Supplementary Material for "Low Rank Independence Samplers in Hierarchical Bayesian Inverse Problems"[*]

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

This Supplementary Material contains additional material referenced in the manuscript that could aid the reader, but is not essential. We empirically demonstrate mixing behavior and efficiency of an LRIS-based Metropolis-Hastings-within-Gibbs sampler as a function of rank of the approximation, elaborate on possibilities for prior modeling of the precision (variance) components in the hierarchical model for the Bayesian inverse problem, and discuss the use of noncentered parameterization in our proposed LRIS algorithm, supported by application to the 2D deblurring example. We further demonstrate computational feasibility afforded by our proposed approach by considering MCMC/LRIS-based reconstruction of a distribution of relaxation times in nuclear magnetic resonance (NMR) relaxometry. Supplementary Figures are at the end.

1. Effect of the Proposal Rank on Chain Mixing. To study how the quality of the low-rank approximation affects the mixing of an MCMC algorithm, we consider the image reconstruction problem originally appearing in [21]. This is a one-dimensional image restoration problem in which the blurred image can be expressed as a Fredholm integral equation of the first kind, $g(s) = \int_{-\pi/2}^{\pi/2} K(s,t) f(t)\,\mathrm{d}t$, where $K(s,t) = (\cos(s) + \cos(t))^2\,(\sin(u)/u)^2$ with $u = \pi(\sin(s) + \sin(t))$, and f is the true one-dimensional image given by $f(t) = 2\exp(-6(t-0.8)^2) + \exp(-2(t+0.5)^2)$. The problem is to reconstruct f given g and K. Using the implementation in the `Matlab` package `Regularization Tools` [10], we discretize the integral via quadrature over $n = 512$ points and corrupt the observations with one percent noise. The resulting model is $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{b}$ is the observed data, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the discretized forward model, $\boldsymbol{x}$ is the discretized solution, and $\boldsymbol{\Gamma} = 0.01^2\|\boldsymbol{b}\|^2\boldsymbol{I}$. The observed data $\boldsymbol{b}$ and solution f are displayed in Figure 1.

In the hierarchical Bayesian model, we use a zero mean Gaussian process (GP) prior [19, 18], $f(\cdot) \sim \mathcal{GP}(0, \sigma^{-1}R(\cdot,\cdot))$. We take the correlation function to be in the power exponential family, $R(t_i, t_j) = \exp\left(-|t_i - t_j|/l\right)$, with correlation length parameter $l = \pi/2$. We use vague Gamma priors about the noise precision μ and prior precision σ with $a_\mu = b_\mu = a_\sigma = b_\sigma = 0.1$. We remark that for this example, the forward model is fast-running so that our computationally-cheap LRIS is actually not necessary. However, the fast-running model makes it feasible to run a large number of replicate MCMC chains in a reasonable amount of time, thus allowing us to empirically assess the mixing behavior of our proposed approach as a

[*] Supplementary material for JUQ MS#M113721.

[†] 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).

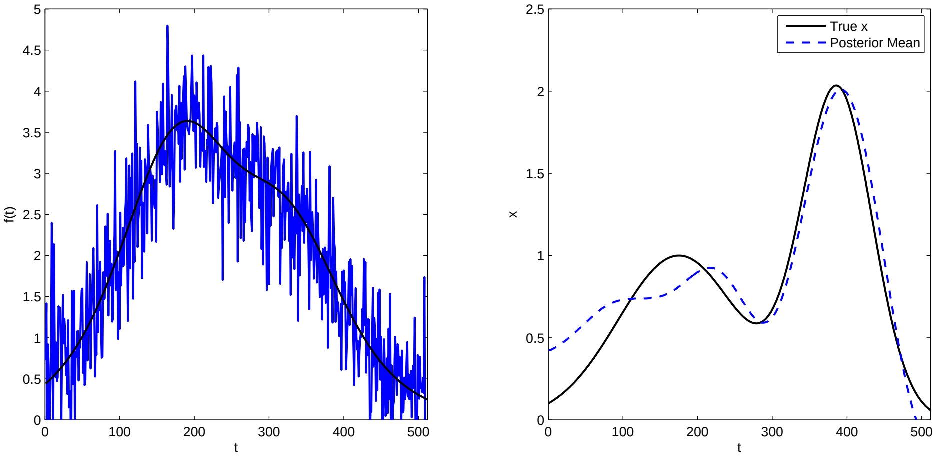

Figure 1. *Observed data (left panel) and true solution (right panel) for the one-dimensional image restoration example. The solid black line in the left panel represents the noise-free observations* $\boldsymbol{Ax}$*. The dashed blue line is the true posterior mean (approximated from a very long block Gibbs Markov chain).*

function of rank of the proposal distribution.

Similar to the analysis in [13], we consider two empirical measures to assess the mixing behavior of the Markov chains as a function of rank of the approximation in the LRIS algorithm. To assess the statistical efficiency of using the mean of the MCMC output to estimate $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{b})$, we use the mean squared error $MSE(\overline{\boldsymbol{x}}_n) := \mathbb{E}[(\overline{\boldsymbol{x}}_n - \mathbb{E}(\boldsymbol{x} \mid \boldsymbol{b}))^2]$, where $\overline{\boldsymbol{x}}_n$ is the sample mean of a chain of length n, $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$, obtained from an MCMC run. To approximate the MSE, we find the sample mean, $\overline{\boldsymbol{x}}^*$, from a very long run of an ordinary block Gibbs sampler and treat this as the true posterior mean $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{b})$. We run an additional m independent Markov chains using LRIS-based Metropolis-Hastings-within-Gibbs, each of length n, whence we can approximate $MSE(\overline{\boldsymbol{x}}_n)$ with $\widehat{MSE}(\overline{\boldsymbol{x}}_n) = m^{-1}\sum_{i=1}^m (\overline{\boldsymbol{x}}_n^{(i)} - \overline{\boldsymbol{x}}^*)^2$, where $\overline{\boldsymbol{x}}_n^{(i)}$ is the sample mean obtained from the i^{th} chain. The second measure we consider is the expected squared Euclidean jump distance, defined as $ESEJD = \mathbb{E}(\|\boldsymbol{x}_{(t+1)} - \boldsymbol{x}_{(t)}\|_2^2)$. This quantity is indicative of how well a Markov chain is exploring the marginal posterior distribution of the estimand $\boldsymbol{x}$. To approximate this expected value, we again run m independent chains, each of length n. For each chain with rank k, we find the mean squared Euclidean jump distance $MSEJD_k := (n-1)^{-1}\sum_{t=1}^{n-1} \|\boldsymbol{x}_{(t+1)} - \boldsymbol{x}_{(t)}\|_2^2$. Then we obtain an estimate of ESEJD with $\widehat{ESEJD}_k = m^{-1}\sum_{i=1}^m MSEJD_k^{(i)}$, where $MSEJD_k^{(i)}$ is mean squared Euclidean jump distance from the i^{th} chain.

