Invited Discussion

Xiaodong Yang* and Jun S. Liu†

1 Introduction

We congratulate Professors Giacomo Zanella and Gareth Roberts for their path-breaking work in analyzing Gibbs sampling algorithms for a class of highly practical Bayesian hierarchical models. Together with their previous work, Papaspiliopoulos and Roberts (2003) and Papaspiliopoulos et al. (2020), their multigrid decomposition strategy elegantly reduces a high-dimensional Gibbs sampling algorithm to independent low-dimensional components so that the convergence rate of the Gibbs sampler can be determined analytically. These are extremely interesting and encouraging results. Throughout of the article, we will refer to this work of Zanella and Roberts (2021) as "Z&R" for simplicity.

The *multigrid decomposition* serves a central role in the whole theory established in the aforementioned series of papers. An intuition behind this decomposition is that lower-level mean statistics are sufficient for posterior inference on upper-level parameters, with lower-level parameters practically marginalized out. For example, Papaspiliopoulos and Roberts (2003) show that, for model (1.1) below, the posterior distribution of $(\mu, \bar{a})$ is independent of that of $(a_1 - \bar{a}, \cdots, a_I - \bar{a})$.

At the first glance, we cannot help notice that the intuition behind Z&R's multigrid decomposition is quite different from that of either the classical deterministic multigrid methods (McCormick, 1987) or *multigrid Monte Carlo* methods (Goodman and Sokal, 1989; Liu and Sabatti, 2000). These latter multigrid strategies, as originally motivated by the design of efficient numerical partial differential equation (PDE) solvers, are typically constructed artificially to accelerate the convergence of the algorithms by iterating between finer-grid and coarser-grid updates. In contrast, Z&R's multigrid decomposition is a decomposition of the given parameter space implied by the algorithm itself (under a specific parametrization). Furthermore, Z&R show that Gibbs sampling for the upper level of their multigrid decomposition converges slower than that for the lower level (Theorem 11), whereas in classical multigrid methods the upper levels are so constructed that their associated MCMC samplers converge faster than those of the lower levels (Goodman and Sokal, 1989; Liu and Sabatti, 2000).

Despite these fundamental differences between the multigrid decomposition and multigrid Monte Carlo, we are very much inspired by Z&R's insightful formulation and will discuss some potential extensions of their work in the rest of the article. To illustrate our main ideas, we start by focusing on the simplest model:

$$y_{ij} = \mu + a_i + \epsilon_{ij}, \quad i \in [1:I], j \in [1:J], \tag{1.1}$$

*School of Gifted Young, University of Science and Technology of China, Hefei, China, yangxiaodong0912@gmail.com

†Department of Statistics, Harvard University, Cambridge, US, jliu@stat.harvard.edu

which can be seen as either a two-level hierarchical model or a one-factor crossed-effects model. In the rest of the article, we use notation $\vec{a}$ to represent a vector. For example, $\vec{a}_i$ used in Section 2 is an ℓ-dimensional vector. Boldface letters are used to represent collections of effects. For example, we write $\boldsymbol{a} = (a_1, \cdots, a_I)$, and $\bar{a}$ for its mean. We also denote $\mathbf{1}_k = (1, \cdots, 1)^\top \in \mathbb{R}^{k\times 1}$ and $\mathbb{I}_k$ for $k \times k$ identity matrix. For a matrix M, $\|M\|_2 = \sqrt{\sigma_{\max}(M^\top M)}$ denotes its spectral norm.

2 Vector hierarchical models

Our main goal here is to extend the framework of (1.1) to consider the vector-version of the model, as shown in (2.1). This type of models is not uncommon in practice and is a prototype of more complex realistic models. For example, the observed vector $\vec{y}_{ij}$ may represent several types of medical measurements (e.g., blood pressure, cholesterol level, weight, height, etc) of individual j in group i, and these measurements are certainly correlated within each individual. After presenting results for (2.1), we will comment on its potential extensions.

2.1 Non-centering model and convergence rate

Let us begin with an extension of model (1.1) by replacing the scalars with vectors to arrive at the following model.

