SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 6, No. 3, pp. 1076–1100

Low-Rank Independence Samplers in Hierarchical Bayesian Inverse Problems*

D. Andrew Brown†, Arvind Saibaba‡, and Sarah Vallélian§

Abstract. In Bayesian inverse problems, the posterior distribution is used to quantify uncertainty about the reconstructed solution. In fully Bayesian approaches in which prior parameters are assigned hyperpriors, Markov chain Monte Carlo algorithms often are used to draw samples from the posterior distribution. However, implementations of such algorithms can be computationally expensive. We present a computationally efficient scheme for sampling high-dimensional Gaussian distributions in ill-posed Bayesian linear inverse problems. Our approach uses Metropolis–Hastings independence sampling with a proposal distribution based on a low-rank approximation of the prior-preconditioned Hessian. We show the dependence of the acceptance rate on the number of eigenvalues retained and discuss conditions under which the acceptance rate is high. We demonstrate our proposed sampler by using it with Metropolis–Hastings-within-Gibbs sampling in numerical experiments in image deblurring, computerized tomography, and NMR relaxometry.

Key words. computerized tomography, image deblurring, low-rank approximation, Metropolis–Hastings independence sampler, prior-preconditioned Hessian

AMS subject classifications. 65F22, 65C05, 65C40

DOI. 10.1137/17M1137218

1. Introduction.

Inverse problems aim to recover quantities that cannot be directly observed, but can only be measured indirectly and in the presence of measurement error. Such problems arise in many applications in science and engineering, including medical imaging [27], earth sciences [2], and particle physics [40]. The deterministic approach to inverse problems involves minimizing an objective function to obtain a point estimate of the unknown parameter. Inverse problems also admit a Bayesian interpretation, facilitating the use of prior information and allowing full quantification of uncertainty about the solutions in the form of a posterior probability distribution. An overview of Bayesian approaches to inverse problems is available in [34, 38, 58]. A recent special issue of *Inverse Problems* also highlights the advances in the Bayesian approach and the broad impacts of its applicability [12].

In the Bayesian statistical framework, the parameters of interest, $\boldsymbol{x}$, and the observed data, $\boldsymbol{b}$, are modeled as random variables. A priori uncertainty about the parameters is quantified in the prior distribution, $\pi(\boldsymbol{x})$. Bayesian inference then proceeds by updating the

*Received by the editors July 5, 2017; accepted for publication (in revised form) May 21, 2018; published electronically July 19, 2018.

http://www.siam.org/journals/juq/6-3/M113721.html

Funding: The work of the first author was partially supported by grants CMMI-1563435 and EEC-1744497 from the National Science Foundation (NSF). This material is based upon work partially supported by the NSF under grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute (SAMSI).

†Department of Mathematical Sciences, Clemson University, Clemson, SC 29634 (ab7@clemson.edu).

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (asaibab@ncsu.edu).

§Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709 (svallelian@samsi.info).

information about these parameters given the observed data. The updated information is quantified in the posterior distribution obtained via Bayes's rule, $\pi(\boldsymbol{x}|\boldsymbol{b}) \propto f(\boldsymbol{b}|\boldsymbol{x})\pi(\boldsymbol{x})$, where $f(\cdot|\boldsymbol{x})$ is the assumed data-generating model determined by the forward operator and $\boldsymbol{x}$, called the likelihood function. Rather than providing a single solution to the inverse problem, the Bayesian approach provides a distribution of plausible solutions. Thus, sampling from the posterior distribution allows for simultaneous estimation of quantities of interest and quantifying the associated uncertainty.

A challenge of the hierarchical Bayesian approach is that the posterior distribution will usually not have a closed form, in which case approximation techniques become necessary. In light of this, an indirect sampling-based approximation often is used to explore the posterior distribution. Since the seminal work of Gelfand and Smith [17], Markov chain Monte Carlo (MCMC), particularly Gibbs sampling [22], has become the predominant technique for Bayesian computation. Several MCMC methods for sampling the posterior distributions obtained from inverse problems have been proposed in the literature [5, 16, 33, 1, 7]. However, these methods can be computationally expensive on large-scale problems due to the need to factorize a large covariance matrix at each iteration, though there are cases in which the choice of the prior and the forward operator lead to a reduction in computational cost [6]. Approximating complex, non-Gaussian posteriors without the computational intensity of MCMC is still an ongoing area of research, e.g., variational Bayes [32] and integrated nested Laplace approximation [50]. Each approach has features and caveats, a full exposition of which is beyond the scope of this paper. In this work, we assume that a researcher has already decided that they will use MCMC to access the posterior distribution.

Our aim in this work is to address the computational burden posed by repeatedly sampling high-dimensional Gaussian random variables as part of a larger MCMC routine, e.g., block Gibbs [36] or one-block [49]. We do so by leveraging the low-rank structure of forward models typically encountered in linear inverse problems. Specifically, we propose a Metropolis–Hastings independence sampler in which the proposal distribution, based on a low-rank approximation to the prior-preconditioned Hessian, is easy to construct and to sample. We also develop a proposal distribution using a randomized approach for computing the low-rank approximation when doing so directly is computationally expensive. We derive explicit formulas for the acceptance rates of our proposed approaches and analyze their statistical properties. We provide a detailed description of the computational costs. Numerical experiments support the theoretical properties of our proposed approaches and demonstrate the computational benefits over standard block Gibbs sampling.

The rest of the paper is organized as follows. In section 2, we formulate a general linear inverse problem in the hierarchical Bayesian framework, with particular attention paid to the computational bottleneck arising in standard MCMC samplers. In section 3, we present our proposed approach of using low-rank approximation as the basis of an independence sampler to accelerate drawing realizations from high-dimensional Gaussian distributions. In section 4, we demonstrate the performance of our approach on simulated examples in image deblurring and CT reconstruction via Metropolis–Hastings-within-Gibbs sampling [39]. The paper concludes with a discussion in section 5 and proofs of stated results in the appendices. Further numerical studies, including convergence and alternative parameterizations for MCMC, are presented in supplementary material to this paper.

2. The Bayesian statistical inverse problem. Assume that the observed data are corrupted by additive noise so that the stochastic model for the forward problem is

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the forward operator, or the parameter-to-observation map, $\boldsymbol{\epsilon}$ is the measurement error, and $\boldsymbol{x}$ is the underlying quantity that we wish to reconstruct. We suppose that $\boldsymbol{\epsilon}$ is a Gaussian random variable with mean zero and covariance $\mu^{-1}\boldsymbol{I}$, independent of the unknown $\boldsymbol{x}$. In some applications, μ may be known. Quite often, however, it is unknown and we assume that is the case here. Under this model, $\boldsymbol{b} \mid \boldsymbol{x}, \mu \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \mu^{-1}\boldsymbol{I})$ so that the likelihood is

$$f(\boldsymbol{b} \mid \boldsymbol{x}, \mu) \;\propto\; \mu^{m/2} \exp\left(-\frac{\mu}{2}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})\right),\ \boldsymbol{b}\in\mathbb{R}^m. \tag{2}$$

The prior distribution for $\boldsymbol{x}$ encodes the structure we expect or wish to enforce on $\boldsymbol{x}$ before taking observed data into account. An often reasonable prior for $\boldsymbol{x}$ is Gaussian with mean zero and covariance $\sigma^{-1}\boldsymbol{\Gamma}_{\text{pr}} \equiv \sigma^{-1}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}$; i.e.,

$$\pi(\boldsymbol{x} \mid \sigma) \;\propto\; \sigma^{n/2} \exp\left(-\frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Gamma}_{\text{pr}}^{-1}\boldsymbol{x}\right),\ \boldsymbol{x}\in\mathbb{R}^n, \tag{3}$$

where the covariance matrix $\boldsymbol{\Gamma}_{\text{pr}}$ is assumed known up to the precision σ.

Different covariance matrices may be chosen depending on what structure one wishes to enforce on the estimand $\boldsymbol{x}$. The prior structure we use in our numerical experiments (section 4) is motivated by Gaussian Markov random fields (GMRFs) [49]. Other popular choices involve Gaussian processes [45], which are parameterized in terms of covariance kernels.

We assume in this work that $\boldsymbol{\Gamma}_{\text{pr}}$ is fixed up to a multiplicative constant. This makes available an a priori factorization that we use to construct a low-rank approximation. However, Gaussian process covariance kernels typically depend on parameters other than the multiplicative precision. For example, the covariance matrix of a Gaussian process often takes the form $\boldsymbol{\Gamma}_{\text{pr}} = \sigma\boldsymbol{R}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ may be the correlation length or other parameters determining the covariance function. When $\boldsymbol{\theta}$ is unknown, one can assign it a prior distribution and estimate it along with the other parameters in the model by, e.g., updating it on each iteration of an MCMC algorithm. Such repeated updates are not feasible for extremely high-dimensional problems since each factorization $\boldsymbol{\Gamma}_{\text{pr}}^{-1} = \boldsymbol{L}^\top\boldsymbol{L}$ is too expensive. However, it is possible to assign a prior to $\boldsymbol{\theta}$ and subsequently obtain an empirical Bayes estimate $\widehat{\boldsymbol{\theta}}$ (e.g., marginal posterior mode as done in [44]). This estimator can be plugged in to the covariance function so that $\boldsymbol{R}(\widehat{\boldsymbol{\theta}})$ is fixed.

Conditional on μ and σ, the Bayesian inverse problem as formulated in (2) and (3) yields $\boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma \sim \mathcal{N}(\boldsymbol{x}_{\text{cond}}, \boldsymbol{\Gamma}_{\text{cond}})$, where $\boldsymbol{\Gamma}_{\text{cond}} = (\mu\boldsymbol{A}^\top\boldsymbol{A} + \sigma\boldsymbol{\Gamma}_{\text{pr}}^{-1})^{-1}$ and $\boldsymbol{x}_{\text{cond}} = \mu\boldsymbol{\Gamma}_{\text{cond}}\boldsymbol{A}^\top\boldsymbol{b}$; i.e.,

$$\pi(\boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma) \propto \exp\left(-\frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 - \frac{\sigma}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2\right). \tag{4}$$

The conditional posterior mode, $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}\in\mathbb{R}^n} \pi(\boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma)$, is the minimizer of the negative log-likelihood $(\mu/2)\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + (\sigma/2)\|\boldsymbol{L}\boldsymbol{x}\|_2^2$ and thus corresponds to Tikhonov regularization in the deterministic linear inverse problem. In a fully Bayesian analysis in

Algorithm 1: An outline of the standard block Gibbs algorithm for sampling the posterior density (5).

Input: Full conditional distributions of $\boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma$ and $(\mu, \sigma) \mid \boldsymbol{x}, \boldsymbol{b}$, sample size N, burn-in period N_b.

Output: Approximate sample from the posterior distribution (5), $\{\boldsymbol{x}_{(t)}, \mu_{(t)}, \sigma_{(t)}\}_{t=N_b+1}^{N}$.

Initialize $\boldsymbol{x}_{(0)}$, $\mu_{(0)}$, and $\sigma_{(0)}$.
for $t = 1$ *to* N **do**
  Draw $\boldsymbol{x}_{(t)} \sim \mathcal{N}(\mu_{(t-1)}\boldsymbol{\Gamma}_{\text{cond}}^{(t)}\boldsymbol{A}^\top\boldsymbol{b}, \boldsymbol{\Gamma}_{\text{cond}}^{(t)})$, where $\boldsymbol{\Gamma}_{\text{cond}}^{(t)} = (\mu_{(t-1)}\boldsymbol{A}^\top\boldsymbol{A} + \sigma_{(t-1)}\boldsymbol{\Gamma}_{\text{pr}}^{-1})^{-1}$.
  Draw $(\mu_{(t)}, \sigma_{(t)}) \sim \mu, \sigma \mid \boldsymbol{x}_{(t)}, \boldsymbol{b}$.
end

which μ and σ are unknown, we assign them a prior $\pi(\mu, \sigma)$ so that they can be estimated along with other parameters. In this case, the joint posterior density becomes

$$\pi(\boldsymbol{x}, \mu, \sigma \mid \boldsymbol{b}) \propto \mu^{m/2}\sigma^{n/2} \exp\left(-\frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \frac{\sigma}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2\right) \pi(\mu, \sigma). \tag{5}$$

In section 4, we consider two different priors on the precision parameters: conditionally conjugate Gamma distributions and a so-called weakly informative prior.