The left panel of Figure 2 displays the first ten eigenvalues of the prior-preconditioned Hessian for the one-dimensional reconstruction example. As most of the information is captured in the first ten eigenvalues, we consider the LRIS algorithm using proposal distributions of ranks $k = 1, \ldots, 10$. For each proposal distribution, we run $m = 100$ independent Markov chains of length $n = 2000$, discarding the first $1,000$ draws as a burn-in period. To approximate the true posterior distribution, we run an ordinary block Gibbs sampler for 10^6 iterations

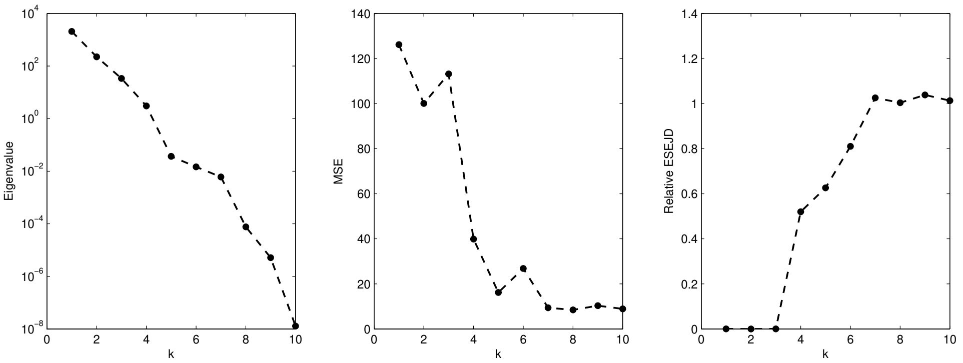

Figure 2. *Left panel: Spectrum (up to the first 10 eigenvalues) of the prior-preconditioned Hessian matrix for the one-dimensional image restoration example; Middle: The approximate mean squared error (MSE) in estimating $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{b})$ versus the rank of the proposal distribution in the LRIS; Right: Approximate expected squared Euclidean jump distance (ESEJD) of the $\boldsymbol{x}$ chain relative to the ESEJD from block Gibbs sampling at convergence.*

and approximate $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{b})$ with the mean of the last 5×10^5 draws. This target is displayed in the right panel of Figure 1.

The middle panel of Figure 2 displays the mean squared errors in the posterior mean estimates versus the rank of the LRIS proposal distribution. We see that the statistical efficiency of the sample mean increases sharply as non-negligible eigenvalues are added to the low-rank approximation, and becomes steady at $k = 7$. The right panel of the Figure displays the approximate expected squared Euclidean jumping distance of the Markov chains relative to the average squared jumping distance of the block Gibbs sampler, $\widehat{ESEJD}_k/ESEJD_{\text{Gibbs}}$, $k = 1, \dots, 10$. Similar to MSE, we can glean that only seven eigenvalues are necessary to obtain within 2,000 iterations a sample whose behavior is equivalent to a converged block Gibbs sampler. This equivalence is further supported in Figure 3, which displays smoothed estimates of the marginal densities of μ and σ as the rank increases. We again see the ability of the LRIS-based algorithm to closely estimate the true marginal density with only 2,000 MCMC iterates. It is interesting to observe that the noise precision μ is well identified regardless of the rank of the approximation, whereas σ is much more difficult to estimate. This reflects identifiability issues that are characteristic of ill-posed Bayesian inverse problems [2].

These results demonstrate sharp improvement in the behavior of an LRIS-based MCMC algorithm that is possible when the spectrum of the prior-preconditioned Hessian decays rapidly. Thus, for computationally expensive forward models, there is the potential to dramatically reduce the computational burden without sacrificing the convergence behavior of the Markov chain.

2. Priors on the Nuisance Parameters.

Consider the Bayesian linear inverse problem with forward operator $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, assuming independent observations with common precision μ. We assume a Gaussian prior on the target $\boldsymbol{x}$ to correspond to the L_2 penalty in regularized inversion, with prior covariance matrix $\boldsymbol{\Gamma}_{\text{pr}}$ known up to a multiplicative constant (precision)

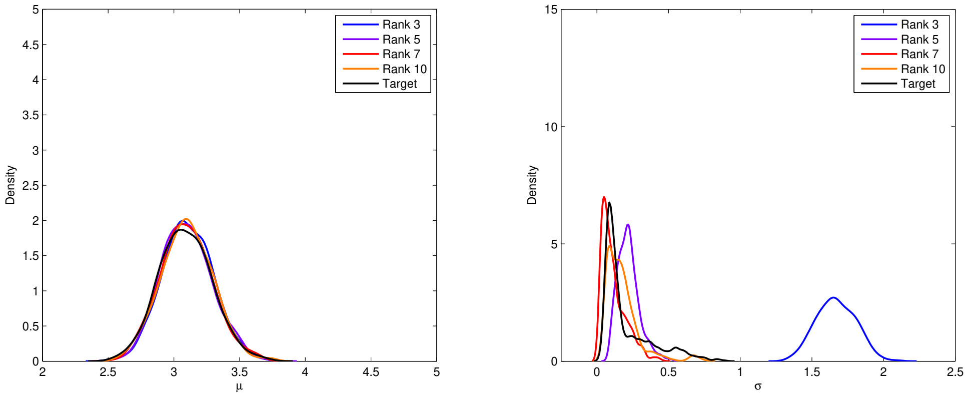

Figure 3. *Smoothed approximate marginal posterior densities of the noise precision μ (left) and prior precision σ (right) in the one-dimensional image restoration example. The target density is estimated with the last 1,000 draws from an ordinary block Gibbs sampler with chain length 10^6.*

σ. Let $\boldsymbol{b}$ denote the observed data. Then the model is

$$\begin{aligned} \boldsymbol{b} \mid \boldsymbol{x}, \mu &\sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \mu^{-1}\boldsymbol{I}) \\ \boldsymbol{x} \mid \sigma &\sim \mathcal{N}(\boldsymbol{0}, \sigma^{-1}\boldsymbol{\Gamma}_{\text{pr}}) \\ (\mu, \sigma) &\sim \Pi, \end{aligned} \tag{1}$$

where Π is some distribution with support $\mathbb{R}^+ \times \mathbb{R}^+$. In the following, we consider two different specifications of Π.

2.1. Conditionally Conjugate Gamma Priors.

The most straightforward case (and by far the most common) is to let $\mu \sim \text{Gamma}(a_\mu, b_\mu)$ and $\sigma \sim \text{Gamma}(a_\sigma, b_\sigma)$ independently, where we use the shape/rate parameterization of a Gamma distribution; e.g.,

$$\text{Gamma}(\mu \mid a, b) \propto \mu^{a-1} \exp(-b\mu), \quad \mu > 0.$$

With these Gamma prior models on μ and σ, the joint posterior distribution of the model specified by (1) is given by

$$\pi(\boldsymbol{x}, \mu, \sigma \mid \boldsymbol{b}) \propto \mu^{m/2+a_\mu-1} \sigma^{n/2+a_\sigma-1} \exp\left(-\frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \frac{\sigma}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2 - b_\mu\mu - b_\sigma\sigma\right). \tag{2}$$

Since the elements of $\boldsymbol{x}$ are expected to be highly correlated in the posterior, it is desirable to update $\boldsymbol{x}$ all at once in a block Gibbs sampler. The full conditional distributions in this case are

$$\begin{aligned} \boldsymbol{x} \mid \boldsymbol{b}, \mu, \sigma &\sim \mathcal{N}(\boldsymbol{x}_{\text{cond}}, \boldsymbol{\Gamma}_{\text{cond}}) \\ \mu \mid \boldsymbol{b}, \boldsymbol{x}, \sigma &\sim \text{Gamma}\left(m/2 + a_\mu, \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + b_\mu\right) \\ \sigma \mid \boldsymbol{b}, \boldsymbol{x}, \mu &\sim \text{Gamma}\left(n/2 + a_\sigma, \frac{1}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2 + b_\sigma\right), \end{aligned} \tag{3}$$

where $\mathbf{\Gamma}_{\text{cond}} = (\mu \boldsymbol{A}^\top \boldsymbol{A} + \sigma \mathbf{\Gamma}_{\text{pr}}^{-1})^{-1}$ and $\boldsymbol{x}_{\text{cond}} = \mu \mathbf{\Gamma}_{\text{cond}} \boldsymbol{A}^\top \boldsymbol{b}$. The remaining question becomes specification of the hyperparameters in the Gamma priors.