Model S2m (Symmetric two-level model with **non-centering** parametrization)**.** *Suppose*

$$\vec{y}_{ij} = \vec{\mu} + \vec{a}_i + \vec{\epsilon}_{ij}, \quad i \in [1:I], j \in [1:J], \tag{2.1}$$

where $\vec{y}_{ij}, \vec{\mu}, \vec{a}_i, \vec{\epsilon}_{ij} \in \mathbb{R}^\ell$*, and* $\vec{\epsilon}_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_e)$ *(i.e., i.i.d. multivariate Gaussian). We impose a flat prior on* $\vec{\mu}$ *and another multivariate Gaussian* $\mathcal{N}(0, \Sigma_a)$ *on each* $\vec{a}$*. Here* Σ_e *and* Σ_a *are two positive definite* $\ell \times \ell$ *matrices.*

For this model, we can write down the joint posterior distribution as

$$p(\vec{\mu}, \vec{\boldsymbol{a}} \mid \vec{y}) \propto \exp\left[-\frac{1}{2}\sum_{i,j}(\vec{y}_{ij} - \vec{\mu} - \vec{a}_i)^\top \Sigma_e^{-1}(\vec{y}_{ij} - \vec{\mu} - \vec{a}_i) - \frac{1}{2}\sum_i \vec{a}_i^\top \Sigma_a^{-1}\vec{a}_i\right]. \tag{2.2}$$

A standard Gibbs Sampler to sample from the posterior distribution $p(\vec{\mu}, \vec{\boldsymbol{a}} \mid \vec{\boldsymbol{y}})$ is defined as follows.

Sampler GS(0). *Initialize* $\vec{\mu}(0)$ *and* $\vec{\boldsymbol{a}}(0)$ *and then iterate*

1. *Sample* $\vec{\mu}(s+1)$ *from* $p(\vec{\mu} \mid \vec{\boldsymbol{a}}(s), \vec{\boldsymbol{y}})$*;*
2. *Sample* $\vec{a}_i(s+1)$ *from* $p(\vec{a}_i \mid \vec{\mu}(s+1), \vec{\boldsymbol{y}})$ *for* $i = 1, \ldots, I$*, independently.*

Using the same notations as in Z&R, we define $\bar{\vec{a}} = \sum_i \vec{a}_i / I$ to be mean and

$$\delta\vec{a}_i = \vec{a}_i - \bar{\vec{a}}, \quad \delta\vec{\boldsymbol{a}} = (\delta\vec{a}_1, \cdots, \delta\vec{a}_I)$$

as the residual. Given this notation, we derive the following factorization

$$p(\vec{\mu}, \vec{\boldsymbol{a}} \mid \vec{\boldsymbol{y}}) = p(\vec{\mu}, \bar{\vec{a}} \mid \vec{\boldsymbol{y}}) \times p(\delta\vec{\boldsymbol{a}} \mid \vec{\boldsymbol{y}}). \tag{2.3}$$

This factorization paves the way for the following multigrid decomposition.

Before stating and proving our result, we introduce a lemma without proof to compute the L^2 convergence rate of some two-component Gaussian Gibbs sampler.

Lemma 2.1. *Let the target distribution $\pi(q_1, q_2)$, where $q_1, q_2 \in \mathbb{R}^\ell$, be a 2ℓ-dimensional Gaussian distribution with $var(q_1) = \Sigma_{11}$, $var(q_2) = \Sigma_{22}$, and $cov(q_1, q_2) = \Sigma_{12}$. The convergence rate of the Gibbs sampler that iterates between conditional sampling $[q_1 \mid q_2]$ and $[q_2 \mid q_1]$ is equal to the squared spectral norm $\|\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\|_2^2$.*

Remark. This lemma is an easy consequence of Theorem 1 in Roberts and Sahu (1997), in which the generated Markov chain is recognized as a multivariate AR(1) process. See also Section 5.1, Liu et al. (1994), for an elementary proof based on *maximal correlations*, as this quantity can also be interpreted as the *maximal correlation* between q_1 and q_2.