With priors on the precision components, the full posterior distribution is no longer Gaussian and generally not available in closed form. Non-Gaussian posteriors can sometimes be approximated by a Gaussian distribution, but such an approximation can be poor, especially with high-dimensional parameter spaces or multimodal posterior distributions [19, Chapter 4]. Thus we appeal to MCMC for sampling from the posterior distribution. A version of the basic block Gibbs sampler for sampling from (5) is given in Algorithm 1. Most often, μ and σ are updated individually (especially when using conditionally conjugate Gamma priors), but this is not necessary. Typically, $\boldsymbol{x}$ is drawn separately from (μ, σ) to take advantage of its conditionally conjugate Gaussian distribution.

For any iterative sampling algorithm in the Bayesian linear inverse problem, the computational cost per iteration is dominated by sampling $\boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma$ in (4). While sampling from this Gaussian distribution is a very straightforward procedure, the fact that it is high-dimensional makes it very computationally intensive. To circumvent the computational burden, we substitute direct sampling with a Metropolis–Hastings independence sampler using a computationally cheap low-rank proposal distribution. We present our proposed approach in section 3.

3. Independence sampling with low-rank proposals.

Here we briefly review independence sampling and discuss a proposal distribution that uses a low-rank approximation to efficiently generate samples from (4).

3.1. Independence sampling.

Let $\boldsymbol{\theta} \in \mathbb{R}^n$ and denote the (possibly unnormalized) target density by $h(\boldsymbol{\theta})$. The Metropolis–Hastings algorithm [37, 26] proceeds iteratively by generating

at iteration t a draw, $\boldsymbol{\theta}_*$, from an available proposal distribution possibly conditioned on the current state, $\boldsymbol{\theta}_{(t-1)}$, and setting $\boldsymbol{\theta}_{(t)} = \boldsymbol{\theta}_*$ with probability $\alpha(\boldsymbol{\theta}_{(t-1)}, \boldsymbol{\theta}_*) = h(\boldsymbol{\theta}_*)q(\boldsymbol{\theta}_{(t-1)} \mid \boldsymbol{\theta}_*)/(h(\boldsymbol{\theta}_{(t-1)})q(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{(t-1)})) \wedge 1$, where $q(\cdot \mid \boldsymbol{\theta}_{(t-1)})$ is the density of the proposal distribution. This algorithm produces a Markov chain $\{\boldsymbol{\theta}_{(t)}\}$ with transition kernel

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}_*)q(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}) + \delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}_*)\left(1 - \int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})d\boldsymbol{\theta}'\right),$$

where $\delta_{\boldsymbol{\theta}}(\cdot)$ is the point mass at $\boldsymbol{\theta}$. Properties of the Metropolis–Hastings algorithm, including convergence to the target distribution, may be found in [47] and elsewhere.

An independence Metropolis–Hastings sampler (IMHS) proposes states from a density that is independent of the current state of the chain. The proposal has density $q(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{(t-1)}) \equiv g(\boldsymbol{\theta}_*)$, and the ratio appearing in $\alpha(\boldsymbol{\theta}_{(t-1)}, \boldsymbol{\theta}_*)$ can be written as

$$\frac{h(\boldsymbol{\theta}_*)g(\boldsymbol{\theta}_{(t-1)})}{h(\boldsymbol{\theta}_{(t-1)})g(\boldsymbol{\theta}_*)} = \frac{w(\boldsymbol{\theta}_*)}{w(\boldsymbol{\theta}_{(t-1)})}, \tag{6}$$

where $w(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta})/g(\boldsymbol{\theta})$. The IMHS is similar to the rejection algorithm. The rejection algorithm draws a candidate value $\boldsymbol{\theta}_*$ from an available generating distribution with density g such that for some $M \geq 1$, $h(\boldsymbol{\theta}) \leq Mg(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. It then accepts the draw with probability $h(\boldsymbol{\theta}_*)/Mg(\boldsymbol{\theta}_*)$. Rejection sampling results in an exact draw from the target distribution.

For both the IMHS and the rejection sampler, it is desirable for g to match the target density as closely as possible and, hence, to have an acceptance rate as high as possible. At least, g should generally follow h, but with tails that are no lighter than h [20, 47]. These guidelines are in contrast to those prescribed for the more common random walk Metropolis–Hastings, in which the best convergence is generally obtained with acceptance rates between 20% and 50% [20, 48]. In what follows, we discuss our proposed generating distribution, both as an independence sampler as well as its use in a rejection algorithm.

3.2. Approximating the target distribution. Samples from the conditional distribution $\mathcal{N}(\boldsymbol{x}_{\text{cond}}, \boldsymbol{\Gamma}_{\text{cond}})$ can be generated as $\boldsymbol{x} = \boldsymbol{x}_{\text{cond}} + \boldsymbol{G}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{G}$ satisfies $\boldsymbol{\Gamma}_{\text{cond}} = \boldsymbol{G}\boldsymbol{G}^\top$. Forming the mean $\boldsymbol{x}_{\text{cond}}$ and computing the random vector $\boldsymbol{G}\boldsymbol{\varepsilon}$ involve expensive operations with the covariance matrix. By leveraging the low-rank nature of the forward operator $\boldsymbol{A}$, we can construct a fast proposal distribution for an independence sampler.

Consider the covariance matrix $\boldsymbol{\Gamma}_{\text{cond}} = (\mu\boldsymbol{A}^\top\boldsymbol{A} + \sigma\boldsymbol{L}^\top\boldsymbol{L})^{-1}$. Factorizing this matrix so that

$$\boldsymbol{\Gamma}_{\text{cond}} = \boldsymbol{L}^{-1}(\mu\boldsymbol{L}^{-\top}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{L}^{-1} + \sigma\boldsymbol{I})^{-1}\boldsymbol{L}^{-\top} \tag{7}$$

yields the so-called *prior-preconditioned Hessian transformation* $\boldsymbol{H} := \boldsymbol{L}^{-\top}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{L}^{-1}$ [11, 15, 43, 52]. For highly ill-posed inverse problems such as those considered here, $\boldsymbol{A}$ either has a rapidly decaying spectrum or is rank deficient. The product singular value inequalities [28, Theorem 3.3.16 (b)] ensure that $\boldsymbol{A}\boldsymbol{L}^{-1}$ has the same rank as $\boldsymbol{A}$ and the same rate of decay of singular values. A detailed discussion on the low-rank approximation of the prior-preconditioned Hessian is provided in [15, section 3].

We approximate $\boldsymbol{H}$ using a truncated eigenvalue decomposition,

$$\boldsymbol{L}^{-\top}\boldsymbol{A}^{\top}\boldsymbol{A}\boldsymbol{L}^{-1} \approx \boldsymbol{V}_k\boldsymbol{\Lambda}_k\boldsymbol{V}_k^{\top}, \tag{8}$$

where $\boldsymbol{V}_k \in \mathbb{R}^{n\times k}$ has orthonormal columns and $\boldsymbol{\Lambda}_k \in \mathbb{R}^{k\times k}$ is the diagonal matrix containing the $k \le n$ largest eigenvalues of $\boldsymbol{H}$. If $\operatorname{rank}(\boldsymbol{A}) = k$, then exact equality holds. The truncation parameter k controls the trade-off between accuracy on the one hand and computational and memory costs on the other.

We approximate the conditional covariance matrix $\boldsymbol{\Gamma}_{\text{cond}}$ by substituting (8) into (7),

$$\widehat{\boldsymbol{\Gamma}}_{\text{cond}} \;\equiv\; \boldsymbol{L}^{-1}\frac{1}{\sigma}\left(\boldsymbol{I} + \frac{\mu}{\sigma}\boldsymbol{V}_k\boldsymbol{\Lambda}_k\boldsymbol{V}_k^{\top}\right)^{-1}\boldsymbol{L}^{-\top}. \tag{9}$$

Using the Woodbury identity and the fact that $\boldsymbol{V}_k$ has orthonormal columns, the right-hand side of (9) becomes

$$\widehat{\boldsymbol{\Gamma}}_{\text{cond}} \;=\; \frac{1}{\sigma}\boldsymbol{L}^{-1}(\boldsymbol{I} - \boldsymbol{V}_k\boldsymbol{D}_k\boldsymbol{V}_k^{\top})\boldsymbol{L}^{-\top}, \qquad \boldsymbol{D}_k = \operatorname{diag}\left(\mu\lambda_j(\mu\lambda_j+\sigma)^{-1} : j = 1,\dots,k\right) \in \mathbb{R}^{k\times k},$$

where λ_j, $j = 1,\dots,k$, are the diagonals of $\boldsymbol{\Lambda}_k$. To approximate the mean $\boldsymbol{x}_{\text{cond}}$, replace $\boldsymbol{\Gamma}_{\text{cond}}$ by $\widehat{\boldsymbol{\Gamma}}_{\text{cond}}$ so that $\hat{\boldsymbol{x}}_{\text{cond}} = \mu\widehat{\boldsymbol{\Gamma}}_{\text{cond}}\boldsymbol{A}^{\top}\boldsymbol{b}$. With these approximations, the proposal distribution for our proposed independence sampler is $\mathcal{N}(\hat{\boldsymbol{x}}_{\text{cond}}, \widehat{\boldsymbol{\Gamma}}_{\text{cond}})$. Optimality of this low-rank approximation was studied in [57].

A factorization of the form $\widehat{\boldsymbol{\Gamma}}_{\text{cond}} = \boldsymbol{G}\boldsymbol{G}^{\top}$ can be used to sample from $\mathcal{N}(\hat{\boldsymbol{x}}_{\text{cond}}, \widehat{\boldsymbol{\Gamma}}_{\text{cond}})$. It can be verified that $\boldsymbol{G} := \sigma^{-1/2}\boldsymbol{L}^{-1}(\boldsymbol{I} - \boldsymbol{V}_k\widehat{\boldsymbol{D}}_k\boldsymbol{V}_k^{\top})$, with $\widehat{\boldsymbol{D}}_k = \boldsymbol{I} \pm (\boldsymbol{I} - \boldsymbol{D}_k)^{1/2}$, satisfies $\widehat{\boldsymbol{\Gamma}}_{\text{cond}} = \boldsymbol{G}\boldsymbol{G}^{\top}$. Since $\widehat{\boldsymbol{D}}_k$ is diagonal and $k \ll n$, we obtain a computationally cheap way of generating draws from the high-dimensional proposal distribution $\mathcal{N}(\hat{\boldsymbol{x}}_{\text{cond}}, \widehat{\boldsymbol{\Gamma}}_{\text{cond}})$. Then we can use a Metropolis–Hastings step to correct for the approximation. This results in our proposed low-rank independence sampler (LRIS).

3.3. Analysis of acceptance ratio. Here, we derive an explicit formula for evaluating the acceptance ratio for our proposed algorithm and provide insight into the conditions under which the proposal distribution closely approximates the target distribution. For simplicity of notation, we suppress the conditioning on $\boldsymbol{b}, \mu$, and σ.