To impose strong prior assumptions and to stabilize the MCMC algorithm, we can rescale the observed data with $\tilde{\boldsymbol{b}} = \boldsymbol{b}/s_b$, where $s_b = \sqrt{(m-1)^{-1}\sum_{i=1}^m (b_i - \bar{b})^2}$ is the sample standard deviation. (See, for instance, [11].) In this case, $\mathbb{V}(\tilde{\boldsymbol{b}})$ is expected to be reasonably close to one, though not exactly equal since correlation in the data induced by dependence on $\boldsymbol{x}$ will cause s_b to over- or under-estimate the true standard deviation. After rescaling, an equivalent model to (1) is

$$\tilde{\boldsymbol{b}} = \boldsymbol{A}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\epsilon}}, \tag{4}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{x}/s_b$ and $\tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \tilde{\mu}^{-1}\boldsymbol{I})$. This is similar to the notion of a standardized regression model. (See, e.g., [14, Section 7.5].) In this case, we set the hyperparameters in the prior on $\tilde{\mu}$ (or simply μ, without loss of generality) to mildly concentrate the density about one; e.g. $a_\mu = b_\mu = 1$. By concentrating μ about 1 and allowing σ to be vague with, say, $a_\sigma = b_\sigma = 0.1$, we do not strongly restrict values of $\lambda = \sigma/\mu$, the corresponding regularization parameter in the MAP estimator. We remark, however, that in a fully Bayesian model the primary goal is to obtain an estimate of $\boldsymbol{x}$, so we are not really interested in μ or σ in their own right (hence the term "nuisance parameters").

If $\mu \sim \text{Gamma}(\epsilon, \epsilon)$, then $\pi(\mu) \propto \mu^{\epsilon-1}e^{-\epsilon\mu} \to \mu^{-1}$ as $\epsilon \to 0$. But $\pi(\mu) = \mu^{-1}$ is the Jeffreys prior for a scale parameter and thus is invariant to reparameterization [3, Ch. 3]. For this reason it is common practice to set ϵ to some small value in a $\text{Gamma}(\epsilon, \epsilon)$ prior, say $\epsilon = 0.1$ or $\epsilon = 0.01$, to approximate the behavior of the objective prior without sacrificing propriety or conjugacy. These are the priors we use in the 2D image deblurring example in Section 4.1 of the manuscript.

2.2. Weakly Informative Priors. Despite the convenience associated with the Gamma priors, it was observed by [8] that there is no limiting posterior distribution associated with taking $\epsilon \to 0$ in a $\text{Gamma}(\epsilon, \epsilon)$ prior, and that using such a hyperprior on the prior-level precision σ can sometimes yield undesirable behavior. (Although Carlin and Louis [5] remarked that it may not make a difference in terms of the estimand of interest, $\boldsymbol{x}$.) To rectify this, Gelman [8] proposed as a default prior the *folded-t* distribution. This prior strikes a good compromise between a completely noninformative prior, which can lead to unreasonable estimates if the data are not informative about a parameter, and a strongly informative prior which prevents the data from 'speaking for themselves' in determining plausible *a posteriori* values. As such, it is called a "weakly informative" prior. Scott and Berger [20] proposed what has since become known as a "proper Jeffreys" prior [17] on the variance components. Defining $\kappa^2 = \mu^{-1}$ and $\tau^2 = \sigma^{-1}$, the proper Jeffreys prior takes

$$\pi(\kappa^2, \tau^2) = (\kappa^2 + \tau^2)^{-2}, \quad \kappa^2, \tau^2 > 0.$$

This prior approximates the improper Jeffreys prior, $\pi(\kappa^2, \tau^2) = (\kappa^2 + \tau^2)^{-1}$ [22]. Scott and Berger [20] observed that the proper Jeffreys prior can be written as $\pi(\kappa^2, \tau^2) = \kappa^{-2}(1 + \tau^2/\kappa^2)^{-2}\kappa^{-2} \equiv \pi(\tau^2 \mid \kappa^2)\pi(\kappa^2)$, so that this model is equivalent to using the usual objective prior on the data-level variance while scaling the prior-level variance by κ^2, following the principle originally suggested by Jeffreys [12]. The conditional prior on $\tau^2 \mid \kappa^2$ is also proper, an important consideration in finite mixture models, or when the data contain limited information about τ^2. Lastly, it was observed by [4] that the proper Jeffreys prior is tail-equivalent to the prior obtained by placing a folded-t_2 on τ, and thus is suitable as a default prior choice. For these reasons, this can be an attractive alternative to conjugate Gamma priors. In the CT example in Section 4.2 of the manuscript, we use the proper Jeffreys prior on the variance components. The sampling algorithm is not quite as simple since we no longer have conjugacy (see Appendix B of the manuscript), but we are still able to use our proposed low-rank independence sampler.

3. Non-Centered Parameterizations.

In model (1), the distribution of $\boldsymbol{x}$ depends on σ, and the distribution of $\boldsymbol{b}$ depends on $\boldsymbol{x}$, but $\boldsymbol{b}$ is conditionally independent of σ, given $\boldsymbol{x}$. Under Gamma priors on μ and σ, this yields convenient conditionally conjugate distributions for use inside a block Gibbs sampler, as discussed in Section 2.1 of the Supplementary Material. However, this also leads to high correlation between the σ and $\boldsymbol{x}$ chains as the dimension of the problem increases, as noted by Bardsley [2] and Agapiou et al. [1].

A framework for potentially reducing the dependence between parameters in an MCMC algorithm is the so-called non-centered parameterization [15, 16]. A non-centered parameterization is one such that parameters are assigned *independent* prior distributions but still result in a model equivalent to the usual case, called a centered parameterization. The “centered” and “non-centered” terminology is a reference to the parameterizations considered by [7] for efficient sampling on normal linear mixed models, where certain parameters were centered or non-centered about other parameters.

Papaspiliopoulos et al. [16] argued that the best choice of parameterization depends on how well the underlying parameters are identified by the data. In our case, if the data were significantly informative about $\boldsymbol{x}$ so that strong Bayesian learning occurred in the posterior, then a centered parameterization would likely be appropriate. On the other hand, if $\boldsymbol{x}$ is only weakly identified by the data alone and hence more dependent on prior information, then there tends to be stronger correlation between σ and $\boldsymbol{x}$ under the centered parameterization. In this case, a non-centered parameterization is likely the better option. Similar behavior was observed by Agapiou et al. [1], where it was shown that the performance of the non-centered parameterization breaks down as the data-level variance becomes small (i.e., the data become more reliable and thus contain more information about the solution). There exist also “partially non-centered” parameterizations, which can be estimated from the data when the appropriate parameterization to use is not clear [15, 16].

3.1. Implementation.

To determine a non-centered parameterization for the hierarchical Bayesian inverse problem, we define a random variable $\boldsymbol{z}$ independent of σ and express the distribution of $\boldsymbol{x}$ in terms of these independent random variables. In our case, we have that $\boldsymbol{x} \overset{\mathrm{d}}{=} \sigma^{-1/2}\boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\mathrm{pr}})$ independent of σ. Substituting this parameterization into

(1), the model becomes

$$\begin{aligned} \boldsymbol{b} \mid \boldsymbol{z}, \mu, \sigma &\sim \mathcal{N}(\sigma^{-1/2}\boldsymbol{A}\boldsymbol{z}, \mu^{-1}\boldsymbol{I}) \\ \boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\mathrm{pr}}) \\ \mu &\sim \mathrm{Gamma}(a_\mu, b_\mu) \\ \sigma &\sim \mathrm{Gamma}(a_\sigma, b_\sigma). \end{aligned} \tag{5}$$

The joint posterior density is

$$\pi(\boldsymbol{z}, \mu, \sigma \mid \boldsymbol{b}) \propto f(\boldsymbol{b} \mid \boldsymbol{z}, \mu, \sigma)\pi(\boldsymbol{z})\pi(\mu)\pi(\sigma), \tag{6}$$

where f is the likelihood, and we adopt the conventional ambiguous use of π, understood to be defined by its arguments.