Theorem 2.1. *Let $\{\vec{\mu}(t), \vec{\boldsymbol{a}}(t)\}$ be the Markov chain generated by either the standard Gibbs sampler. Then the functionals $\{\delta\vec{\boldsymbol{a}}(t)\}$ and $\{\vec{\mu}(t), \bar{\vec{a}}(t)\}$ evolve as two independent Markov chains. Furthermore, the L^2-convergence rate of the sampler is*

$$\rho_0 = \left\| \left(J\Sigma_e^{-1}\right)^{1/2} \left(\Sigma_a^{-1} + J\Sigma_e^{-1}\right)^{-1/2} \right\|_2^2. \tag{2.4}$$

Proof. The decomposition directly follows from the following two identities

$$p\left[\vec{\mu}(s+1) \mid \vec{\boldsymbol{a}}(s), \vec{\boldsymbol{y}}\right] = p\left[\vec{\mu}(s+1) | \vec{a}(s), \vec{\boldsymbol{y}}\right], \tag{2.5}$$

$$p\left[\vec{a}(s+1), \delta\vec{\boldsymbol{a}}(s+1) \mid \vec{\mu}(s+1), \vec{\boldsymbol{y}}\right] = p\left[\vec{a}(s+1) | \vec{\mu}(s), \vec{\boldsymbol{y}}\right] \times p\left[\delta\vec{\boldsymbol{a}}(s+1) \mid \vec{y}\right]. \tag{2.6}$$

Moreover, the latter identity further implies that $\{\delta\vec{\boldsymbol{a}}(t)\}$ carries out exact sampling. So the convergence rate of $\{\vec{\mu}(t), \vec{\boldsymbol{a}}(t)\}$ is actually determined by the rate of $\{\vec{\mu}(t), \bar{\vec{a}}(t)\}$. The latter chain converges to the following joint-normal stationary distribution

$$\begin{aligned} p(\vec{\mu}, \bar{\vec{a}} \mid \vec{y}) \propto & \exp\left[-\frac{IJ}{2}\vec{\mu}^\top\Sigma_e^{-1}\vec{\mu} - \frac{1}{2}\bar{\vec{a}}^\top\left(I\Sigma_a^{-1} + IJ\Sigma_e^{-1}\right)\bar{\vec{a}}\right] \\ & \times \exp\left[-IJ\vec{\mu}^\top\Sigma_e^{-1}\bar{\vec{a}} + IJ\bar{\vec{y}}^\top\Sigma_e^{-1}(\vec{\mu} + \bar{\vec{a}})\right], \end{aligned}$$

where we write $\bar{\vec{y}} \triangleq \sum_{i,j}\vec{y}_{ij}/IJ$. This is a Markov chain in a 2ℓ-dimensional space induced by the block-wise two-component Gibbs sampler. In contrast, the original chain is of dimension $(I+1)\ell$. The final result then follows from Lemma 2.1. □

Remark. If we choose dimension $\ell = 1$ and replace Σ_e and Σ_a with σ_e^2 and σ_a^2, respectively, the convergence rate becomes

$$\rho_0 = \frac{J\sigma_e^{-2}}{\sigma_a^{-2} + J\sigma_e^{-2}},$$

which coincides with Proposition 3 in Papaspiliopoulos et al. (2020).

2.2 Convergence rate for centering model

Inspired by Z&R, we seek to give a theoretical guidance towards centering (2.1) or non-centering (2.7) parametrizations.

Model S2m (Symmetric two-level model with **centering** parametrization)**.** *Suppose*

$$\vec{y}_{ij} \sim \mathcal{N}(\vec{\alpha}_i, \Sigma_e), \quad \vec{\alpha}_i \sim \mathcal{N}(\vec{\mu}, \Sigma_a),\ i \in [1:I], j \in [1:J], \tag{2.7}$$

where $\vec{y}_{ij}, \vec{\mu}, \vec{\alpha}_i \in \mathbb{R}^\ell$*. Same as before, a flat prior is imposed on* $\vec{\mu}$*. Here* Σ_e *and* Σ_a *are two positive definite* $\ell \times \ell$ *matrices.*

Sampler GS(1). *Initialize* $\vec{\mu}(0)$ *and* $\vec{\boldsymbol{\alpha}}(0)$ *and then iterate*

1. *Sample* $\vec{\mu}(s+1)$ *from* $p(\vec{\mu} \mid \vec{\boldsymbol{\alpha}}(s), \vec{\boldsymbol{y}})$*;*
2. *Sample* $\vec{\alpha}_i(s+1)$ *from* $p(\vec{\alpha}_i \mid \vec{\mu}(s+1), \vec{\boldsymbol{y}})$ *for* $i = 1, \ldots, I$ *independently.*

Almost in the same manner, we offer the following theorem.