The target density is

$$h(\boldsymbol{x}) \;:=\; \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Gamma}_{\text{cond}})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_{\text{cond}})^{\top}\boldsymbol{\Gamma}_{\text{cond}}^{-1}(\boldsymbol{x} - \boldsymbol{x}_{\text{cond}})\right), \tag{10}$$

and the proposal density, $g(\boldsymbol{x})$, replaces $\boldsymbol{x}_{\text{cond}}$ with $\hat{\boldsymbol{x}}_{\text{cond}}$ and $\boldsymbol{\Gamma}_{\text{cond}}$ with $\widehat{\boldsymbol{\Gamma}}_{\text{cond}}$ in $h(\boldsymbol{x})$. The following result gives a practical way to compute the acceptance ratio. It can be verified with a little algebra, so the proof is omitted.

Proposition 1. *Let $\boldsymbol{x}$ be the current state of the LRIS chain, and let $\boldsymbol{z}$ be the proposed state. Then the acceptance ratio can be computed as $\eta(\boldsymbol{z}, \boldsymbol{x}) = w(\boldsymbol{z})/w(\boldsymbol{x})$, where $w(\boldsymbol{x}) = \exp\left(-\boldsymbol{x}^{\top}(\boldsymbol{\Gamma}_{cond}^{-1} - \widehat{\boldsymbol{\Gamma}}_{cond}^{-1})\boldsymbol{x}/2\right)$.*

An efficient implementation and the cost of computing this ratio is discussed in subsection 3.5. The quality of the low-rank approximation to the target distribution can be seen through the acceptance ratio.

Proposition 2. *Let $\boldsymbol{x}$ be the current state of the LRIS chain, and let $\boldsymbol{z}$ be the proposed state. Then the LRIS acceptance ratio can be expressed as*

$$\eta(\boldsymbol{z},\boldsymbol{x}) = \exp\left(-\frac{\mu}{2}\sum_{j=k+1}^{n}\lambda_j\left[\left(\boldsymbol{v}_j^\top\boldsymbol{L}\boldsymbol{z}\right)^2-\left(\boldsymbol{v}_j^\top\boldsymbol{L}\boldsymbol{x}\right)^2\right]\right). \tag{11}$$

Proof. See Appendix A. ∎

This proposition asserts that the acceptance ratio is high when either μ is small or the discarded eigenvalues $\{\lambda_j\}_{j=k+1}^n$ are small. The dependence of the acceptance ratio on the eigenvectors can be seen explicitly by writing $(\boldsymbol{v}_j^\top\boldsymbol{L}\boldsymbol{z})^2-(\boldsymbol{v}_j^\top\boldsymbol{L}\boldsymbol{x})^2 = [\boldsymbol{v}_j^\top\boldsymbol{L}(\boldsymbol{z}+\boldsymbol{x})][\boldsymbol{v}_j^\top\boldsymbol{L}(\boldsymbol{z}-\boldsymbol{x})]$. Thus, if $\boldsymbol{z}\pm\boldsymbol{x}\perp\boldsymbol{L}^\top\boldsymbol{v}_j,\ j=k+1,\dots,n$, then the acceptance ratio is 1.

While Proposition 2 provides insight into realizations of the acceptance ratio, the actual acceptance ratio is a random variable. The expected behavior and variability of this quantity can be understood through Theorem 1. To this end, define the constants

$$N_\ell := \exp\left(\frac{\mu^2}{2\sigma}\sum_{j=k+1}^{n}\frac{\ell\mu\lambda_j}{\ell\mu\lambda_j+\sigma}(\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{L}^{-1}\boldsymbol{v}_j)^2\right)\prod_{j=k+1}^{n}\left(1+\frac{\ell\mu}{\sigma}\lambda_j\right)^{1/2},\quad \ell=1,2,\dots. \tag{12}$$

Theorem 1. *Let $\boldsymbol{x}$ be the current state of the LRIS chain, and let $\boldsymbol{z}$ be the proposed state. Then*

$$\begin{aligned} e_\eta &:= \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\eta(\boldsymbol{z},\boldsymbol{x})] = \frac{1}{N_1 w(\boldsymbol{x})},\\ v_\eta^2 &:= \mathbb{V}_{\boldsymbol{z}|\boldsymbol{x}}[\eta(\boldsymbol{z},\boldsymbol{x})] = \frac{1}{w^2(\boldsymbol{x})}\left(\frac{1}{N_2}-\frac{1}{N_1^2}\right), \end{aligned} \tag{13}$$

where $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}(\cdot)$ denotes expectation conditional on $\boldsymbol{x}$, $\mathbb{V}_{\boldsymbol{z}|\boldsymbol{x}}(\cdot)$ denotes the variance conditional on $\boldsymbol{x}$, and $w(\boldsymbol{x})$ is as defined in Proposition 1.

Proof. See Appendix A. ∎

Using this result, a straightforward application of Chebyshev's inequality [46] shows that for any $\epsilon>0$, $\Pr_{\boldsymbol{z}|\boldsymbol{x}}\left(\left|\eta(\boldsymbol{z},\boldsymbol{x})-(N_1w(\boldsymbol{x}))^{-1}\right|\geq\epsilon\right)\leq\left(N_2^{-1}-N_1^{-2}\right)/[\epsilon^2w^2(\boldsymbol{x})]$, where $\Pr_{\boldsymbol{z}|\boldsymbol{x}}(\cdot)$ denotes probability conditional on the current state $\boldsymbol{x}$. Thus, we can construct conditional prediction intervals about the realized acceptance rate. For instance, at any given state $\boldsymbol{x}$, $\Pr_{\boldsymbol{z}|\boldsymbol{x}}(\eta(\boldsymbol{z},\boldsymbol{x})\in[e_\eta\pm 4.47v_\eta])\geq 0.95$. In Appendix A, we derive expressions for all moments of the acceptance ratio.

It is clear from Theorem 1 that if the eigenvalues $\{\lambda_j\}_{j=k+1}^n$ are zero, then the acceptance probability is 1. Likewise, if the eigenvalues are nonzero but small in magnitude, then the acceptance rate is close to 1. Further, consider the SVD of $\boldsymbol{A}\boldsymbol{L}^{-1}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Then $\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{L}^{-1}\boldsymbol{v}_j$ in (12) is equal to $\sigma_j\boldsymbol{b}^\top\boldsymbol{u}_j$, where $\boldsymbol{u}_j$ is the jth singular vector of $\boldsymbol{A}\boldsymbol{L}^{-1}$. Thus, the acceptance

rate may be close to 1 even if the components of the measurement $\boldsymbol{b}$ along the left singular vectors of $\boldsymbol{A}\boldsymbol{L}^{-1}$ are small. This is closely related to filter factors that are used to analyze deterministic inverse problems [25].

These results establish the moments of the acceptance rate for fixed precision parameters μ and σ and fixed rank k of the proposal distribution. In practice, when running MCMC, μ and σ will change on each iteration, meaning that the actual acceptance rate will vary from one iteration to the next. Thus, it may not be clear a priori which truncation level to use to achieve an acceptable acceptance rate while minimizing the computational cost. Of course, if the low-rank matrix is obtained from a rank-deficient forward model by discarding only the zero eigenvalues, then the acceptance rate is one for all μ and σ. Otherwise, a practitioner can employ an *adaptive* LRIS in which the acceptance rate is tracked during an initial burn-in period, adding rank to the distribution every, say, 100 iterations if the acceptance rate is too low. This allows finding the minimum number of eigenvalues needed to achieve high acceptance over the high probability region of μ and σ. Provided the adaptation stops after a finite number of iterations, convergence to the stationary distribution is still guaranteed [13]. An outline of the adaptive LRIS approach, along with practical guidelines to determine the target rank k, is given in the supplementary material.

Convergence and the rejection algorithm. Our proposed candidate generating distribution $g(\boldsymbol{x})$ bounds the target distribution up to a fixed constant as a function of the remaining eigenvalues in the low-rank approximation, as asserted by the next proposition.

Proposition 3. *The target density $h(\boldsymbol{x})$ (10) and the proposal density $g(\boldsymbol{x})$ can be bounded as $h(\boldsymbol{x}) \leq N_1 g(\boldsymbol{x})$ for all $\boldsymbol{x}$, where $N_1 \geq 1$ is given in (12).*

Proof. See Appendix A. ∎

Proposition 3 establishes that the subchain produced by our proposed sampler has stationary distribution $\pi(\cdot \mid \boldsymbol{b}, \mu, \sigma)$ and is uniformly ergodic by [47, Theorem 7.8]; i.e., for $p \in \mathbb{N}$,

$$\|K^p(\boldsymbol{x}, \cdot) - \pi(\cdot \mid \boldsymbol{b}, \mu, \sigma)\|_{TV} \leq 2\left(1 - N_1^{-1}\right)^p \quad \forall \boldsymbol{x} \in \operatorname{supp} \pi, \tag{14}$$

where $K^p(\boldsymbol{x}, \cdot)$ is the p-step LRIS transition kernel starting from $\boldsymbol{x}$ and $\|\cdot\|_{TV}$ denotes the total variation norm. Thus, if one runs several subiterations of the LRIS, the realizations will converge to a draw from the true full conditional distribution at a rate independent of the initial state. Convergence is faster as the remaining eigenvalues from the low-rank approximation become small, and is immediate when the remaining eigenvalues are zero.

Equation (14) explicitly quantifies convergence of the subchain to the full conditional distribution as a function of the quality of the approximation to the target, quantified in N_1. However, when the LRIS is used inside a larger MCMC algorithm (e.g., Metropolis–Hastings-within-Gibbs), convergence of the entire Markov chain to its stationary distribution is affected not only by the LRIS proposal distribution, but also by modeling choices on the remaining parameters and the manner in which they are updated. There exist results for establishing geometric ergodicity of componentwise Metropolis–Hastings independence samplers and so-called two-stage Metropolis–Hastings-within-Gibbs algorithms [30] for which Proposition 3 could be useful. To the best of our knowledge, though, more general effects of the proposal distribution on the convergence of a Metropolis–Hastings-within-Gibbs algorithm are

unknown. While exploring this issue is beyond the scope of this work, we carry out in the supplementary material an empirical study in which we assess convergence of a Metropolis–Hastings-within-Gibbs chain as a function of the rank of the proposal. We observe that as the number of eigenvalues retained increases, the convergence of the LRIS algorithm becomes more rapid.

Proposition 3 suggests also that the approximating distribution can be used in a rejection algorithm instead of LRIS. The proof of the proposition shows that $\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}}) \geq \det(\boldsymbol{\Gamma}_{\text{cond}})$, but each determinant is a generalized variance [31]. When there are nonzero eigenvalues left out of the low-rank approximation, the proposal density will have heavier tails than the target density, a desirable property for a candidate distribution in a rejection algorithm [19]. Otherwise, the approximation is exact. We remark, however, that for a given candidate density g, LRIS is more efficient than a rejection algorithm in terms of variances of the concomitant estimators [35]. Further, the rejection sampler requires knowledge of N_1, which depends on eigenvalues that may be unavailable.

3.4. Generating low-rank approximations. A major cost of our proposed sampler is in the precomputation associated with constructing the low-rank approximation. The standard approach for computing this low-rank approximation is to use a Krylov subspace solver (e.g., Lanczos method [51]) for computing a partial eigenvalue decomposition. Alternatively, we can compute the rank-k SVD $\boldsymbol{A}\boldsymbol{L}^{-1} \approx \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^\top$. Then the approximate low-rank decomposition can be computed as $\boldsymbol{H} \approx \boldsymbol{V}_k\boldsymbol{\Sigma}_k^2\boldsymbol{V}_k^\top$. Here we discuss a computationally efficient alternative.

