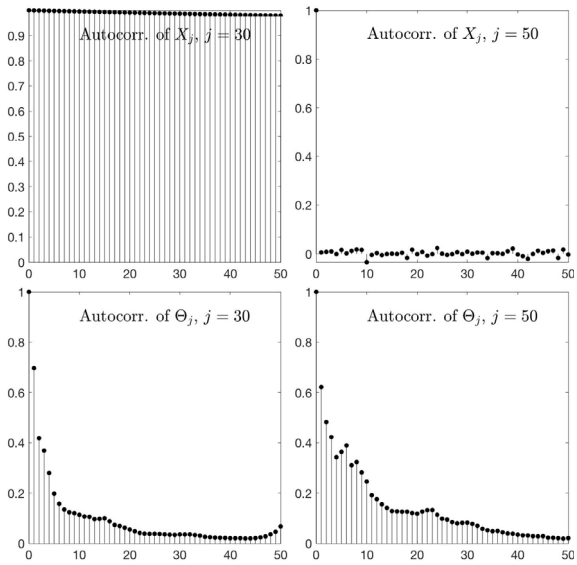

Figure 9. *Autocorrelation functions of the sample corresponding to the hyperprior with $r = -1$ corresponding to the modified pCN sampler. The top row shows the autocorrelation functions of the variables X_j corresponding to a background value $j = 30$ and to a jump value $j = 50$, and the bottom row for the corresponding varibles Θ_j.*

The results show that while the MAP estimates for all models can be considered sparse solutions, the posterior means with $r = 1/2$ and $r = 1$ in particular, do not reflect the sparsity promoting nature of the prior models. This suggests that the posterior mean is not a particularly good representative summary of the posterior distribution, even if the individual samples would reflect the sparsity prior.