The conditional distributions for Gibbs sampling can be derived in a similar manner to the centered case discussed in Section 2.1 of the Supplementary Material, with the exception of σ. The non-centered parameterization loses the conditional conjugacy on this parameter, making an indirect sampling approach necessary. The conditional density of $\boldsymbol{z}$ can be derived as the usual Normal-Normal model $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}_\sigma\boldsymbol{z}, \mu^{-1}\boldsymbol{I})$ and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\mathrm{pr}})$, where $\boldsymbol{A}_\sigma = \sigma^{-1/2}\boldsymbol{A}$. Thus,

$$\boldsymbol{z} \mid \boldsymbol{b}, \mu, \sigma \sim \mathcal{N}(\boldsymbol{z}_{\mathrm{cond}}, \boldsymbol{\Gamma}_{\mathrm{cond}}),$$

where $\boldsymbol{z}_{\mathrm{cond}} = (\mu\boldsymbol{A}_\sigma^\top\boldsymbol{A}_\sigma + \boldsymbol{L}^\top\boldsymbol{L})^{-1} = ((\mu/\sigma)\boldsymbol{A}^\top\boldsymbol{A} + \boldsymbol{L}^\top\boldsymbol{L})^{-1}$ and $\boldsymbol{z}_{\mathrm{cond}} = \mu\boldsymbol{\Gamma}_{\mathrm{cond}}\boldsymbol{A}_\sigma^\top\boldsymbol{b} = (\mu/\sigma^{1/2})\boldsymbol{\Gamma}_{\mathrm{cond}}\boldsymbol{A}^\top\boldsymbol{b}$. One can easily show the same results for this parameterization as in the manuscript with appropriate substitutions of μ and σ. In particular, the same simplification of the MH acceptance ratio for the $\boldsymbol{z}$ chain holds as in Proposition 1. The full conditional of μ is also straightforward. It is derived similarly to the centered case:

$$\begin{aligned} &\pi(\mu \mid \boldsymbol{b}, \boldsymbol{z}, \sigma) \propto \mu^{m/2+a_\mu-1} \exp\left[-\mu\left(\frac{1}{2}\|\boldsymbol{A}_\sigma\boldsymbol{z} - \boldsymbol{b}\|_2^2 + b_\mu\right)\right] \\ &\Rightarrow \mu \mid \boldsymbol{b}, \boldsymbol{z}, \sigma \sim \mathrm{Gamma}\left(m/2 + a_\mu, \frac{1}{2}\|\boldsymbol{A}_\sigma\boldsymbol{z} - \boldsymbol{b}\|_2^2 + b_\mu\right). \end{aligned}$$

The most substantial difference between the centered and non-centered parameterizations is the loss of conditional conjugacy on σ. The conditional density for σ is

$$\pi(\sigma \mid \boldsymbol{b}, \boldsymbol{z}, \mu) \propto \exp\left[-\frac{\mu}{2}\|\sigma^{-1/2}\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\|_2^2\right]\sigma^{a_\sigma-1}e^{-b_\sigma\sigma}. \tag{7}$$

While there is no obvious simplification or standard distribution for σ, we can use a random walk Metropolis step to sample from it. Simplifying (7) slightly, we have

$$\pi(\sigma \mid \boldsymbol{b}, \boldsymbol{z}, \mu) \propto \sigma^{a_\sigma-1}e^{-b_\sigma\sigma} \exp\left[-\frac{\mu}{2\sigma}\left(\boldsymbol{z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{z} - 2\sigma^{1/2}\boldsymbol{z}^\top\boldsymbol{A}^\top\boldsymbol{b}\right)\right], \quad \sigma > 0.$$

To eliminate boundary constraints on σ and thus facilitate Gaussian proposals in the Metropolis algorithm, reparameterize the model with $\omega = \log(\sigma)$. Then the density for ω becomes

$$\pi_\omega(\omega \mid \boldsymbol{b}, \boldsymbol{z}, \mu) \propto e^{(a_\sigma-1)\omega - b_\sigma e^\omega} \exp\left[-\frac{\mu}{2}\left(e^{-\omega}\boldsymbol{z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{z} - 2e^{-\omega/2}\boldsymbol{z}^\top\boldsymbol{A}^\top\boldsymbol{b}\right)\right].$$

To implement the Metropolis step inside the block Gibbs algorithm, we can explicitly separate the burn-in phase from the sampling phase. During the first phase, we seek a suitable proposal distribution for σ (i.e., a suitable Gaussian proposal for ω). This is done by adaptively controlling the variance c, adjusting its scale based on the acceptance rate of the σ samples. A skeleton of this procedure is as follows:

```
1. Initialize $c = c_0$ and $accept = 0$.
2. For $i \in$ (burn-in iterations)
    (a) Draw $\omega^* \sim N(\omega^{(i-1)}, c)$
    (b) Accept/reject $\omega^*$ according to the Metropolis ratio. If accepted, $accept \leftarrow accept + 1$.
    (c) If $\mod(i, 100) = 0$, then
        i. If $accept/100 < 0.35,\ c \leftarrow 0.75c$,
        ii. Else, if $accept/100 > 0.5,\ c \leftarrow 1.75c$
        iii. $accept \leftarrow 0$
    (d) Repeat
3. For $i \in$ (sampling iterations)
    (a) Draw $\omega^* \sim N(\omega^{(i-1)}, c)$, where $c$ is fixed at the value determined from the burn-in period.
    (b) Accept/reject $\omega^*$ according to the Metropolis ratio.
    (c) Repeat
```

Agaipiou et al. [1] also considered a non-centered parameterization in a Bayesian Gaussian linear inverse problem, similar to that considered in this work. They relied on Metropolis sampling, but with a different proposal mechanism. Instead of sampling σ directly, they considered $\zeta := \sigma^{-1/2}$. After a change of variables, the full conditional distribution of ζ is

$$\begin{aligned}\pi(\zeta \mid \boldsymbol{b}, \boldsymbol{z}, \mu) &\propto f(\boldsymbol{b} \mid \boldsymbol{z}, \mu, \zeta)\pi(\zeta) \\ &\propto \exp\left[-\frac{\mu}{2}(\boldsymbol{b} - \zeta \boldsymbol{A}\boldsymbol{z})^\top(\boldsymbol{b} - \zeta \boldsymbol{A}\boldsymbol{z})\right]\left(\frac{1}{\zeta^2}\right)^{a_\sigma + 1/2} e^{-b_\sigma/\zeta^2}.\end{aligned}$$

We see that the likelihood contribution to this density, written as a function of ζ, is proportional to a Gaussian density. That is,

$$g(\zeta) := \exp\left[-\frac{\mu}{2}(\boldsymbol{b} - \zeta \boldsymbol{A}\boldsymbol{z})^\top(\boldsymbol{b} - \zeta \boldsymbol{A}\boldsymbol{z})\right] \propto \exp\left[-\frac{\mu \boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z}}{2}\left(\zeta - \frac{\boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{b}}{\boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z}}\right)^2\right], \tag{8}$$

which is the density of a normal distribution with mean $\zeta_c := \boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{b}(\boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z})^{-1}$ and variance $\xi_c := (\mu \boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z})^{-1}$. Agaipiou et al. [1] use this Gaussian distribution as a proposal for an independence sampler, except that ζ is restricted to be positive. In other words, their proposal distribution is a truncated Gaussian with density

$$q(\zeta^*) = \mathcal{N}(\zeta^* \mid \zeta_c, \xi_c)\left(1 - \Phi(-\zeta_c/\sqrt{\xi_c})\right)^{-1},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. It is important to note that this approach assumes $\boldsymbol{A}$ is of full rank, so that $\boldsymbol{A}\boldsymbol{z} \neq \boldsymbol{0}$ for all $\boldsymbol{z}$ (*a.e.*).