Randomized SVD, reviewed in [23], is a computationally efficient approach for computing a low-rank approximation to the prior-preconditioned Hessian $\boldsymbol{H}$. The basic idea of the randomized SVD approach is to draw a random matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times(k+p)}$, where the entries of $\boldsymbol{\Omega}$ are i.i.d. standard Gaussian random variables. Here, k is the target rank and p is an oversampling parameter. An approximation to the column space of $\boldsymbol{H}$ is computed by the matrix product $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\Omega}$. A thin-QR factorization $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{R}$ is computed, and the resulting low-rank approximation to $\boldsymbol{H}$ is given by

$$\boldsymbol{H} \approx \widehat{\boldsymbol{H}} := \boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{H}\boldsymbol{Q}\boldsymbol{Q}^\top. \tag{15}$$

This can be postprocessed to obtain an approximate low-rank decomposition of the form (8). This is summarized in Algorithm 2.

Algorithm 2: Randomized SVD algorithm for computing low-rank decomposition.

Input: Matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ and random matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times(k+p)}$.
Output: Approximate eigenvectors $\boldsymbol{V}$ and approximate eigenvalues $\boldsymbol{\Lambda}$.
1 Compute $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\Omega}$ and thin-QR factorization $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{R}$.
2 Compute $\boldsymbol{T} = \boldsymbol{Q}^\top\boldsymbol{H}\boldsymbol{Q}$ and its eigendecomposition $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$.
3 Compute $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{U}$.

Similar to Theorem 1, we can bound the expected value of the acceptance ratio under the randomized SVD approach.

Theorem 2. *Suppose we compute the low-rank approximation* $\widehat{\boldsymbol{H}}$ *using Algorithm 2 with guess* $\boldsymbol{\Omega} \in \mathbb{R}^{n\times(k+p)}$. *Let* $p \geq 2$ *be the oversampling parameter. Then*

$$\mathbb{E}_{\boldsymbol{\Omega}|\boldsymbol{z},\boldsymbol{x}}\left[\eta(\boldsymbol{z},\boldsymbol{x})\right] \geq \frac{1}{w(\boldsymbol{x})}\exp\left(-\mu\|\boldsymbol{L}\boldsymbol{z}\|_2^2\left[\alpha\lambda_{k+1}+\beta\left(\sum_{j=k+1}^{n}\lambda_j^2\right)^{1/2}\right]\right),$$

where $\alpha = 1+\sqrt{\frac{k}{p-1}}, \beta = \frac{e\sqrt{k+p}}{p}$ *and* $\mathbb{E}_{\boldsymbol{\Omega}|\boldsymbol{z},\boldsymbol{x}}$ *denotes expectation with respect to* $\boldsymbol{\Omega}$ *given the current state* $\boldsymbol{x}$ *and the proposed step* $\boldsymbol{z}$.

Proof. See Appendix A. ∎

The interpretation of this result is similar to Theorem 1. That is, if the eigenvalues of the prior-preconditioned Hessian $\boldsymbol{H}$ are rapidly decaying or zero beyond the index k, then the expected acceptance rate, averaged over all random matrices $\boldsymbol{\Omega}$, is high.

In practice, an oversampling parameter of $p \lesssim 20$ is recommended [23]. As proposed, Algorithm 2 requires $2(k+p)$ matrix-vector products (matvecs) with $\boldsymbol{H}$. The second round of matvecs required in step 2 can be avoided by using the approximation [53, section 2.3]

$$\boldsymbol{T} \approx (\boldsymbol{\Omega}^\top\boldsymbol{Q})^{-1}(\boldsymbol{\Omega}^\top\boldsymbol{Y})(\boldsymbol{Q}^\top\boldsymbol{\Omega})^{-1}.$$

This is an example of the so-called single-pass algorithm. Other single-pass algorithms are discussed in [59]. In practice, the target rank k may not be known, in which case a modified approach may be used to adaptively estimate the subspace [23, Algorithm 4.2].

3.5. Computational costs. Denote the computational cost of a matvec with $\boldsymbol{A}$ by $T_{\boldsymbol{A}}$, and the cost of a matvec with $\boldsymbol{L}$ and $\boldsymbol{L}^{-1}$ as $T_{\boldsymbol{L}}$ and $T_{\boldsymbol{L}^{-1}}$, respectively. For simplicity, we assume the cost of the transpose operations of the respective matrices is the same as that of the original matrix.

It is difficult to accurately estimate the cost of the Krylov subspace method a priori, but the cost is roughly 2 sets of matvecs with $\boldsymbol{A}$ and $\boldsymbol{L}^{-1}$ and an additional $\mathcal{O}(nk^2)$ operations. The quantities $\boldsymbol{A}^\top\boldsymbol{b}$ and $\boldsymbol{L}^{-\top}\boldsymbol{A}^\top\boldsymbol{b}$ can also be precomputed at a cost of $T_{\boldsymbol{A}}$ and $T_{\boldsymbol{A}}+T_{\boldsymbol{L}^{-1}}$ flops, respectively. Generally speaking, this is the same asymptotic cost for randomized SVD. In practice, however, randomized SVD can be much cheaper; see [23] for details.

The cost of computing the mean $\hat{\boldsymbol{x}}_{\text{cond}}$ involves the application of $\boldsymbol{L}^{-1}$ and $(\boldsymbol{I}-\boldsymbol{V}_k\boldsymbol{D}_k\boldsymbol{V}_k^\top)$. This costs $T_{\boldsymbol{L}^{-1}}+4nk$ flops. Similarly, the cost of $\boldsymbol{G}\boldsymbol{\varepsilon}$ is also $T_{\boldsymbol{L}^{-1}}+4nk$ flops. The important point here is that generating a sample from the proposal distribution does not require a matvec with $\boldsymbol{A}$. This is useful for applications in which $T_{\boldsymbol{A}}$ can be extremely high. The computational cost of computing the acceptance ratio can be examined in light of Proposition 1. On each iteration, the weight $w(\boldsymbol{x})$ will already be available from the previous iteration, so we only need to compute $w(\boldsymbol{z})$. We can simplify this expression as $\log w(\boldsymbol{z}) = -\mu\boldsymbol{z}^\top(\boldsymbol{A}^\top\boldsymbol{A}-\boldsymbol{L}^\top\boldsymbol{V}_k\boldsymbol{\Lambda}_k\boldsymbol{V}_k^\top\boldsymbol{L})\boldsymbol{z}$, which requires one matvec with $\boldsymbol{A}$ and $\boldsymbol{L}$ each, two inner products and $4n$ flops, and an additional $2nk$ flops. Aside from the precomputational cost of the low-rank factorization, only the evaluation of the acceptance ratio requires accessing the forward operator $\boldsymbol{A}$. The resulting costs are summarized in Table 1.

Table 1
Summary of computational costs of various steps in the LRIS.

Operation	Formula	Cost
Precomputation	Equation (8)	$2k(T_{\boldsymbol{A}} + T_{\boldsymbol{L}^{-1}}) + \mathcal{O}(nk^2)$
Computing mean	$\hat{\boldsymbol{x}}_{\text{cond}} = \mu\widehat{\boldsymbol{\Gamma}}_{\text{cond}}\boldsymbol{A}^\top\boldsymbol{b}$	$T_{\boldsymbol{L}^{-1}} + 4nk$
Generating sample	$\boldsymbol{x} \sim \mathcal{N}(\hat{\boldsymbol{x}}_{\text{cond}}, \widehat{\boldsymbol{\Gamma}}_{\text{cond}})$	$T_{\boldsymbol{L}^{-1}} + 4nk$
Acceptance ratio	Proposition 1	$T_{\boldsymbol{A}} + T_{\boldsymbol{L}} + 2n(k+2)$

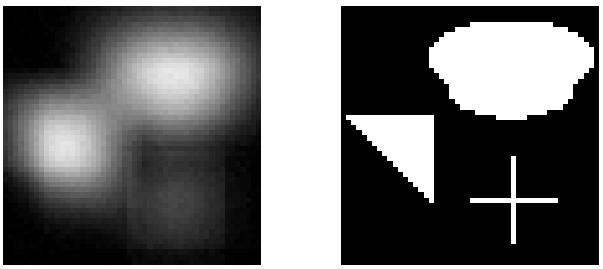

Figure 1. *Observed image (left panel) and true image (right panel) in the 2D image deblurring example.*

4. Illustrations. Here we demonstrate our proposed approach on two simulated examples. The first example is a standard two-dimensional deblurring problem in which we compare the performance of our proposed low-rank independence sampler to conventional block Gibbs sampling to demonstrate the competitive solutions and the ability to access the posterior distribution in an efficient manner. The second example is a more challenging application motivated by medical imaging with a rank-deficient forward model. We apply our proposed approach there to demonstrate feasibility and to consider a different prior on the precisions than the conventional independent conjugate Gammas.

To ensure meaningful inferences based on the MCMC output, it is important to assess whether the Markov chain is sufficiently close to its stationary distribution. It is well known that an MCMC procedure will generally not result in an immediate draw from the target distribution, unless the initial distribution is the stationary distribution. Usually it is not possible to prove that a chain has converged to its limiting distribution, except in special cases (e.g., perfect sampling [14]). However, diagnostic tools can be used to assess whether or not a chain is sufficiently close so that one can safely treat its output as draws from the target distribution. To diagnose convergence, we use (scalar and multivariate) potential scale reduction factors (PSRF/MPSRF) [21, 9], trace plots, and autocorrelation plots. The reader is referred to [47, Chapter 12], [13, Chapter 3], or [19, Chapter 11] for further discussions of convergence diagnostics for MCMC.

4.1. 2D image deblurring. We take as our target image a 50×50 pixel grayscale image of geometric shapes so that $n = 2{,}500$ in (3). We blur the image by convolution with a Gaussian point spread function. The forward model $\boldsymbol{A}$ and true image $\boldsymbol{x}$ are created using the `Regularization Tools` package [24]. The data are generated by adding Gaussian noise with variance $0.01^2\|\boldsymbol{Ax}\|_\infty^2$. Figure 1 displays the target image and the noisy data.

We model smoothness on $\boldsymbol{x}$ a priori by taking $\boldsymbol{L} = -\Delta + \delta\boldsymbol{I}$ in (3), where $-\Delta$ is the

discrete Laplacian and δ is a small constant to ensure positive definiteness [34]. For the prior and noise precision parameters, we assign a vague Gamma prior, Gamma$(0.1, 0.1)$, which approximates the scale invariant objective prior while maintaining conditional conjugacy. We compute the eigenvalues of the prior preconditioned Hessian matrix $\boldsymbol{H}$ via SVD to determine an appropriate cutoff. Figure 7 (discussed further below) indicates rapid decay within the first few eigenvalues, followed by a smoother decay, and another sharp decay. We use the first $k = 500$ eigenvalues of the matrix to construct our low-rank approximation. We analyze below the effect of truncation level on the acceptance rate of the sampling algorithm. For comparison, we also compute a low-rank approximation using the randomized SVD approach described in subsection 3.4.

Convergence and UQ metrics. We implement a Metropolis–Hastings-within-Gibbs algorithm in which step 3 of Algorithm 1 is substituted with our proposed low-rank independence sampler (LRIS) presented in section 3. Three different chains are run in parallel, with each chain initialized by drawing $\boldsymbol{x}, \mu$, and σ randomly from their prior distributions. Each chain is run for $N = 50{,}000$ iterations, with the first 25,000 iterations discarded as a burn-in period. For comparison, the Gibbs sampler is implemented identically to the low-rank procedure with three independent chains run in parallel with widely dispersed initial values. All simulations are done in MATLAB running on OS X Yosemite (8GB RAM, Intel Core i5 2.66GHz processor).