Theorem 2.2. *Let* $\{\vec{\mu}(t), \vec{\boldsymbol{\alpha}}(t)\}$ *be the Markov chain generated by the sampler GS(1). Then the functionals* $\{\delta\vec{\boldsymbol{\alpha}}(t)\}$ *and* $\{\vec{\mu}(t), \bar{\vec{\alpha}}(t)\}$ *evolve as two independent Markov chains. Furthermore, the* L^2*-convergence rate of* $\{\vec{\mu}(t), \vec{\boldsymbol{\alpha}}(t)\}$ *is*

$$\rho_1 = \left\| \left(\Sigma_a^{-1}\right)^{1/2} \left(\Sigma_a^{-1} + J\Sigma_e^{-1}\right)^{-1/2} \right\|_2^2. \tag{2.8}$$

Optimal Parameterization Strategy: If $\rho_0 \le \rho_1$, then choose the non-centering parameterization (2.1); otherwise, choose the centering parameterization (2.7).

When dimension $\ell = 1$, (2.8) becomes $\rho_1 = \sigma_a^{-2}/(\sigma_a^{-2} + J\sigma_e^{-2})$. This strategy can be adaptively used when the variances are unknown. Specifically, in one iteration, after sampling $\hat{\sigma}_a^2, \hat{\sigma}_e^2$, we compare $J\hat{\sigma}_e^{-2}/(\hat{\sigma}_a^{-2} + J\hat{\sigma}_e^{-2})$ and $\hat{\sigma}_a^{-2}/(\hat{\sigma}_a^{-2} + J\hat{\sigma}_e^{-2})$, and choose the optimal parameterization accordingly. Back to the case of known variances, a direct benefit is that we can always achieve a convergence rate bounded by $1/2$ since $\rho_0 + \rho_1 = 1$, regardless of what values σ_a^2, σ_e^2 are (Papaspiliopoulos and Roberts, 2003). Corollary 2 in Z&R proposes an optimal parametrization strategy for 3-level models and gives a constant rate upper bound $2/3$ therein.

However, in a multi-dimensional case with $\ell > 1$, the rates found in Theorem 2.1 and Theorem 2.2 do not necessarily sum up to 1. Though the parameterization strategy still applies, it does not necessarily give a constant rate upper bound. If both covariance matrices are diagonal, i.e., $\Sigma_a = \mathrm{diag}(1/\tau_1^a, \cdots, 1/\tau_\ell^a)$ and $\Sigma_e = \mathrm{diag}(1/\tau_1^e, \cdots, 1/\tau_\ell^e)$, then we have

$$\rho_0 = \max_{1 \le i \le \ell} \left[\frac{J\tau_i^e}{\tau_i^a + J\tau_i^e} \right], \quad \rho_1 = \max_{1 \le i \le \ell} \left[\frac{\tau_i^a}{\tau_i^a + J\tau_i^e} \right].$$

Applying the optimal parametrization strategy component-wise is of interest in this non-correlated case. That is, we may introduce a "centering" indicator variable C of

dimension ℓ, indicating which of the ℓ components use centering and which use non-centering parameterization. In this way, we may still be able to obtain the rate bound 1/2.

When Σ_a and Σ_e become general non-diagonal covariance matrices, the picture becomes more complicated. It will be of great interest to develop some methodological guidance on how to approach this problem. The constant rate bound 1/2 as discussed above is no longer guaranteed, and it is entirely possible that both rates are close to 1. We speculate that one may extend the "centering" indicator C to be a continuous vector to allow "partial-centering" (more about this issue in Section 4).