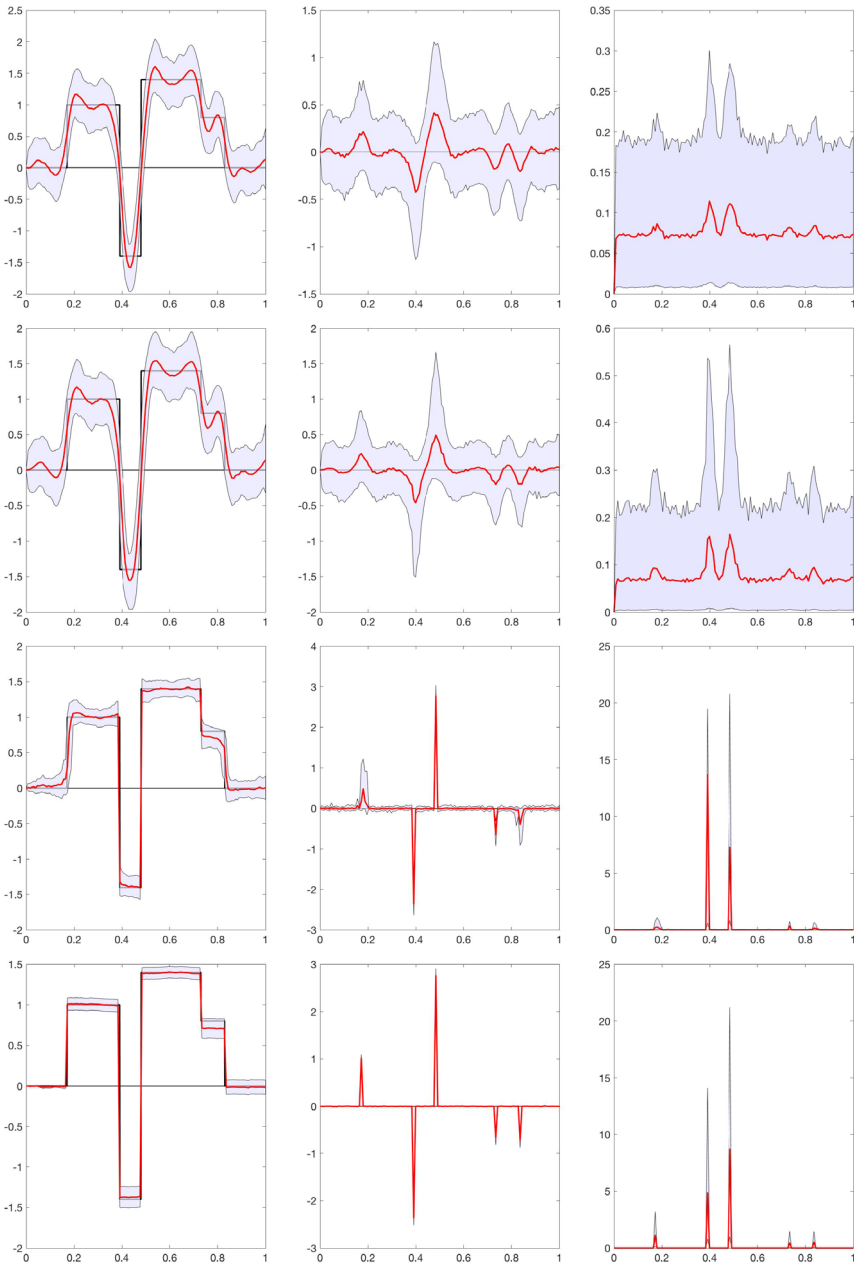

Figure 10. *Posterior means (red curve) and the 90% credible envelopes for z (left), x (middle), and θ (right) with hypermodels $r = 1$ (top), $r = 1/2$, $r = -1/2$, and $r = -1$ (bottom). The inverse gamma sample is based on the modified pCN algorithm.*

The smoothness of the posterior mean raises the question to what extent profiles sampled from the posterior density are compressible. To further investigate this issue, let θ^j denote the jth sample vector of the variance parameter, and define δ-compressibility by the formula

$$\|\theta^j\|_{0,\delta} = \mathrm{card}\{1 \le k \le n \mid \theta_k^j > \delta\},$$

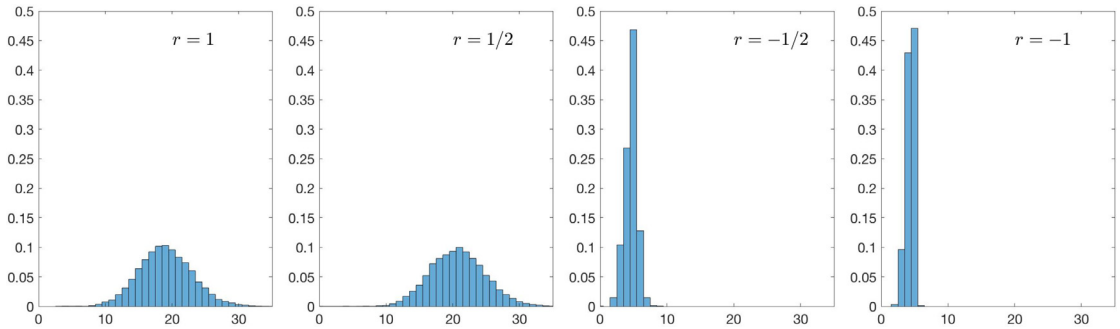

Figure 11. *Histograms of the number of components in the vectors θ^j that are above the threshold value δ (22), thus indicating the level of compressibility of the sample vectors. Observe that the number of nonzero increments in the generative model is 5, corresponding to the maximum for $r = -1/2$ and $r = -1$.*

where $\delta > 0$ is a given threshold. We set the threshold to correspond to one standard deviation above the mean of the gamma distribution,

$$\delta = \beta_1 \vartheta + \sqrt{\beta_1}\vartheta. \tag{22}$$

Figure 11 shows the histograms of the number of components in the vectors θ^j exceeding the threshold δ. The samples based on models $r = 1$ and $r = 1/2$ identify significantly more increments above the threshold than the cardinality of the support of the generative model, the maximum of the histograms being around twenty, while for both $r = -1/2$ and $r = -1$, the maximum of the histogram is at 5, which coincides with the number of nonzero increments in the generative model.

5. Discussion. The difficulties of sampling from posterior densities corresponding to hierarchical Bayesian models are twofold: The curse of dimensionality makes efficient sampling hard, and the strong correlation between the parameters at different levels adds an extra bottleneck for samplers. In this article, the former problem has been addressed by transforming the problem so that the pCN sampling scheme can be applied. As the examples with strongly nonconvex energy functionals show, the transformation does not completely remove the second problem; however, the transformation gives enough insight to allow further developments of the sampling strategy so that at least in the case of the inverse gamma hyperprior model, a relatively efficient algorithm can be found. The proposed coordinate transformation generates nonlinearities, potentially complicating the likelihood density. The effects of these nonlinearities on the sampler are a topic for further studies. Moreover, a natural question to be addressed in the future is if the proposed algorithm can be generalized for hyperprior models more general than the inverse gamma distribution.

The sampling analysis for sparsity promoting hypermodels reveals that the concept of sparsity promotion is more complex than the analysis of the maximum a posteriori estimates reveals: The MAP estimate may identify the correct number of nonzero entries in a sparse vector, but sampling from the posterior density may not consistently support the level of sparsity. This was clearly demonstrated by the computed examples with parameter values $r = 1$ and $r = 1/2$: While the MAP estimates localize the discontinuities well, the draws

from the posterior seem to have significantly more discontinuities. This finding underlines also the observation that for sparse recovery, the MAP estimate may be a better summary of the posterior than the posterior mean, which in our example resembles more a smooth reconstruction than a discontinuous one.