Figures 2 and 3 display trace plots and autocorrelation functions, respectively, for the last 25,000 iterations of the μ and σ chains for both ordinary block Gibbs sampling and our proposed algorithm. As is known to occur with block Gibbs sampling in high-dimensional linear inverse problems [5], we observe near independence within the μ chains and strong autocorrelation in the σ chains. Despite the high autocorrelation, we still are able to achieve approximate convergence and a sufficient effective sample size (ESS) from the σ chains by running each chain long enough. By combining the three independent chains after approximate convergence, we effectively triple the ESS and thus the number of independent pieces of information available about the target posterior. Thinning the chains to, e.g., every 10th, 50th, or 100th draw would dramatically reduce the autocorrelation of the chains. However, it was argued by Carlin and Louis [13] that such thinning is not necessary and does not improve estimates of quantities of interest. Figure 4 illustrates the approximate convergence of the ergodic averages $\widehat{\mu}_{(n)} = n^{-1}\sum_{t=1}^{n}\mu_{(t)}$ and $\widehat{\sigma}_{(n)} = n^{-1}\sum_{t=1}^{n}\sigma_{(t)}$, $n = 1, \ldots, 25{,}000$, despite the high autocorrelation of the σ chain. The limiting values from both approaches closely agree.

While assessing convergence of high-dimensional parameters is more difficult than for scalar quantities, we can track realized values of the data-misfit part of the log-likelihood as a proxy for monitoring convergence. These realizations also should settle down as the chain approaches the target distribution. Figure 4 displays these plots for both algorithms along with the multivariate PSRFs. Again, we see consistency between ordinary block Gibbs and our own approach, as well as approximate convergence according to the rule of thumb that the PSRF should be approximately less than or equal to 1.1 [19].

The advantage of using the LRIS approach is clear in Table 2, which displays the total wall time to complete the 50,000 MCMC iterations for both block Gibbs and our proposed low-rank sampling approach. Table 2 also displays the *cost per effective sample* (CES), defined

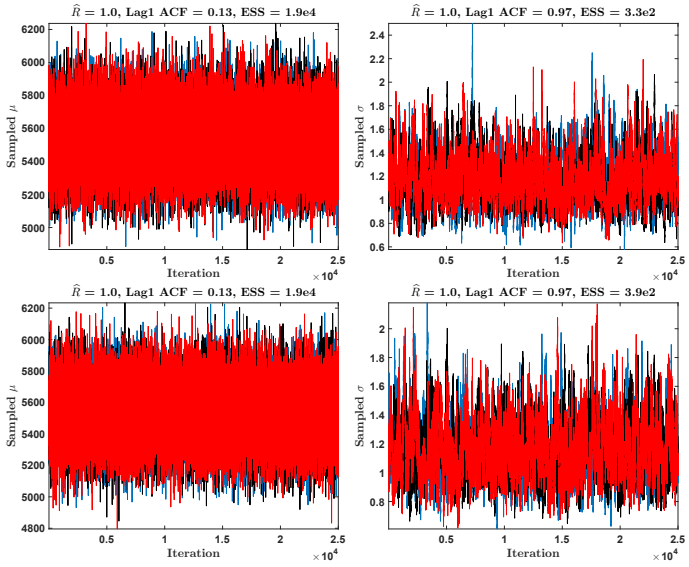

Figure 2. *Trace plots of μ (left) and σ (right) obtained from the MCMC output of both block Gibbs sampling (top row) and the LRIS-based algorithm (bottom row), where each of the three colors represents a different chain. The potential scale reduction factors ($\widehat{R}$) are displayed above each plot, along with the lag 1 autocorrelation coefficients and effective sample size (ESS) estimated from one chain each under both approaches.*

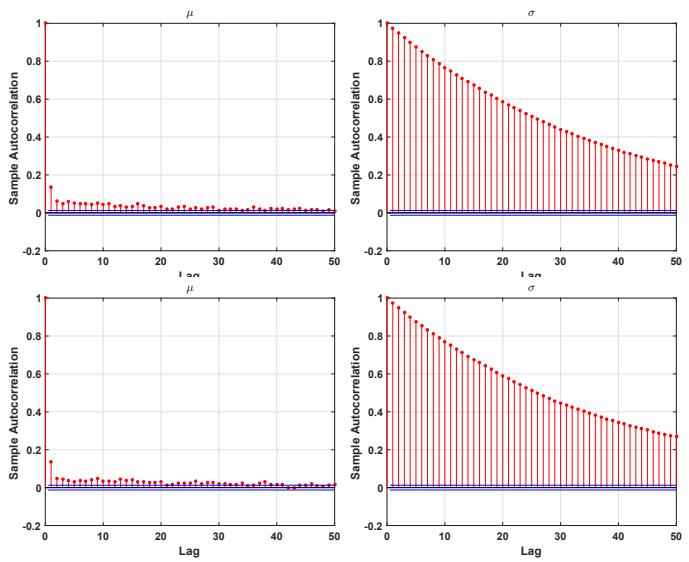

Figure 3. *Estimated autocorrelation functions of μ (left) and σ (right) obtained from one of the chains each under block Gibbs sampling (top row) and our proposed LRIS algorithm (bottom row).*

as the total computation time divided by the effective sample size [16], for one of the σ chains obtained under both algorithms as well as the randomized SVD approach. CES is a measure of the average computational effort required between effectively independent draws. The LRIS approach yields a 76% reduction in computation time compared to the standard block

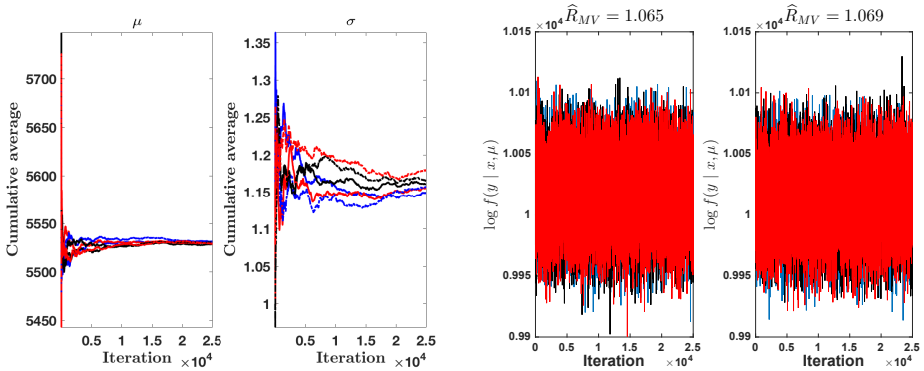

Figure 4. *Cumulative averages of the μ chains (far left) and σ chains (middle left) obtained from the MCMC output in the 2D deblurring example. The dotted lines represent the three chains from block Gibbs; the solid lines correspond to the LRIS algorithm output. Plot of the data-misfit part of the log-likelihood values calculated from the MCMC output of both block Gibbs sampling (middle right) and our low-rank independence sampling algorithm (far right), where each of the three colors corresponds to a different chain. The multivariate potential scale reduction factors are displayed above.*

Table 2

Total wall time to complete 50,000 *MCMC iterations under block Gibbs sampling and the LRIS approach for the 2D deblurring example, along with the estimated cost per effective sample (CES) for one of the σ chains in each case.*

Algorithm	Wall time (s)	CES for σ
Block Gibbs	27907	84.30
LRIS	5134	13.20
LRIS (randomized SVD)	5307	13.86

Gibbs sampler, along with an approximate 80% reduction in computational effort between independent draws of σ. The average acceptance rate over the three chains using our low-rank proposals is 98%, for both "exact" and randomized SVD. The acceptance rate versus rank is discussed further below.

We attain this dramatic reduction in computational effort without sacrificing the quality of posterior inferences, as evident in Figure 5. This figure displays the approximate posterior means of $\boldsymbol{x}$ from both block Gibbs and our low-rank approach. The estimators we obtain with randomized SVD are similar and hence omitted. We show also the (μ, σ) scatterplots and approximate marginal densities obtained from both algorithms in Figure 6, again showing agreement. The strong Bayesian learning that occurred about these parameters is evident in Figure 11 of the supplementary material. Table 3 gives the relative errors to quantify the quality of the reconstructions. We observe nearly identical solutions under both MCMC approaches, both graphically and quantitatively.

Acceptance rate versus rank. To explore the effect of the retained number of eigenvalues on the acceptance rate for our algorithm, we estimate the predicted and empirical acceptance rates of sampling from the proposal distribution, over a range of truncation levels, at a given state of the chain. We fix the state by initializing $(\boldsymbol{x}, \mu, \sigma)$ as the last sample from one of the chains obtained from the LRIS. At each truncation level k, we compute the expected

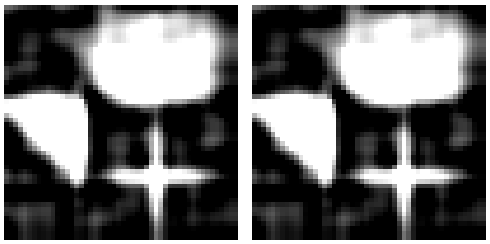

Figure 5. *Approximate posterior mean images obtained from block Gibbs sampling (left) and the proposed low-rank sampling algorithm (right) for the 2D deblurring example.*

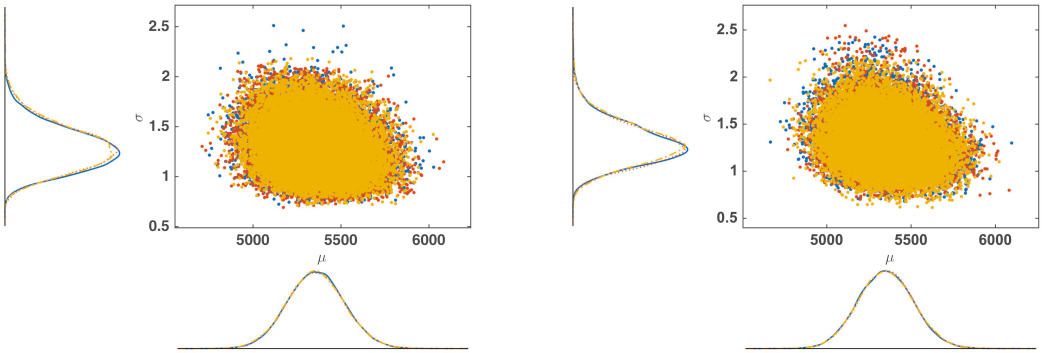

Figure 6. *Scatterplots of the (μ, σ) realizations obtained from block Gibbs (left panel) and the low-rank approach (right panel) for the 2D deblurring example, where the three different colors correspond to different chains. The smoothed marginal posterior densities are displayed in the margins of the plots.*

Table 3
Relative error (RE) of the estimates for the 2D image deblurring example.

Estimator	RE
Posterior mean (block Gibbs)	0.4453
Posterior mean (LRIS)	0.4455

value of the acceptance ratio using Theorem 1. We draw 2,000 samples from the proposal distribution and compute the acceptance ratio of each using Proposition 2. From these we estimate the empirical failure rates to compare with their expected values as the truncation level increases. Figure 7 displays the results. The close agreement between the predicted and empirical acceptance rates support the theoretical results in section 3.