It is also not too difficult to extend these results to more complex structures such as three-level vector hierarchical models and vector crossed-effects models, although the formulae would grow more complicated and the design of the optimal parameterization may no longer be possible. The authors' insights and suggestions along this direction would be very much welcome.

3 Incorporating regression covariates

Zanella and Roberts mainly focus on hierarchical models with certain symmetry conditions for data without individual-level covariates. Mixed-effects models, which accommodate individual-level variability and are very commonly used in practice, seem to have not been directly covered by Z&R. Our goal here is to consider possible ways to extend the authors' multigrid decomposition technique to this more complex class of models.

3.1 Linear mixed effects models

To extend and see the limits of multigrid decomposition, we consider the following simple extension, which just replaces the intercept term μ with a linear combination of p covariates with a fixed coefficient vector. Previously, Gao and Owen (2019) attempted to tackle the computational efficiency of this model (3.1). But their results give loose bounds while requiring mild conditions.

Model SR (Symmetric two-level mixed-effect model)**.** *Suppose*

$$y_{ij} = X_{ij}^\top \beta + a_i + \epsilon_{ij}, \quad i \in [1:I], j \in [1:J], \tag{3.1}$$

where ϵ_{ij} is i.i.d. normal random variables with mean 0 and variance σ_e^2. Moreover, $X_{ij}, \beta \in \mathbb{R}^p$ (column vectors) are known covariates and unknown coefficients respectively. We then impose a standard Bayesian model specification assuming $a_i \sim \mathcal{N}(0, \sigma_a^2)$ and $\beta \sim \mathcal{N}(0, \Sigma_0)$.

Essential full-rank conditions should be imposed on the design matrix. Requiring $p < I$, we denote the $I \times p$ matrix as

$$\bar{X} \triangleq (\bar{X}_1, \ldots, \bar{X}_I)^\top,$$

where $\bar{X}_i = J^{-1}\sum_j X_{ij} \in \mathbb{R}^p$. A further natural requirement is that $\bar{X}$ is of rank p. Then, we can define a $p \times I$ matrix $P = (\bar{X}^\top \bar{X})^{-1/2}\bar{X}^\top$. We also introduce another $(I-p)\times I$ matrix L such that $L^\top L + P^\top P = \mathbb{I}_I$ (i.e., the identity matrix of dimension I). Note that $P^\top P = \bar{X}(\bar{X}^\top \bar{X})^{-1}\bar{X}^\top$ and $PP^\top = \mathbb{I}_p$. Let $\boldsymbol{X} = \{X_{ij}\}$.

Sampler GS (Regression)**.** *Initialize* $\beta(0)$ *and* $\boldsymbol{a}(0)$ *and then iterate*

1. *Sample* $\beta(s+1)$ *from* $p(\beta \mid \boldsymbol{a}(s), \boldsymbol{X}, \boldsymbol{y})$*;*
2. *Sample* $a_i(s+1)$ *from* $p(a_i \mid \boldsymbol{a}(s+1), \boldsymbol{X}, \boldsymbol{y})$ *for all* i*.*

Theorem 3.1. *Let* $\{\beta(t), \boldsymbol{a}(t)\}$ *be the Markov chain generated by the standard Gibbs sampler. Then the two functionals* $\{L\boldsymbol{a}(t)\}$ *and* $\{\beta(t), \bar{X}^\top \boldsymbol{a}(t)\}$ *evolve as two independent Markov chains. Furthermore, the* L^2*-convergence rate of* $\{\beta(t), \boldsymbol{a}(t)\}$ *is*

$$\rho = \frac{J^2\sigma_e^{-4}}{\sigma_a^{-2} + J\sigma_e^{-2}} \left\| \left(\bar{X}^\top \bar{X}\right)^{1/2} \left(\Sigma_0^{-1} + \sum_{i,j} X_{ij} X_{ij}^\top \sigma_e^{-2} \right)^{-1/2} \right\|_2^2 . \tag{3.2}$$

Proof. It is easy to write down the likelihood function and prior:

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \beta, \boldsymbol{a}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J} \exp\left[-\frac{1}{2\sigma_e^2}(y_{ij} - X_{ij}^\top \beta - a_i)^2 \right],$$

$$p(\beta, \boldsymbol{a}) \propto \exp\left[-\frac{1}{2}\beta^\top \Sigma_0^{-1}\beta - \frac{1}{2\sigma_a^2}\sum_{i=1}^{I} a_i^2 \right].$$

The posterior distribution is

$$\begin{aligned} p(\beta, \boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}) &\propto \exp\left[-\frac{1}{2}\beta^\top \Sigma_0^{-1}\beta - \frac{1}{2\sigma_a^2}\sum_i a_i^2 - \frac{1}{2\sigma_e^2}\sum_{i,j}(y_{ij} - X_{ij}^\top \beta - a_i)^2 \right] \\ &\propto \exp\left[-\frac{1}{2}\beta^\top \left(\Sigma_0^{-1} + \sum_{i,j} X_{ij} X_{ij}^\top \sigma_e^{-2} \right)\beta - \frac{1}{2}\left(\frac{1}{\sigma_a^2} + \frac{J}{\sigma_e^2} \right)\sum_i a_i^2 \right] \\ &\quad \times \exp\left[-\frac{1}{\sigma_e^2}\sum_{i,j} a_i X_{ij}^\top \beta \frac{J}{\sigma_e^2}\sum_i a_i \bar{y}_i + \frac{1}{\sigma_e^2}\sum_{ij} y_{ij} X_{ij}^\top \beta \right]. \end{aligned}$$

We should especially focus on the cross term

$$\sum_{ij} a_i X_{ij}^\top \beta = \sum_{i=1} a_i (J\bar{X}_i^\top)\beta = J\boldsymbol{a}^\top \bar{X}\beta.$$

Furthermore, we also find that

$$\sum_i a_i^2 = \boldsymbol{a}^\top \boldsymbol{a} = \|P\boldsymbol{a}\|^2 + \|L\boldsymbol{a}\|^2 = \boldsymbol{a}^\top \bar{X}\left(\bar{X}^\top \bar{X}\right)^{-1}\bar{X}^\top \boldsymbol{a} + \|L\boldsymbol{a}\|^2.$$

The distribution of $\boldsymbol{a}$ is actually equivalent to the joint distribution of $(\bar{X}^\top \boldsymbol{a}, L\boldsymbol{a})$, since $(\bar{X}, L^\top)$ is an invertible $I \times I$ matrix. Hence, we derive the following factorization

$$p(\beta, \boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}) = p(\beta, \bar{X}^\top \boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}) \times p(L\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}). \tag{3.3}$$

We shall also deduce the following identities

$$p\left[\beta(s+1) \mid \boldsymbol{a}(s), \boldsymbol{y}, \boldsymbol{X}\right] = p\left[\beta(s+1) \mid \bar{X}^\top \boldsymbol{a}(s), \boldsymbol{y}, \boldsymbol{X}\right], \tag{3.4}$$

$$p\left[\bar{X}^\top \boldsymbol{a}(s+1), L\boldsymbol{a}(s) \mid \beta(s), \boldsymbol{y}, \boldsymbol{X}\right] = p\left[\bar{X}^\top \boldsymbol{a}(s+1) \mid \beta(s), \boldsymbol{y}, \boldsymbol{X}\right] p\left[L\boldsymbol{a}(s) \mid \boldsymbol{y}, \boldsymbol{X}\right], \tag{3.5}$$

which imply the multigrid decomposition. Again, convergence rate ρ is controlled by the convergence rate of $\{\beta(t), \bar{X}^\top \boldsymbol{a}(t)\}$. The joint target distribution of $\{\beta, \bar{X}^\top \boldsymbol{a}\}$ is

$$p(\beta, \bar{X}^\top \boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left[-\frac{1}{2}\beta^\top \left(\Sigma_0^{-1} + \sum_{i,j} X_{ij}^\top X_{ij} \sigma_e^{-2}\right) \beta - \frac{J}{\sigma_e^2} \boldsymbol{a}^\top \bar{X} \beta\right]$$
$$\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_a^2} + \frac{J}{\sigma_e^2}\right) \boldsymbol{a}^\top \bar{X} \left(\bar{X}^\top \bar{X}\right)^{-1} \bar{X}^\top \boldsymbol{a}\right]$$