The main focus of this article is to investigate to what extent the class of hierarchical models that constitute the core of the IAS algorithm are concentrated on sparse solutions, and the sampler proposed in this article is tailored for this particular class. We point out that while efficient generic samplers for hierarchical models exist, e.g., HMC-NUTS [20], running them in high dimensions is not without challenges. A comparison of the performance of different samplers, as well as other bespoken methods for quantifying the uncertainty (e.g., [1]) in this class of hierarchical models is beyond the scope of this article and will be addressed in future works.

Finally, we point out that the analysis above did not address the question of data sensitivity. It has been demonstrated that if the data have variable sensitivity to different components of the unknown, both MAP and posterior mean estimates may fail to recognize some of the nonzero components in the generative model. While the sensitivity analysis could have been included in the discussion here, it was omitted in order to keep the focus on the sampling techniques proposed in this article, but may be the topic of future work.

REFERENCES

[1] S. Agrawal, H. Kim, D. Sanz-Alonso, and A. Strang, *A variational inference approach to inverse problems with gamma hyperpriors*, SIAM/ASA J. Uncertain. Quantif., 10 (2022), pp. 1533–1559.

[2] S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart, *Analysis of the Gibbs sampler for hierarchical inverse problems*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 511–544.

[3] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Trans. Signal Process., 54 (2006), pp. 4311–4322.

[4] M. Betancourt and M. Girolami, *Hamiltonian Monte Carlo for hierarchical models*, Curr. Trends Bayesian Methodol. Appl., 79 (2015), pp. 2–4.

[5] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal, *Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical Bayesian approach. Part* II: *Theoretical analysis*, SIAM J. Imaging Sci., 13 (2020), pp. 1990–2028.

[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 Topog., 32 (2019), pp. 363–393.

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

[8] 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.

[9] D. Calvetti, M. Pragliola, and E. Somersalo, *Sparsity promoting hybrid solvers for hierarchical Bayesian inverse problems*, SIAM J. Sci. Comput., 42 (2020), pp. A3761–A3784.

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

[11] D. Calvetti and E. Somersalo, *Bayesian Scientific Computing*, Springer, New York, 2023.

[12] E. Candes and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 12 (2005), pp. 4203–4215.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[14] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.
[15] P. Damlen, J. Wakefield, and S. Walker, *Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables*, J. R. Stat. Soc. Ser. B Stat. Methodol., 61 (1999), pp. 331–344.
[16] J. E. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.
[17] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[18] A. Durmus, É Moulines, and M. Pereyra, *A proximal Markov chain Monte Carlo method for Bayesian inference in imaging inverse problems: When Langevin meets Moreau*, SIAM Rev., 64 (2022), pp. 991–1028.
[19] J. Glaubitz, A. Gelb, and G. Song, *Generalized sparse Bayesian learning and application to image reconstruction*, SIAM/ASA J. Uncertain. Quantif., 11 (2023), pp. 262–284.
[20] M. D. Hoffman and A. Gelman, *The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*, J. Mach. Learn. Res., 15 (2014), pp. 1593–1623.
[21] J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, *Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography*, Inverse Problems, 16 (2000), 1487.
[22] Y. Li and D. W. Oldenburg, 3-*D inversion of magnetic data*, Geophysics, 61 (1996), pp. 394–408.
[23] Y. Li and D. W. Oldenburg, 3-*D inversion of gravity data*, Geophysics, 63 (1998), pp. 109–119.
[24] F. H. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen, *Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates*, NeuroImage, 31 (2006), pp. 160–171.
[25] J. Liu and M. West, *Combined parameter and state estimation in simulation-based filtering*, in Sequential Monte Carlo Methods in Practice, Springer, New York, 2001, pp. 197–223.
[26] F. Lucka, *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L_1-type priors*, Inverse Problems, 28 (2012), 125012.
[27] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed., Academic Press, Amsterdam, 2009.
[28] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld, *A general framework for the parametrization of hierarchical models*, Statist. Sci., 22 (2007), pp. 59–73.
[29] G. O. Roberts and S. K. Sahu, *Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler*, J. R. Stat. Soc. Ser. B Stat. Methodol., 59 (1997), pp. 291–317.
[30] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 58 (1996), pp. 267–288.
[31] 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 J. Imaging Sci., 13 (2020), pp. 1945–1989.
[32] Y. Xiao and J. Glaubitz, *Sequential image recovery using joint hierarchical Bayesian learning*, J. Sci. Comput., 96 (2023) 4.