4.2. CT image reconstruction. Computed x-ray tomography (CT) is a common medical imaging modality in which x-rays are passed through a body from a source to a sensor along parallel lines indexed by an angle ω and offset y with respect to a fixed coordinate system and origin. The intensities of the rays are attenuated according to an unknown absorption function as they pass through tissue. The attenuated intensity I is recorded while the lines are rotated around the origin so that $I(S) = I(0)\exp\{-\int_0^S \alpha(x(s))\,\mathrm{d}s\}$, where $s = 0$ is the source of the x-ray, $s = S$ is the receiver location, $x(\cdot)$ indicates the line position, and α is

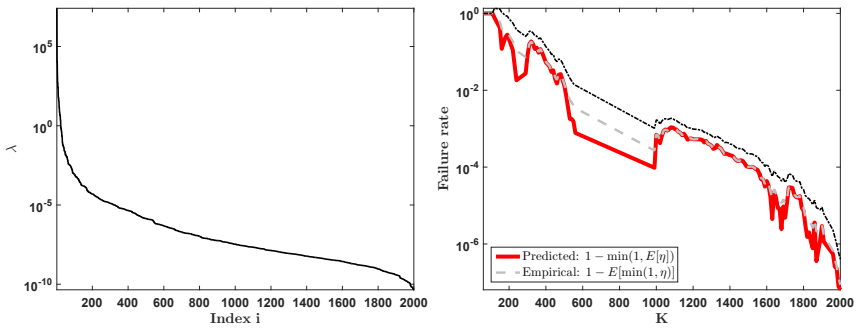

Figure 7. *(left) Eigenvalues of* $\boldsymbol{H}$ *in the* 2*D image deblurring example. (right) Predicted and empirical failure rates for different truncation levels* K*. The empirically determined upper* 0.975 *quantile is given by the dashed line; the lower quantile often attained zero values, so it is not displayed.*

the absorption function. The observed data are a transformation of the intensities, yielding the Radon transform model for CT [34, 4], $z(\omega, y) = \int_{L(\omega,y)} \alpha(x(s))\,\mathrm{d}s$, where $L(\omega, y)$ is the line along which the x-ray passes through the body. The inverse problem is to reconstruct the absorption function, which provides an image of the scanned body. Discretization of the integral yields the model in (1). This is typically an underdetermined system with infinitely many solutions, resulting in an ill-posed inverse problem.

Our target image is the Shepp–Logan phantom [56]. The forward model is implemented in MATLAB on the same computer as in subsection 4.1 with code available online [3]. The data are simulated by adding Gaussian noise with variance $0.01^2\|\boldsymbol{A}\boldsymbol{x}\|_\infty^2$. The target α is discretized to a relatively fine grid of size 128×128 so that $\dim(\boldsymbol{x}) = 16{,}384$. We suppose that the data are observed over lines and angles such that $\dim(\boldsymbol{b}) = 5{,}000$. Thus, $\operatorname{rank}(\boldsymbol{A}) = 5{,}000 \ll \dim(\boldsymbol{x})$, guiding our choice of eigenvalue truncation in the low-rank approximation to $\boldsymbol{H}$. An approximate eigendecomposition of the prior preconditioned Hessian is computed using randomized SVD with $\ell = 5{,}000$, as discussed in section 3, since computing the "exact" SVD is considerably more expensive. As in subsection 4.1, we take $\boldsymbol{L} = -\Delta + \delta\boldsymbol{I}$.

For the nuisance parameters, we use a weakly informative prior [18], namely, the proper Jeffreys prior proposed by Scott and Berger [55]. For convenience, we parameterize the model in terms of variance components instead of precisions, $\tau^2 := \sigma^{-1}$ and $\kappa^2 := \mu^{-1}$. Then the proper Jeffreys prior on (κ^2, τ^2) is[1]

$$\begin{aligned} \pi_{SB}(\kappa^2, \tau^2) &= (\kappa^2 + \tau^2)^{-2} \\ &= (\kappa^2)^{-1}(1 + \tau^2/\kappa^2)^{-2} \times (\kappa^2)^{-1} \\ &\equiv \pi(\tau^2 \mid \kappa^2)\pi(\kappa^2), \end{aligned} \tag{16}$$

so that the scale invariant prior is used for κ^2 while scaling τ^2 by the data level variance, as advocated by Jeffreys [29]. The implementation of this prior as a modification to Algorithm 1,

[1]We write "proportional to" ($\pi(\theta) \propto g(\theta)$) for proper priors to indicate that $\pi(\theta) = cg(\theta)$, where $c^{-1} = \int g(\theta)d\theta < \infty$ uniquely determines the density. However, the normalizing constant for an improper prior does not exist, so the prior is not unique. In the scale invariant case, any prior $\pi(\kappa^2) = a/\kappa^2$ for any $a > 0$ works. Since a is arbitrary, we simply set it equal to 1 for convenience and take $\pi(\kappa^2) = 1/\kappa^2$. See [8, Chapter 3].

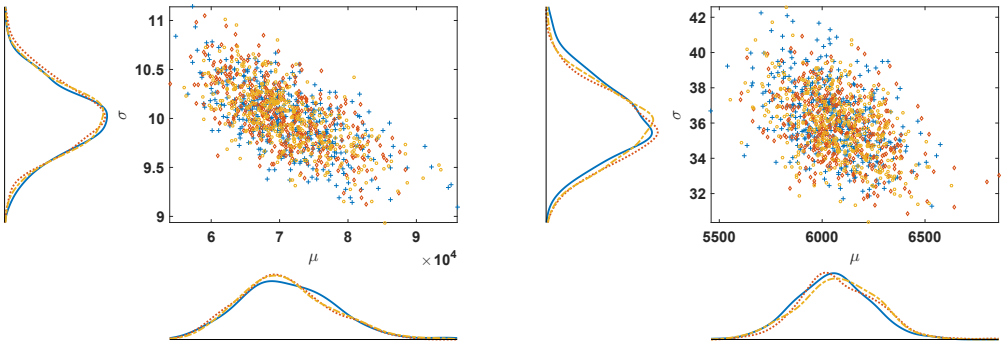

Figure 8. *Estimated joint densities for the precision parameters using the proper Jeffreys prior (left) and the conjugate Gamma priors (right).*

presented in Appendix B, is similar to the approach of [10]. Section 2 of the supplementary material contains further discussion of prior specification for the nuisance parameters.

We simulate three Markov chains using our proposed LRIS approach for 40,000 iterations (average acceptance rate $\sim 100\%$). Each chain is initialized with values drawn randomly from the prior. We thin the chains by retaining every 50th draw to reduce the autocorrelation, making it easier to diagnose convergence. We discard the first 400 draws of the thinned chains as a burn-in period. Trace plots and autocorrelation plots are used to verify approximate convergence of the chains. Relevant diagnostic plots are displayed in Figures 12, 13, and 14 of the supplementary material. The total computation time for our sampling approach is 197,517 seconds, or about 55 hours. This is noteworthy since the algorithm repeatedly updates a large, nontrivial covariance matrix and samples an approximately 16,000-dimensional Gaussian distribution 40,000 times. An ordinary block Gibbs sampler is simply not feasible for this problem.

We compare the results with samples obtained using the conjugate Gamma model with the same vague priors on μ and σ as in subsection 4.1. Convergence diagnostics are displayed in Figures 15 and 16 of the supplementary material. Figure 8 compares the approximate joint distributions for the precision parameters (μ, σ) under the conjugate model to the distribution based on the proper Jeffreys model, after back-transforming κ^2 and τ^2. Here we see the effect of prior selection in that both μ and σ tend to concentrate around different values, with much greater uncertainty in μ in the proper Jeffreys case. The differences between the two marginal posteriors of (μ, σ) affect the quality of the reconstructed images, displayed in Figure 9 and quantified in Table 4. This echoes Gelman's observation [18] that even a supposedly noninformative prior on the hyperprecision can have a disproportionate influence on the results. In this case, using a weakly informative prior for σ that depends on μ results in a higher quality reconstruction.

These simulations show that by exploiting the low-rank structure of the preconditioned Hessian of the forward model, we are able to substantially reduce the computational burden compared to block Gibbs sampling. Even when the forward model is of full row rank, the results illustrate the potential for efficiency gains using our proposed LRIS approach, provided the system is underdetermined. Our approach using either proper Jeffreys or conjugate priors

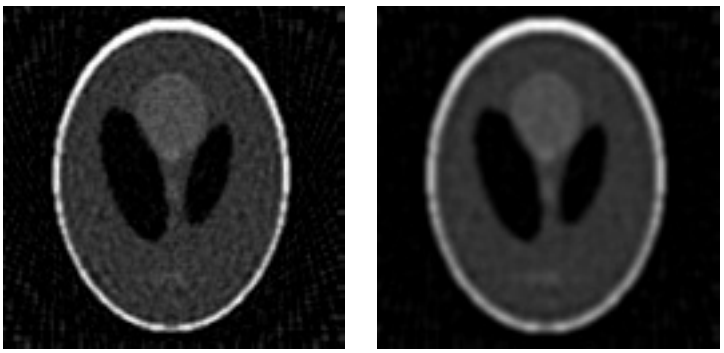

Figure 9. *Posterior mean estimators of the true image in the CT image reconstruction example, using the proper Jeffreys prior (left) and the conjugate Gamma priors (right).*

Table 4

Relative error (RE) of the estimates for the CT image reconstruction example.

Scale parameter prior	RE
Proper Jeffreys	0.3270
Conjugate Gamma	0.4229

is much more feasible than a block Gibbs sampler, for which the computational demands can be prohibitively expensive. A comparison of the conventional conjugate Gamma priors to a weakly informative prior suggests that even a "noninformative" prior may exert considerable influence on the results, despite strong Bayesian learning in the posterior.

In the supplementary material, we further consider the challenging problem of nuclear magnetic resonance (NMR) relaxometry. There, we demonstrate that our approach still produces within reasonable computation time a solution that is comparable to those obtained from deterministic iterative procedures such as conjugate gradient least squares. This is achieved with randomized SVD and without explicitly forming the forward operator $\boldsymbol{A}$ in the LRIS algorithm.

5. Discussion. When approximating the posterior distribution via Markov chain Monte Carlo in the hierarchical Bayesian linear inverse problem, the bottleneck is in repeatedly sampling high-dimensional Gaussian random variables. Sampling from the joint posterior with standard MCMC is challenging due to the high dimensionality of the estimand, since drawing from the full conditional involves expensive operations with the covariance matrix.

In this work we propose a computationally efficient sampling algorithm which is well suited for a fully Bayesian approach in which the noise precision and the prior precision parameters are unknown and assigned prior distributions. Our proposed low-rank independence sampler uses a proposal distribution constructed via low-rank approximation to the preconditioned Hessian. We show that the acceptance rate is high when the magnitudes of the discarded eigenvalues of the Hessian are small, a feature of severely ill-posed problems. When it is not obvious how to determine an appropriate truncation of rank due to the dependence of the known acceptance rates on other parameters in the model, we discuss how to adaptively determine the truncation level as part of the MCMC algorithm to find the minimal rank with

high acceptance rates. We demonstrate both theoretically and empirically that the quality of the approximation is directly related to the acceptance rate of the sampler, as intuition would suggest. We illustrate our approach on several examples, demonstrating convergence as a function of rank, as well as computational improvements to accessing the full posterior distribution.

A known issue with block Gibbs sampling in Bayesian inverse problems is the deterioration of the chains due to correlation between the hyperparameters and the estimand $\boldsymbol{x}$ as the dimension of the problem increases [5]. Several approaches have been proposed to ameliorate this by breaking the dependence between the hyperparameters and $\boldsymbol{x}$ in the algorithm. These include the one-block algorithm [49], partially collapsed samplers [60, 33], noncentered parameterization (NCP) [41, 42], and marginal then conditional (MTC) sampling [16]. Noncentered parameterization is easily incorporated into our proposed approach without sacrificing gains in efficiency. (See the supplementary material for discussion of NCP and an illustration of combining it with our low-rank sampler.) One-block sampling and MTC sampling, on the other hand, require expressions for marginal densities that no longer hold when substituting the true full conditional of $\boldsymbol{x}$ with an approximation, as well as approximation of determinants of large covariance matrices to which the results in this work are not directly applicable. Combining our proposed low-rank sampling approach with these algorithms is the subject of ongoing research, to appear in future work.

Appendix A. Proofs.