By Lemma 2.1, the L^2 convergence rate is equal to the squared maximal correlation between β and $\bar{X}^\top \boldsymbol{a}$. □

Remark 1. If we set $p = 1, X_{ij} \equiv 1$, then $\bar{X}_i = 1$, $\bar{X}^\top \bar{X} = I$ and $\sum_{ij} X_{ij}^\top X_{ij} = IJ$. By placing a flat prior on μ, we just replace Σ_0^{-1} with 0 in (3.2). Henceforth, Theorem 3.1 reduces to $\rho = J\sigma_e^{-2}/(\sigma_a^{-2} + J\sigma_e^{-2})$, in this case.

Remark 2. Theorem 3.1 implies that p summary statistics $\bar{X}^\top \boldsymbol{a}$ of the lower level parameters are sufficient for the inference of upper level parameters β, with $L\boldsymbol{a}$ marginalized out.

Remark 3. Further note that (3.2) is invariant if the variance terms are scaled simultaneously. Specifically, (3.2) remains the same if we replace $\left(\Sigma_0, \sigma_a^2, \sigma_e^2\right)$ by $\left(r\Sigma_0, r\sigma_a^2, r\sigma_e^2\right)$ where $r > 0$. Moreover, another common rotation invariance in Bayesian linear regression applies to our result: (3.2) remains the same if the pair (Σ_0, X_{ij}) is replaced with $\left(R^\top \Sigma_0 R, RX_{ij}\right)$, where R is a $p \times p$ orthogonal matrix.

We further note that the multigrid decomposition techniques do not naturally extend to more complex structures. Roughly speaking, both nested structures (such as $y_{ijk} = X_{ijk}^\top \beta + a_i + b_{ij} + \epsilon_{ijk}$) and crossed structures (such as $y_{ijk} = X_{ijk}^\top \beta + a_i + b_j + \epsilon_{ijk}$) would bring in a new cross term "$\boldsymbol{a}^\top \boldsymbol{b}$", which is hard to handle. Can we still obtain an elegant decomposition for these models?

Indeed, many researchers have studied the general linear mixed-effects model:

$$y = X^\top \beta + Z^\top u + \epsilon, \tag{3.6}$$

where, in the first part, β is common to all individuals as in a typical linear regression framework, and u represents random effects (e.g., Z can be dummy variables). For example, if Z represents one categorical variable with I categories (using a dummy variable

representation), this general form (3.6) reduces to the simple model (3.1) considered before.

Model (3.6) with arbitrary Z, however, has an identical mathematical representation as a standard linear regression model (i.e., one can simply treat (X, Z) as covariates) although the prior distributions for β and u may differ substantially. Compared with the models handled in Z&R, a key thing we have lost in the general model (3.6) seems to be the strong symmetry that can be used to decompose the involved variables into meaningful levels. A curious question is: how far we can push so that we can still have certain meaningful decomposition?

3.2 Implications for general linear regression models

Linear model formulation of two-level hierarchical model

We can recast the multigrid decomposition of Z&R for both centering and non-centering parameterizations of model (1.1) in the context of general Bayesian linear regression via covariate orthogonalization.

Non-centering parametrization By setting $\beta = (a_1, \cdots, a_I)^\top$ and

$$y = (y_{11}, y_{12}, \cdots, y_{1I}, y_{21}, \cdots, y_{IJ})^\top \in \mathbb{R}^{IJ\times 1}, X = (\mathbb{I}_I \otimes \mathbf{1}_J)^\top \quad \text{(Kronecker product)}, \tag{3.7}$$

the simple linear model $y = \mu\mathbf{1}_{IJ} + X^\top\beta + \epsilon$ is equivalent to model (1.1). The decomposition can be seen as imposing a linear transformation by replacing β with $A\beta$, where the first row of A is $\frac{1}{\sqrt{I}}\mathbf{1}_I^\top$ and A is $I \times I$ orthogonal. In the following, we omit the terms involving y when dealing with the posterior, cause these terms do not affect the covariance of unknown parameters. With flat prior on μ and independent $\mathcal{N}(0, 1/\tau_a)$ on each a_i, the posterior is