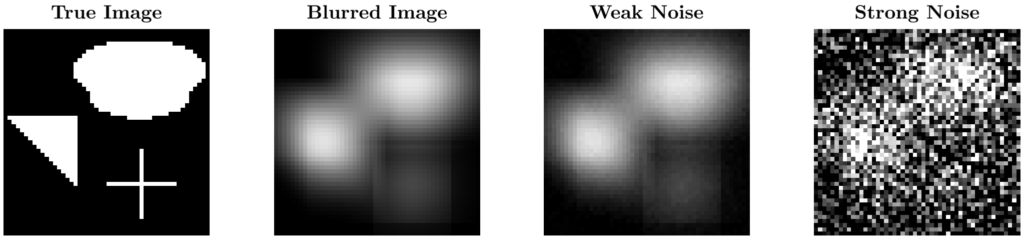

Figure 4. *From left to right: target image, blurred image, reliable data, and corrupted data, used for comparing noncentered and centered parameterizations under low-rank independence sampling for 2D deblurring.*

In cases where $\boldsymbol{A}$ is rank deficient, one can use the random walk Metropolis step previously discussed, but other approaches are possible.

3.2. Illustration with Image Deblurring. To illustrate use of the non-centered parameterization in our proposed LRIS algorithm, we consider again the 2D image deblurring example from Section 4.1 of the manuscript. Here, the target image $\boldsymbol{x}$ and the blurring operator $\boldsymbol{A}$ are the same as before. However, we create two different observed datasets $\boldsymbol{b}_i = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{\epsilon}_i, \;\; i = 1, 2$, with two different levels of noise. One set is strongly corrupted with 50% noise, $Var(\boldsymbol{\epsilon}_1) = 0.5^2\|\boldsymbol{b}\|_\infty^2\boldsymbol{I}$, and the other contains much less noise, $Var(\boldsymbol{\epsilon}_2) = 0.01^2\|\boldsymbol{b}\|_\infty^2\boldsymbol{I}$, and thus is more informative about the true solution $\boldsymbol{x}$. The target image, blurred image, and noisy data sets are displayed in Supplementary Figure 4.

For the centered parameterization, we use the same priors on $\boldsymbol{x}, \mu$ and σ in model (1) as in the manuscript, namely $\boldsymbol{\Gamma}_{\text{pr}}^{-1} = \boldsymbol{L}^\top\boldsymbol{L}$ with $\boldsymbol{L} = -\Delta + \delta\boldsymbol{I}$ and $\epsilon = 0.1$ in the Gamma(ϵ, ϵ) priors on μ and σ to approximate scale invariant objective priors. We use the same low-rank Metropolis-Hastings-within-Gibbs algorithm as in the manuscript. To implement the non-centered parameterization, we use our proposed low-rank independent sampler to update $\boldsymbol{x}$ and the independent sampler proposed by [1] to update $\sigma^{-1/2}$. Our interest here is not in posterior inference about the target image, but in the mixing behavior of these two parameterizations applied to data with different amounts of noise. Thus, rather than running to and diagnosing convergence, we run only 5,000 iterations of each Markov chain to study autocorrelation and how quickly the chains appear to be moving through their support. Under each configuration, we run three chains in parallel, with μ and $\boldsymbol{x}$ initialized by drawing them randomly from their prior distributions. The prior precision σ is initialized at $0.1, 6$, and 25 for chains 1, 2, and 3, respectively.

Supplementary Figure 5 displays the trace plots of the three σ chains under each of the four combinations of data and parameterization. With severely noisy data, the noncentered parameterization shows stronger mixing than the centered parameterization. The opposite is true with the reliable data containing only 1% noise. The differences under the reliable data are particulary striking, where we see improvement in the centered paramterization but a very severe degradation in performance of the noncentered parameterization. The drift in all of the chains is indicative of considerable autocorrelation, and this is confirmed by examining the autocorrelation functions plotted in Supplementary Figure 6 and the estimated lag 1 and lag 50 correlation coefficients in Supplementary Table 1. Each chain suffers from high lag 1

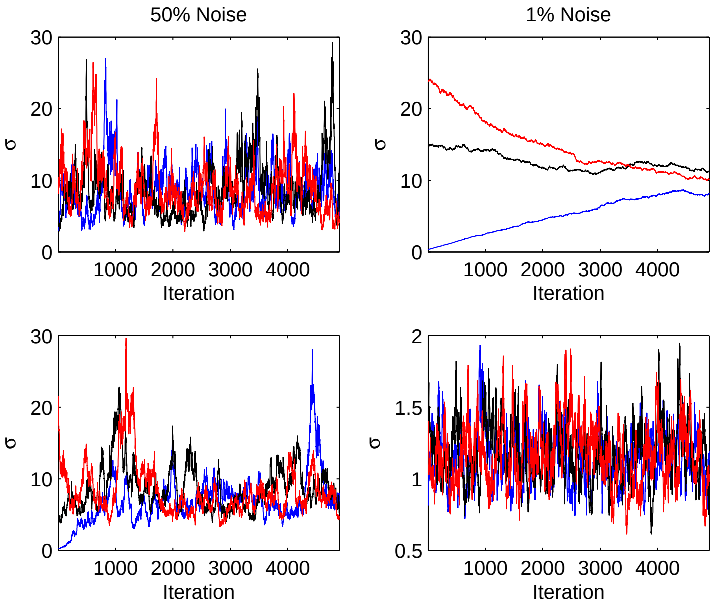

Figure 5. *Trace plots of the σ chains obtained from centered and non-centered parameterizations using both datasets displayed in Supplementary Figure 4. The left column corresponds to the noisy data and right column to the reliable data. The top row are plots obtained from the noncentered parameterization, and the bottom row is obtained from the centered parameterization.*

autocorrelation, but it decays faster under the non-centered parameterization for the noisy data, and faster for the centered parameterization under the more reliable data. The decay of the autocorrelation is particularly poor for the noncentered parameterization with the reliable data.

Table 1

Estimated autocorrelation coefficients of one σ chain, for each data / parameterization combination in Supplementary simulation study.

	Reliable		Noisy	
	Lag 1	Lag 50	Lag 1	Lag 50
CP	0.968	0.174	0.995	0.774
NCP	0.999	0.974	0.982	0.348

The results of this illustration demonstrate the ease with which either the non-centered or centered parameterization can be used in combination with our proposed LRIS. Further, the relative performance of non-centered versus centered parameterizations previously observed in [15, 16, 1] is still present when using our more computationally efficient alternative. In particular, for strongly corrupted data, $\boldsymbol{x}$ is more strongly determined through the prior than the data, so that a non-centered parameterization is preferable. More reliable data impose the constraint that $\boldsymbol{b} \approx \sigma^{-1/2}\boldsymbol{A}\boldsymbol{z}$, severely degrading the performance of the non-centered parameterizaton. Thus, when using our proposed approach, a practitioner can still appeal to

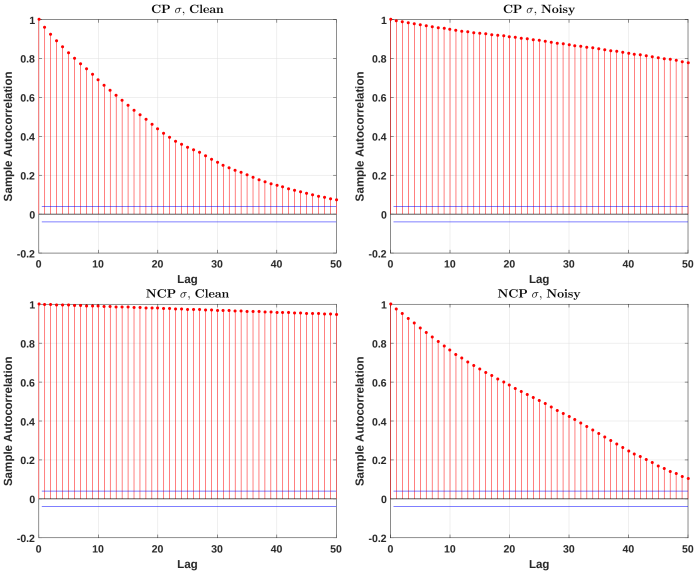

Figure 6. *Estimated autocorrelation functions of one σ chain from the Supplementary 2D deblurring simulation study, each obtained from centered (top row) and non-centered (bottom row) parameterizations using both weakly (left column) and strongly (right column) corrupted data.*

the same considerations when choosing a more effective parameterization for convergence of their MCMC algorithm.