Proof of Proposition 2. The difference between the true and the approximate covariance matrices can be expressed as

$$\begin{aligned}\boldsymbol{\Gamma}_{\text{cond}}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1} &= \mu \boldsymbol{A}^\top \boldsymbol{A} + \sigma \boldsymbol{L}^\top \boldsymbol{L} - \left(\mu \boldsymbol{L}^\top \boldsymbol{V}_k \boldsymbol{\Lambda}_k \boldsymbol{V}_k^\top \boldsymbol{L} + \sigma \boldsymbol{L}^\top \boldsymbol{L}\right), \\ &= \mu \boldsymbol{L}^\top \left(\boldsymbol{L}^{-\top} \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{L}^{-1} - \boldsymbol{V}_k \boldsymbol{\Lambda}_k \boldsymbol{V}_k^\top\right) \boldsymbol{L}, \\ &= \mu \boldsymbol{L}^\top \left(\sum_{j=k+1}^{n} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top\right) \boldsymbol{L},\end{aligned}$$

giving that $\log w(\boldsymbol{x}) = -\frac{\mu}{2} \sum_{j=k+1}^{n} \lambda_j \left(\boldsymbol{v}_j^\top \boldsymbol{L} \boldsymbol{x}\right)^2 \leq 0$, and hence the acceptance ratio is given by (11). ∎

Lemma 3. *Suppose $\boldsymbol{M}$ is symmetric positive definite. Then*

$$\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} \boldsymbol{z}^T \boldsymbol{M} \boldsymbol{z} + \boldsymbol{J}^\top \boldsymbol{z}\right) d\boldsymbol{z} \;=\; \frac{(2\pi)^{n/2}}{\det(\boldsymbol{M})^{1/2}} \exp\left(\frac{1}{2} \boldsymbol{J}^\top \boldsymbol{M}^{-1} \boldsymbol{J}\right).$$

Proof. See [54, Lemma B.1.1]. ∎

Lemma 4. *The moments of the acceptance ratio are*

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\eta^m(\boldsymbol{z}, \boldsymbol{x})] \;=\; \frac{1}{N_m w^m(\boldsymbol{x})},$$

where N_m is defined in (12).

Proof. The proof proceeds in four steps.

1. *Simplifying* $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[w^m(\boldsymbol{z})]$. We focus on $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[w^m(\boldsymbol{z})]$. By definition, this is

$$\int_{\mathbb{R}^n} w^m(\boldsymbol{z})g(\boldsymbol{z})d\boldsymbol{z} = \frac{\exp\left(-\frac{1}{2}(\hat{\boldsymbol{x}}_{\text{cond}})^\top \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}\hat{\boldsymbol{x}}_{\text{cond}}\right)}{(2\pi)^{n/2}\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}})^{1/2}} \int \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z} + \boldsymbol{J}^\top\boldsymbol{z}\right) d\boldsymbol{z},$$

where, by using $\hat{\boldsymbol{x}}_{\text{cond}} = \mu\widehat{\boldsymbol{\Gamma}}_{\text{cond}}\boldsymbol{A}^\top\boldsymbol{b}$, we get

$$\boldsymbol{M} = m(\boldsymbol{\Gamma}_{\text{cond}}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}) + \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1} \quad \text{and} \quad \boldsymbol{J} = \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}\hat{\boldsymbol{x}}_{\text{cond}} = \mu\boldsymbol{A}^\top\boldsymbol{b}.$$

Applying Lemma 3 and rearranging, we have

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[w^m(\boldsymbol{z})] = \frac{\exp\left(\frac{\mu^2}{2}(\boldsymbol{A}^\top\boldsymbol{b})^\top(\boldsymbol{M}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}})\boldsymbol{A}^\top\boldsymbol{b}\right)}{\det(\boldsymbol{M})^{1/2}\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}})^{1/2}}. \tag{17}$$

We focus on the numerator and denominator of (17) separately.

2. *Denominator of* (17). Note that

$$\boldsymbol{M} = m(\boldsymbol{\Gamma}_{\text{cond}}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}) + \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1} = m\boldsymbol{\Gamma}_{\text{cond}}^{-1} + (1-m)\widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1},$$

and furthermore

$$\boldsymbol{M} = \boldsymbol{L}^\top\left(\mu\sum_{j=1}^{k}\lambda_j\boldsymbol{v}_j\boldsymbol{v}_j^\top + m\mu\sum_{j=k+1}^{n}\lambda_j\boldsymbol{v}_j\boldsymbol{v}_j^\top + \sigma\boldsymbol{I}\right)\boldsymbol{L}.$$

Using the properties of determinants, it can be shown that

$$\det(\boldsymbol{M}) = \sigma^n\det(\boldsymbol{L})^2\prod_{j=1}^{k}\left(1+\frac{\mu}{\sigma}\lambda_j\right)\prod_{j=k+1}^{n}\left(1+\frac{m\mu}{\sigma}\lambda_j\right).$$

Similarly,

$$\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}) = \sigma^n\det(\boldsymbol{L})^2\prod_{j=1}^{k}\left(1+\frac{\mu}{\sigma}\lambda_j\right).$$

Combining these results, the denominator of (17) becomes

$$\det(\boldsymbol{M})^{1/2}\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}})^{1/2} = \sqrt{\frac{\det(\boldsymbol{M})}{\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}})}} = \prod_{j>k}\left(1+\frac{m\mu}{\sigma}\lambda_j\right)^{1/2}.$$

3. *Numerator of* (17). Consider $\boldsymbol{M}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}}$. By the Woodbury matrix identity,

$$\begin{aligned}\boldsymbol{M}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}} &= \left(m\boldsymbol{\Gamma}_{\text{cond}}^{-1} + (1-m)\widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1}\right)^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}} \\ &= \frac{1}{\sigma}\boldsymbol{L}^{-1}\left(\boldsymbol{I} - \sum_{j=1}^{k}\frac{\mu\lambda_j}{\mu\lambda_j+\sigma}\boldsymbol{v}_j\boldsymbol{v}_j^\top - \sum_{j=k+1}^{n}\frac{m\mu\lambda_j}{m\mu\lambda_j+\sigma}\boldsymbol{v}_j\boldsymbol{v}_j^\top\right)\boldsymbol{L}^{-\top} \\ &\quad - \frac{1}{\sigma}\left(\boldsymbol{I} - \sum_{j=1}^{k}\frac{\mu\lambda_j}{\mu\lambda_j+\sigma}\boldsymbol{v}_j\boldsymbol{v}_j^\top\right)\boldsymbol{L}^{-\top} \\ &= -\frac{1}{\sigma}\sum_{j=k+1}^{n}\frac{m\mu\lambda_j}{m\mu\lambda_j+\sigma}\boldsymbol{L}^{-1}\boldsymbol{v}_j\boldsymbol{v}_j^\top\boldsymbol{L}^{-\top}.\end{aligned}$$

The numerator is therefore

$$\exp\left(\frac{\mu^2}{2}(\boldsymbol{A}^\top\boldsymbol{b})^\top(\boldsymbol{M}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}})\boldsymbol{A}^\top\boldsymbol{b}\right) = \exp\left(-\frac{\mu^2}{2\sigma}\sum_{j=k+1}^{n}\frac{m\mu\lambda_j}{m\mu\lambda_j+\sigma}(\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{L}^{-1}\boldsymbol{v}_j)^2\right).$$

4. *Combining intermediate results.* Plugging the results of steps 2 and 3 into (17) gives us $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[w^m(\boldsymbol{z})] = \frac{1}{N_m}$, where N_m is defined in (12). The proof readily follows because $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\eta^m(\boldsymbol{z},\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[w^m(\boldsymbol{z})]/w^m(\boldsymbol{x})$. ∎

Proof of Theorem 1. From Lemma 4, we have $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\eta^m(\boldsymbol{z},\boldsymbol{x})] = \frac{1}{N_m w^m(\boldsymbol{x})}$. The first result follows immediately by plugging in $m = 1$. For the second result, we use the fact that for a random variable X with $\mathbb{E}(X^2) < \infty$, $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. The result follows from (13) and applying Lemma 4 with $m = 2$. ∎

Proof of Proposition 3. Using h as defined in (10) and the definition of the proposal density g,

$$\frac{h(\boldsymbol{x})}{g(\boldsymbol{x})} = \sqrt{\frac{\det(\widehat{\boldsymbol{\Gamma}}_{\text{cond}})}{\det(\boldsymbol{\Gamma}_{\text{cond}})}}\, w(\boldsymbol{x}) \exp\left(-\frac{\mu^2}{2}\boldsymbol{b}^\top\boldsymbol{A}(\boldsymbol{\Gamma}_{\text{cond}} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}})\boldsymbol{A}^\top\boldsymbol{b}\right).$$

From the proof of Lemma 4, $\boldsymbol{M}^{-1} = \boldsymbol{\Gamma}_{\text{cond}}$ when $m = 1$. Comparing terms with (17), this gives us $h(\boldsymbol{x}) = N_1 g(\boldsymbol{x}) w(\boldsymbol{x})$, where N_1 is defined in (12). From the proof of Proposition 2, $\log w(\boldsymbol{x}) \le 0$, and therefore $w(\boldsymbol{x}) \le 1$. The desired result follows. Note that the bound is tight because $w(\boldsymbol{0}) = 1$. ∎

Proof of Theorem 2. From Proposition 1, we need to consider the quadratic form

$$\frac{1}{2}\boldsymbol{z}^\top(\boldsymbol{\Gamma}_{\text{cond}}^{-1} - \widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1})\boldsymbol{z} = \frac{\mu}{2}\boldsymbol{z}^\top\boldsymbol{L}^\top(\boldsymbol{H} - \widehat{\boldsymbol{H}})\boldsymbol{L}\boldsymbol{z},$$

where $\widehat{\boldsymbol{H}}$ is the low-rank approximation; see (15). This follows from $\widehat{\boldsymbol{\Gamma}}_{\text{cond}}^{-1} = (\mu\boldsymbol{L}^\top\widehat{\boldsymbol{H}}\boldsymbol{L} + \sigma\boldsymbol{L}^\top\boldsymbol{L})$. Using the Cauchy–Schwarz inequality, we can bound (in the spectral norm)

$$\boldsymbol{z}^\top\boldsymbol{L}^\top(\boldsymbol{H} - \widehat{\boldsymbol{H}})\boldsymbol{L}\boldsymbol{z} \le \|\boldsymbol{L}\boldsymbol{z}\|_2^2\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\|_2.$$

Arguing as in [23, section 5.3], we have $\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\|_2 \leq 2\|\boldsymbol{H} - \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{H}\|_2$. Applying [23, Theorem 10.6] we have

$$\mathbb{E}_{\boldsymbol{\Omega}} \|\boldsymbol{H} - \widehat{\boldsymbol{H}}\|_2 \leq 2 \left[\alpha \lambda_{k+1} + \beta \left(\sum_{j=k+1}^{n} \lambda_j^2 \right)^{1/2} \right], \tag{18}$$

with constants α and β given in the statement of the result. Applying Jensen's inequality,

$$\mathbb{E}_{\boldsymbol{\Omega}} [\eta(\boldsymbol{z}, \boldsymbol{x})] \geq \exp\left(-\mu \|\boldsymbol{L}\boldsymbol{z}\|^2 \mathbb{E}_{\boldsymbol{\Omega}} \|\boldsymbol{H} - \widehat{\boldsymbol{H}}\|_2 \right).$$

Plug (18) into the above equation to complete the proof. ∎

Appendix B. Implementing the proper Jeffreys prior. Here we briefly discuss an implementation of the LRIS algorithm when the Scott–Berger prior (16) is used instead of the independent conjugate Gamma priors.