4. Adaptive LRIS. The target rank k for the low-rank approximation may not be known in practice, or may depend explicitly on the parameters μ and σ. Here we outline a simple adaptive strategy for determining the target rank. The basic idea is to increment the rank till the acceptance ratio meets the desired tolerance.

1. Inputs: user defined tolerance δ, initial rank k_{init}, rank increment k_{inc}
2. Start $k = k_{\text{init}}$
3. While not `converged`
4. Check if acceptance ratio is higher than $1-\delta$. If, yes, mark `converged`.
5. Increment the rank $k \leftarrow k + k_{\text{inc}}$.
6. End While

Monitoring the acceptance ratio is important to determine a stopping criterion for the algorithm described above. We suggest several different strategies:

1. The acceptance ratio can be monitored empirically as the independence sampler is exploring the distribution. If the acceptance ratio is too small, the target rank may be incremented.
2. In the case that the exact eigenpairs are used, we derive a lower bound for $w(\boldsymbol{x})$. From the proof of Theorem 1, we find that if $\widehat{\boldsymbol{H}} = \boldsymbol{V}_k\boldsymbol{\Lambda}_k\boldsymbol{V}_k^\top$

$$w(\boldsymbol{x}) \geq \exp\left(-\frac{\mu}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2\|\boldsymbol{H}-\widehat{\boldsymbol{H}}\|_2\right) = \exp\left(-\frac{\mu\lambda_{k+1}}{2}\|\boldsymbol{L}\boldsymbol{x}\|_2^2\right).$$

Here λ_{k+1} is the largest eigenvalue that is discarded. This suggests that the target rank k is the minimum index which satisfies

$$\lambda_{k+1} \leq \frac{2}{\mu\|\boldsymbol{Lx}\|_2^2} \log \frac{1}{1-\delta}.$$

This will ensure $w(\boldsymbol{x}) \geq 1-\delta$, where δ is a user-defined parameter. It is worth mentioning that the entire eigendecomposition need not be computed, nor recommended. In practice, the eigenpairs can be computed in an incremental fashion.

3. In the randomized low-rank approach, λ_{k+1} may not be available. However, it may be easy to estimate $\|\boldsymbol{H}-\widehat{\boldsymbol{H}}\|_2$ cheaply using a randomized estimator; see [9, Lemma 4.1]. Then to ensure that $w(\boldsymbol{x}) \geq 1-\delta$, it is required that $\|\boldsymbol{H}-\widehat{\boldsymbol{H}}\|_2 \leq \varepsilon$, where ε satisfies

$$\varepsilon \leq \frac{2}{\mu\|\boldsymbol{Lx}\|_2^2} \log \frac{1}{1-\delta}.$$

The low-rank approximation can be adaptively and efficiently determined using the adaptive range finding algorithm, see [9, Algorithm 4.2].

5. NMR Relaxometry Simulation.

In this section, we investigate the performance of our approach on a large-scale ill-posed inverse problem with a different low-rank nature than the CT example. We consider the problem of nuclear magnetic resonance (NMR) relaxometry in which nuclear magnetic moments are used to infer physical or chemical properties of a medium. The deterioration (or *relaxation*) of the signal over time is analyzed using various electromagnetic pulse sequences and acquisition strategies. The longitudinal and traverse relaxation times of the medium, denoted T_1 and T_2, respectively, can be recovered from these acquisitions. The goal of 2D NMR relaxometry is to reconstruct the joint distribution of the relaxation times, $f(T_1, T_2)$, from measured signals gathered at different experimental times, denoted $g(\tau_1, \tau_2)$. Mathematically, the forward model is a Fredholm integral equation of the first kind, $g(\tau_1, \tau_2) = \int_0^{\hat{T}_2} \int_0^{\hat{T}_1} K(\tau_1, \tau_2, T_1, T_2) f(T_1, T_2)\, \mathrm{d}T_1 \mathrm{d}T_2$. As in the 1D image deblurring example, the measurement noise is typically assumed to be Gaussian and the problem is to reconstruct f from g and K.

We use the implementation in the `Matlab` package `IR Tools` [6] to simulate NMR data. In this example, the kernel is separable, and after discretization we obtain the linear inverse problem $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{b}$ is the vectorization of the data g and $\boldsymbol{x}$ the vectorization of the unknown f. Unlike the other examples we consider in this work, the forward map $\boldsymbol{A}$ and its transpose are not constructed explicitly, but their matrix-vector products are computed using available function handles. We again enforce smoothness in $\boldsymbol{x}$ by specifying $\boldsymbol{L} = -\Delta + \delta \boldsymbol{I}$ and corrupt the data with one percent Gaussian random noise. With the default discretizations, the data have dimension $m = 65,536$ and the unknown has dimension $n = 16,384$. The noisy data $\boldsymbol{b}$ and unknown $\boldsymbol{x}$ are displayed in Figure 7. We compute approximate eigenvalues of the prior preconditioned Hessian using Randomized SVD with $\ell = 1000$. The sharp decay in the eigenvalues, also displayed in Figure 7, illustrates the severe ill-posedness of the NMR relaxometry inverse problem.

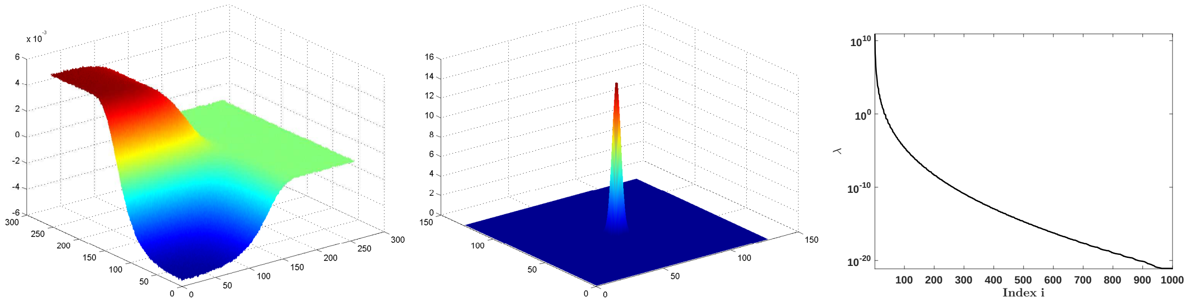

Figure 7. *Noisy measured data* $\boldsymbol{b}$ *(left) and true relaxation distribution* $\boldsymbol{x}$ *(center) in the NMR relaxometry example. Approximate spectrum of* $\boldsymbol{H}$ *(right) shows significant decay within the first few hundred eigenvalues.*

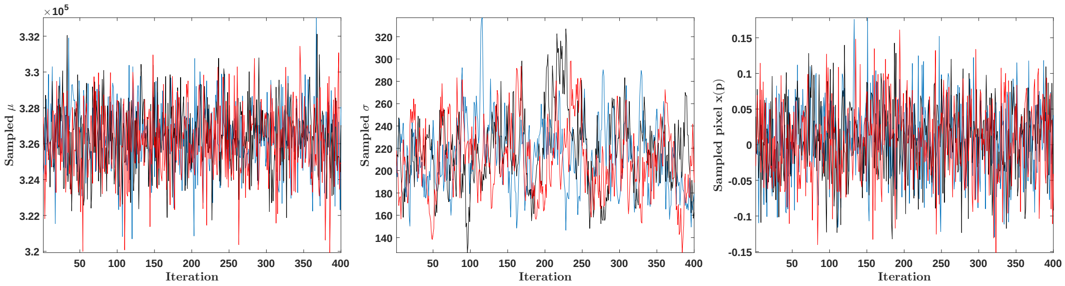

Figure 8. *Trace plots for the three thinned chains using Gamma priors in the NMR relaxometry example. Left: noise precision* μ*. Center: prior precision* σ*. Right: a randomly chosen pixel of the image* $\boldsymbol{x}$*.*

For implementation, we perform the same experiment as in Section 4.2 of the manuscript with the same vague conjugate Gamma priors on μ and σ. Analyzing the acceptance rates for different eigenvalue truncation levels as well as the spectral decay in Figure 7, we specify the truncation level $k = 300$ (average acceptance rate $\approx 100\%$). We simulate three Markov chains for 40,000 iterations each, initializing each chain with random draws from the priors. To reduce the autocorrelation in the chains, they are thinned to retain every 50th draw, after which the first 400 draws are discarded as the burn-in period. Approximate convergence of the chains is diagnosed via trace plots and autocorrelation plots, displayed in Figures 8 and 9. The PSRFs for μ and σ are 0.99 and 1.01, respectively. The computations are carried out in `MATLAB 2013a` on a Dell Precision T3600 desktop PC running Windows 7 Enterprise with an Intel Xeon E5-1660 3.30GHz CPU and 64GB RAM. The total computation time is 13,263 seconds, or about 3.7 hours.

Figure 10 displays the posterior mean of $\boldsymbol{x}$ based upon the output of the MCMC at approximate convergence, as well as the approximate densities of μ and σ. While we obtain a reasonable estimate of the distribution of relaxation times, it is strongly attenuated compared to the true solution and exhibits background fluctuations (including negative values). The attenuation, though, is similar to the solution obtained in [6] via conjugate gradient least squares, though our posterior mean produces a smoother estimate. Indeed, [6] remark that the NMR problem is extremely challenging and that even the iterative methods they consider require many thousands of iterations to produce interpretable estimates of $\boldsymbol{x}$.

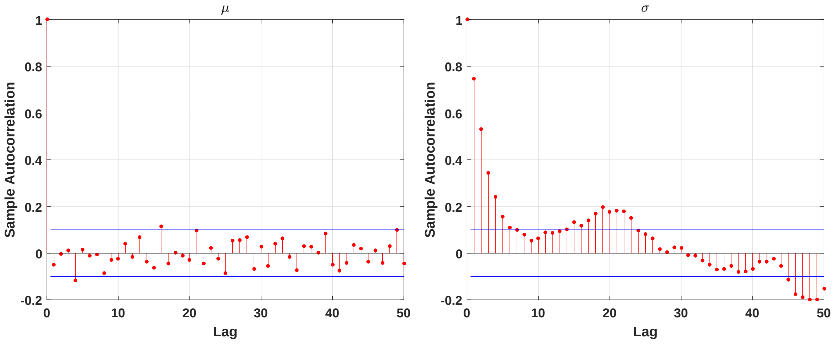

Figure 9. *Autocorrelation plots for one thinned chain using Gamma priors in the NMR relaxometry example. Left: noise precision μ. Right: prior precision σ.*

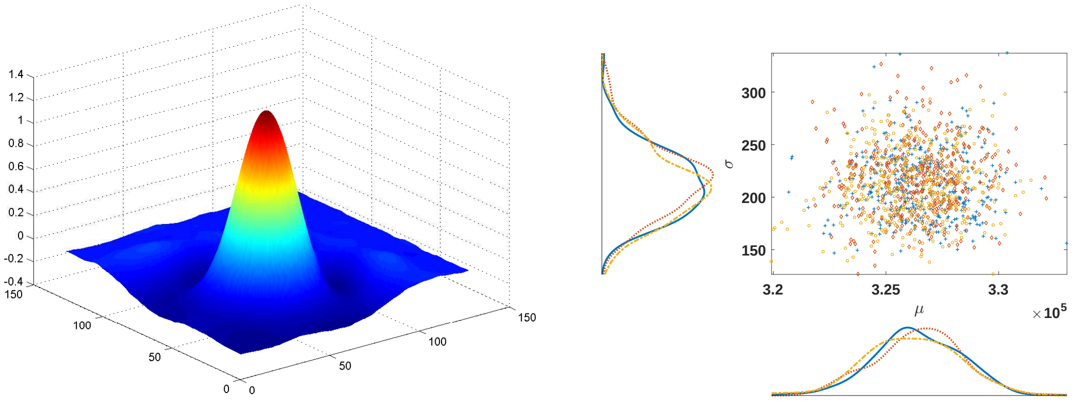

Figure 10. *Posterior mean estimate of $\boldsymbol{x}$ (left) and smoothed histograms of μ and σ (right) in the NMR relaxometry example.*

Our low-rank approach offers considerable computational advantages over standard MCMC, thus making it feasible to reconstruct a competitive solution along with appropriate measures of uncertainty. It is demonstrated in [6] that non-negativity constraints on $\boldsymbol{x}$ can yield an improved solution. In our framework, this suggests that a truncated Gaussian or lognormal prior on $\boldsymbol{x}$ may produce solutions superior to that produced by the Gaussian prior. Such an investigation is beyond the scope of this work and left for future research.

6. Supplementary Figures.

In Supplementary Figure 11, we compare the (vague) Gamma priors on μ and σ with the marginal posterior distributions estimated from the MCMC samples of the joint posterior. The prior for μ has been scaled by a factor of 10,000 for visualization. Note the significant difference between the distributions, indicating that strong Bayesian learning has occurred.

Trace plots of the samples from the thinned MCMC chains using weakly informative priors in the CT example (Section 4.2) are displayed in Supplementary Figure 12, and autocorrelation plots are displayed in Supplementary Figure 13. In Supplementary Figure 14, we plot

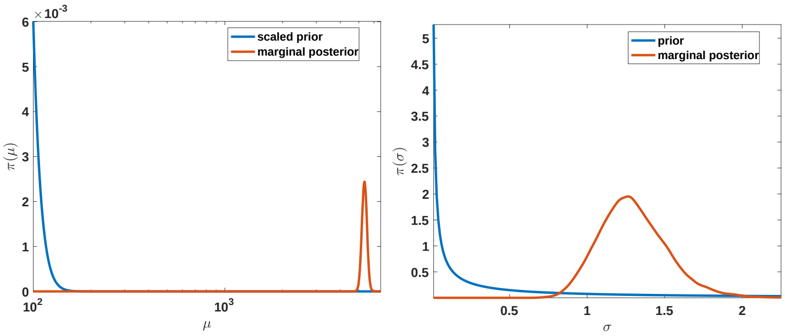

Figure 11. *Priors and marginal posterior distributions for μ (left) and σ (right) in the 2D deblurring example.*

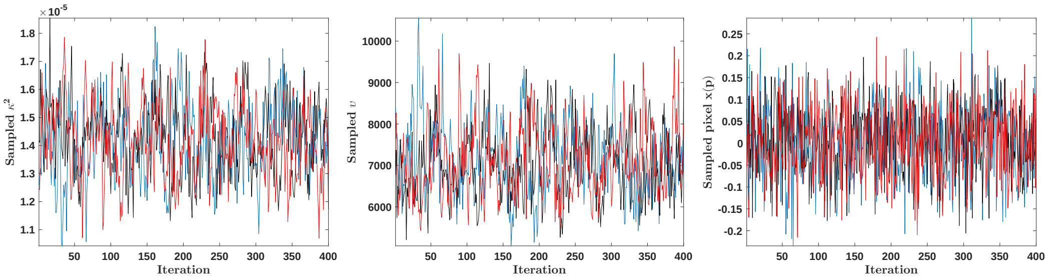

Figure 12. *Trace plots for the thinned chains in the CT image reconstruction example. Left: noise variance κ^2. Center: Variance ratio $\upsilon = \tau^2/\kappa^2$. Right: A randomly chosen pixel of the image $\boldsymbol{x}$.*

the cumulative averages of the parameters κ^2 and υ of these samples. The PSRFs for these parameters were both 1.00. Even after thinning, we see some correlation between the samples.