In this case, the model (16) becomes

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{x}, \kappa^2 &\sim N(\boldsymbol{A}\boldsymbol{x}, \kappa^2 \boldsymbol{I}), \\ \boldsymbol{x} \mid \tau^2 &\sim N(\boldsymbol{0}, \tau^2 \boldsymbol{\Gamma}), \\ \pi(\tau^2 \mid \kappa^2) &= \kappa^{-2}(1 + \tau^2/\kappa^2)^{-2}, \\ \pi(\kappa^2) &= \kappa^{-2}. \end{aligned} \tag{19}$$

To obtain the full conditional densities necessary for Gibbs sampling, it is convenient to reparameterize the model with $\upsilon = \tau^2/\kappa^2$. After the change of variables, the joint posterior (5) becomes

$$\begin{aligned} \pi(\boldsymbol{x}, \kappa^2, \upsilon \mid \boldsymbol{b}) \propto{}& (\kappa^2)^{-\frac{m+n}{2}-1} \upsilon^{-n/2} (1+\upsilon)^{-2} \\ &\times \exp\left(-\frac{1}{2\kappa^2} \left[\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{1}{\upsilon} \|\boldsymbol{L}\boldsymbol{x}\|_2^2 \right] \right). \end{aligned} \tag{20}$$

The full conditional on $\boldsymbol{x}$ is a standard result for the normal-normal model (4); i.e.,

$$\boldsymbol{x} \mid \kappa^2, \upsilon, \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{x}_{\text{cond}}, \boldsymbol{\Gamma}_{\text{cond}}),$$

where $\boldsymbol{\Gamma}_{\text{cond}} = \kappa^2 \left(\boldsymbol{A}^\top \boldsymbol{A} + \upsilon^{-1} \boldsymbol{\Gamma}_{\text{pr}}^{-1} \right)^{-1}$ and $\boldsymbol{x}_{\text{cond}} = \left(\boldsymbol{A}^\top \boldsymbol{A} + \upsilon^{-1} \boldsymbol{\Gamma}_{\text{pr}}^{-1} \right)^{-1} \boldsymbol{A}^\top \boldsymbol{b}$. We can sample from this density using our proposed low-rank approximation approach. (See section 3.) The full conditional on κ^2 is

$$\kappa^2 \mid \boldsymbol{x}, \upsilon, \boldsymbol{y} \sim \text{InvGamma}\left(\frac{m+n}{2}, \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{1}{2\upsilon} \|\boldsymbol{L}\boldsymbol{x}\|_2^2 \right).$$

To draw from this density, draw $W \sim \text{Gamma}\left(\frac{m+n}{2}, \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{1}{2\upsilon} \|\boldsymbol{L}\boldsymbol{x}\|_2^2 \right)$ and set $\kappa^2 = 1/W$. The full conditional on υ is

$$\pi(\upsilon \mid \boldsymbol{x}, \kappa^2, \boldsymbol{b}) \propto \upsilon^{-(n/2+1)-1} \exp\left[-\frac{1}{2\kappa^2 \upsilon} \|\boldsymbol{L}\boldsymbol{x}\|_2^2 \right] \left(\frac{\upsilon}{1+\upsilon} \right)^2 \tag{21}$$

$$\equiv h(\upsilon) \left(\frac{\upsilon}{1+\upsilon} \right)^2, \tag{22}$$

where $h(v)$ is the density of an InvGamma$((n+2)/2, \|\boldsymbol{Lx}\|_2^2/(2\kappa^2))$ distribution. Thus we can use an independence Metropolis–Hastings algorithm with candidate distribution

$$\text{InvGamma}((n+2)/2, \|\boldsymbol{Lx}\|_2^2/(2\kappa^2)).$$

Acknowledgments. The authors thank the editors, an associate editor, and anonymous referees for comments and suggestions that improved this manuscript. The authors also would like to thank Duy Thai, Vered Madar, Johnathan Bardsley, and Ray Falk for useful conversations.

REFERENCES

[1] 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, https://doi.org/10.1137/130944229.

[2] K. E. Andersen, S. P. Brooks, and M. B. Hansen, *Bayesian inversion of geoelectrical resistivity data*, J. R. Stat. Soc. Ser. B Stat. Methodol., 65 (2003), pp. 619–642.

[3] J. M. Bardsley, *WMRNSD for Medical Imaging Examples*, http://www.math.umt.edu/bardsley/codes.html (accessed 2016-06-23).

[4] J. M. Bardsley, *Applications of nonnegatively constrained iterative method with statistically based stopping rules to CT, PET, and SPECT imaging*, Electron. Trans. Numer. Anal., 38 (2011), pp. 34–43.

[5] J. M. Bardsley, *MCMC-based image reconstruction with uncertainty quantification*, SIAM J. Sci. Comput., 34 (2012), pp. A1316–A1332, https://doi.org/10.1137/11085760X.

[6] J. M. Bardsley, M. Howard, and J. G. Nagy, *Efficient MCMC–based image deblurring with Neumann boundary conditions*, Electron. Trans. Numer. Anal., 40 (2013), pp. 476–488.

[7] J. M. Bardsley and A. Luttman, *A Metropolis-Hastings method for linear inverse problems with Poisson likelihood and Gaussian prior*, Int. J. Uncertain. Quantif., 6 (2016), pp. 35–55.

[8] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.

[9] S. P. Brooks and A. Gelman, *General methods for monitoring convergence of iterative simulations*, J. Comput. Graph. Statist., 7 (1998), pp. 434–455.

[10] D. A. Brown, G. S. Datta, and N. A. Lazar, *A Bayesian generalized CAR model for correlated signal detection*, Statist. Sinica, 27 (2017), pp. 1125–1153.

[11] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems, Part* I: *The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523, https://doi.org/10.1137/12089586X.

[12] D. Calvetti, J. P. Kaipio, and E. Somersalo, eds., *Inverse problems in the Bayesian framework*, Inverse Problems, 30 (2014).

[13] B. P. Carlin and T. A. Louis, *Bayesian Methods for Data Analysis*, 3rd ed., Chapman & Hall/CRC, Boca Raton, FL, 2009.

[14] R. V. Craiu and X.-L. Meng, *Perfection within reach: Exact MCMC sampling*, in Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds., Chapman & Hall/CRC Press, Boca Raton, FL, 2011, pp. 199–226.

[15] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432, https://doi.org/10.1137/090780717.

[16] C. Fox and R. A. Norton, *Fast sampling in a linear-Gaussian inverse problem*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 1191–1218, https://doi.org/10.1137/15M1029527.

[17] A. E. Gelfand and A. F. M. Smith, *Sampling-based approaches to calculating marginal densities*, J. Amer. Statist. Assoc., 85 (1990), pp. 398–409.

[18] A. Gelman, *Prior distributions for variance parameters in hierarchical models*, Bayesian Anal., 1 (2006), pp. 515–533.
[19] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC, Boca Raton, FL, 2014.
[20] A. Gelman, G. Roberts, and W. Gilks, *Efficient Metropolis jumping rules*, in Bayesian Statistics 5, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., Oxford University Press, New York, 1995, pp. 599–607.
[21] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences (with discussion)*, Statist. Sci., 7 (1992), pp. 457–511.
[22] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Trans. Pattern Anal., 6 (1984), pp. 721–741.
[23] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.
[24] P. C. Hansen, *Regularization tools version* 4.0 *for Matlab* 7.3, Numer. Algorithms, 46 (2007), pp. 189–194.
[25] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, Fundam. Algorithms 7, SIAM, Philadelphia, 2010, https://doi.org/10.1137/1.9780898718836.
[26] W. Hastings, *Monte Carlo sampling methods using Markov chains and their application*, Biometrika, 57 (1970), pp. 97–109.
[27] O. Hauk, *Keep it simple: A case for using classical minimum norm estimation in the analysis of EEG and MEG data*, NeuroImage, 21 (2004), pp. 1612–1621.
[28] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[29] H. Jeffreys, *Theory of Probability*, 3rd ed., Oxford University Press, Cambridge, UK, 1961.
[30] A. Johnson, G. L. Jones, and R. C. Neath, *Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition*, Statist. Sci., 28 (2013), pp. 360–375.
[31] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2007.
[32] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, *Introduction to variational methods for graphical models*, Mach. Learn., 37 (1999), pp. 183–233.
[33] K. T. Joyce, J. M. Bardsley, and A. Luttman, *Point spread function estimation in X-ray imaging with partially collapsed Gibbs sampling*, SIAM J. Sci. Comput., 40 (2018), pp. B766–B787, https://doi.org/10.1137/17M1149250.
[34] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.
[35] J. S. Liu, *Metropolized independent sampling with comparisons to rejection sampling and importance sampling*, Statist. Comput., 6 (1996), pp. 113–119.
[36] J. S. Liu, W. H. Wong, and A. Kong, *Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes*, Biometrika, 84 (1994), pp. 27–40.
[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1091.
[38] K. Mosegaard and A. Tarantola, *Probabilistic approach to inverse problems*, in International Handbook of Earthquake & Engineering Seismology, Part A, Academic Press, New York, 2002, pp. 237–265.
[39] P. Müller, *A Generic Approach to Posterior Integration and Gibbs Sampling*, Technical Report, Purdue University, West Lafayette, IN, 1991.
[40] C. Nahkhleh, D. Higdon, C. K. Allen, and R. Ryne, *Bayesian reconstruction of particle beam phase space*, in Bayesian Theory and Applications, Oxford University Press, Oxford, 2013, pp. 673–686.
[41] O. Papaspiliopoulos and G. O. Roberts, *Non-centered parameterisations for hierarchical models and data augmentation*, Bayesian Statist., 7 (2003), pp. 307–326.
[42] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld, *A general framework for the parametrization of hierarchical models*, Statist. Sci., 22 (2007), pp. 59–73.
[43] N. Petra, J. Martin, G. Stadler, and O. Ghattas, *A computational framework for infinite-dimensional Bayesian inverse problems, Part* II: *Stochastic Newton MCMC with application to ice*

sheet flow inverse problems, SIAM J. Sci. Comput., 36 (2014), pp. A1525–A1555, https://doi.org/10.1137/130934805.

[44] P. Z. G. Qian and C. F. J. Wu, *Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments*, Technometrics, 50 (2008), pp. 192–204.

[45] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.

[46] S. I. Resnick, *A Probability Path*, Birkhäuser, Boston, 1999.

[47] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, New York, 2004.

[48] J. S. Rosenthal, *Optimal proposal distributions and adaptive MCMC*, in Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds., Chapman & Hall/CRC, Boca Raton, FL, 2011, pp. 93–111.

[49] H. Rue and L. Held, *Gaussian Markov Random Fields*, Chapman & Hall/CRC, Boca Raton, FL, 2005.

[50] H. Rue, S. Martino, and N. Chopin, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*, J. R. Stat. Soc. Ser. B Stat. Methodol., 71 (2009), pp. 319–392.

[51] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Classics Appl. Math. 66, SIAM, Philadelphia, 2011.

[52] A. K. Saibaba and P. K. Kitanidis, *Fast computation of uncertainty quantification measures in the geostatistical approach to solve inverse problems*, Adv. Water Resour., 82 (2015), pp. 124–138.

[53] A. K. Saibaba, J. Lee, and P. K. Kitanidis, *Randomized algorithms for generalized Hermitian eigenvalue problems with application to computing Karhunen–Loève expansion*, Numer. Linear Algebra Appl., 23 (2016), pp. 314–339.

[54] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer-Verlag, New York, 2003.

[55] J. G. Scott and J. O. Berger, *An exploration of aspects of Bayesian multiple testing*, J. Statist. Plann. Inference, 136 (2006), pp. 2144–2162.

[56] L. A. Shepp and B. F. Logan, *The Fourier reconstruction of a head section*, IEEE Trans Nucl. Sci., 21 (1974), pp. 21–43.

[57] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 37 (2015), pp. A2451–A2487, https://doi.org/10.1137/140977308.

[58] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[59] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *Randomized Single-View Algorithms for Low-Rank Matrix Approximation*, preprint, https://arxiv.org/abs/1609.00048v1, 2016.

[60] D. van Dyk and T. Park, *Partially collapsed Gibbs samplers*, J. Amer. Statist. Assoc., 103 (2008), pp. 790–796.