We next present convergence diagnostics for the CT example with the conjugate Gamma prior instead of proper Jeffreys to justify the comparison in the manuscript. We use vague Gamma priors for μ and σ as in the 2D example, and the same prior for $\boldsymbol{x}$ as in the original CT experiment. We again use Randomized SVD with target rank $\ell = 5000$ in the low-rank proposal distribution for $\boldsymbol{x}$. We simulate three (randomly initialized) Markov chains using the MCMC algorithm with our proposed approach for 20,000 iterations. These chains are then thinned and the burn-in period discarded to produce an equal Monte Carlo sample size as in the previous experiment. Trace plots, autocorrelation plots, and PSRFs are used to determine approximate convergence of the chains. The trace plots are displayed in Supplementary Figure 15 and the autocorrelation plots for μ and σ are in Supplementary Figure 16. The PSRFs for μ and σ were 1.00 and 1.01 respectively.

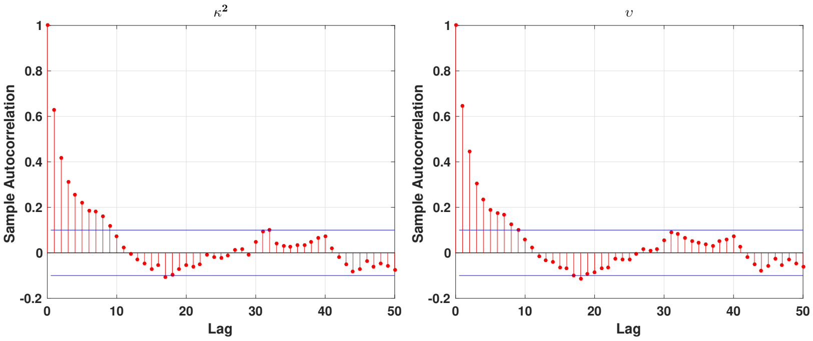

Figure 13. *Autocorrelation plots for the thinned chains in the CT image reconstruction example. Left: noise variance κ^2. Right: Variance ratio $\upsilon = \tau^2/\kappa^2$.*

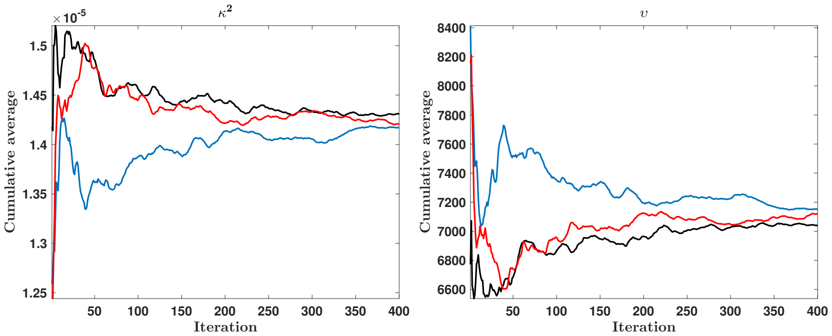

Figure 14. *Cumulative averages of κ^2 and υ for each of the thinned MCMC chains in the CT example.*

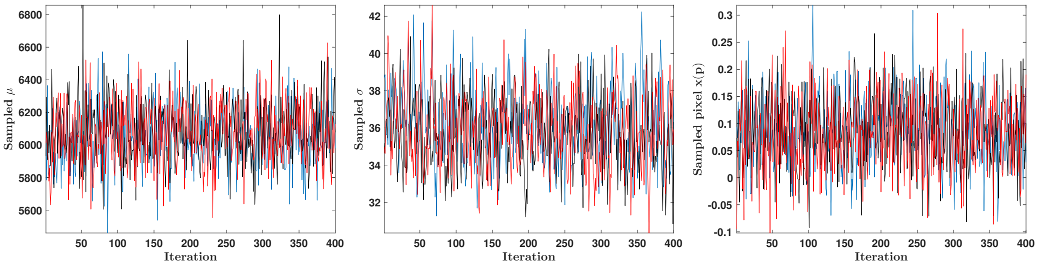

Figure 15. *Trace plots for the three thinned chains using Gamma priors in the CT image reconstruction example. Left: noise precision μ. Center: prior precision σ. Right: a randomly chosen pixel of the image $\boldsymbol{x}$.*

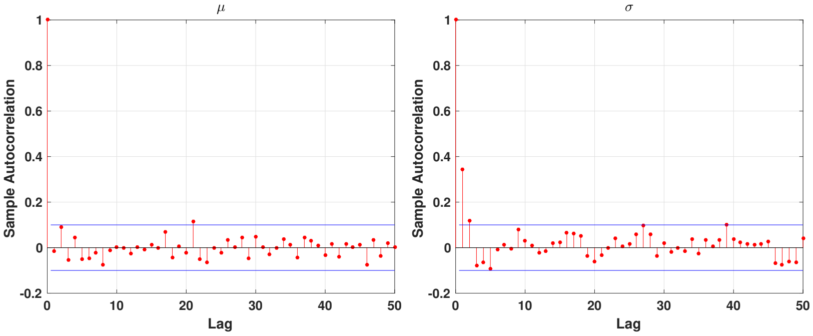

Figure 16. *Autocorrelation plots for one thinned chain using Gamma priors in the CT image reconstruction example. Left: noise precision μ. Right: prior precision σ.*

REFERENCES

[1] S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart. Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):511–544, 2014.

[2] J. M. Bardsley. MCMC-based image reconstruction with uncertainty quantification. *SIAM J. Sci. Comput.*, 34(3):A1316–A1332, 2012.

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

[4] D. A. Brown, G. S. Datta, and N. A. Lazar. A Bayesian generalized CAR model for correlated signal detection. *Stat. Sinica*, 27:1125–1153, 2017.

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

[6] S. Gazzola, P. C. Hansen, and J. G. Nagy. Ir tools: A matlab package of iterative regularization methods and large-scale test problems. *arXiv preprint arXiv:1712:05602*, 2017.

[7] A E Gelfand, S K Sahu, and B P Carlin. Efficient parameterisations for normal linear mixed models. *Biometrika*, 82(3):479–488, 1995.

[8] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–533, 2006.

[9] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[10] P. C. Hansen. Regularization tools version 4.0 for Matlab 7.3. *Numer. Algorithms*, 46:189–194, 2007.

[11] D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *J. Am. Stat. Assoc.*, 103:570–583, 2008.

[12] H. Jeffreys. *Theory of Probability.* Oxford University Press, Cambridge, 3rd edition, 1961.

[13] A. Johnson, G. L. Jones, and R. C. Neath. Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Stat. Sci.*, 28(3):360–375, 2013.

[14] Kutner, M. and Nachtsheim, C. and Neter, J. and Li, W. *Applied Linear Statistical Models.* McGraw-Hill/Irwin, Boston, 5th edition, 2005.

[15] O. Papaspiliopoulos and G. O. Roberts. Non-Centered Parameterisations for Hierarchical Models and Data Augmentation. *Bayesian Statistics*, 7:307–326, 2003.

[16] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A General Framework for the Parametrization of Hierarchical Models. *Stat. Sci.*, 22(1):59–73, 2007.

[17] N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse Bayesian regression and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 9*, pages 501–538. Oxford University Press, 2010.

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

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

[20] J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *J. Stat. Plan. Infer.*, 136(7):2144–2162, 2006.

[21] C. B. Shaw. Improvements of the resolution of an instrument by numerical solution of an integral equation. *J. Math. Anal. Appl.*, 37:83–112, 1972.

[22] G. C. Tiao and W. Tan. Bayesian analysis of random-effect models in the analysis of variance, I. Posterior distribution of variance components. *Biometrika*, 51:37–53, 1965.