$$\begin{aligned} p(\beta, \mu \mid y, X) &\propto \exp\left(-\frac{1}{2}\beta^\top(\tau_e XX^\top + \tau_a\mathbb{I})\beta - \tau_e\mu\mathbf{1}_{IJ}^\top X^\top\beta - \frac{IJ\tau_e}{2}\mu^2\right) \\ &= \exp\left(-\frac{1}{2}(A\beta)^\top(\tau_e AXX^\top A^\top + \tau_a\mathbb{I})(A\beta) - \tau_e\mu[A\beta]_1 - \frac{IJ\tau_e}{2}\mu^2\right). \end{aligned}$$

Moreover, $[AXX^\top A^\top]_{i1} = [AXX^\top A^\top]_{1i} = 0$ for any $i \geq 2$, which means that the first column of $X^\top A^\top$ is orthogonal to the other columns. Thus, $(\mu, [A\beta]_1)$ and $[A\beta]_{2:I}$ are independent *a posteriori*. The first component corresponds to $(\mu, \bar{a})$ and the latter one is a representation of the residual δa. The multigrid decomposition is then built upon this orthogonalization. To investigate the potential of this orthogonalization-based view, we consider the following general linear regression model.

Centering parametrization Model (1.1) can also be written as

$$y_{ij} \sim \mathcal{N}(\alpha_i, 1/\tau_e), \quad \alpha_i \sim \mathcal{N}(\mu, 1/\tau_a),\ i \in [1:I], j \in [1:J]. \tag{3.8}$$

Set y, X, and β exactly the same way as (3.7), we have an equivalent model:

$$y = X^\top \beta + \epsilon, \quad \beta \sim \mathcal{N}(\mu \mathbf{1}_I, 1/\tau_a \mathbb{I}_I), \quad \epsilon \sim \mathcal{N}(0, 1/\tau_e \mathbb{I}_{IJ}). \tag{3.9}$$

Intuitively, we use a new prior on β with a latent variable μ. With flat prior on μ, the posterior is

$$p(\beta, \mu \mid y, X) \propto \exp\left[-\frac{1}{2}\beta^\top \left(\tau_e X X^\top + \tau_a \mathbb{I}\right)\beta + \tau_a \mu \mathbf{1}_I^\top \beta - \frac{I\tau_a}{2}\mu^2\right].$$

We can apply the same linear transformation A as before.

Extension to general linear models

Model LM. *Suppose $X_1 \in \mathbb{R}^{p_1 \times n}$, $X_2 \in \mathbb{R}^{p_2 \times n}$ are two sets of covariates and consider*

$$y = X_1^\top \beta_1 + X_2^\top \beta_2 + \epsilon, \tag{3.10}$$

where $\beta_i \in \mathbb{R}^{p_i}, (i = 1, 2)$ are unknown coefficients. Error $\epsilon \in \mathbb{R}^n$ is modeled as i.i.d. $\mathcal{N}(0, 1/\tau_e)$. Independent priors $\mathcal{N}(0, 1/\tau_1 \mathbb{I}_{p_1})$ and $\mathcal{N}(0, 1/\tau_2 \mathbb{I}_{p_2})$ are imposed on β_1 and β_2 respectively.

Assume $r = \operatorname{rank}(X_1 X_2^\top)$, we conduct SVD to find $B_i \in \mathbb{R}^{r \times p_i}, (i = 1, 2)$ with orthonormal rows and diagonal $Q = \operatorname{diag}(\lambda_1, \cdots, \lambda_r)$ such that

$$X_1 X_2^\top = B_1^\top Q B_2. \tag{3.11}$$

By constructing orthogonal matrices $A_i \in \mathbb{R}^{p_i \times p_i}$, $i = 1, 2$, as completions of B_1 and B_2, respectively, i.e., A_i and B_i share the same r first rows, we have the following result.

Theorem 3.2. *Consider a Markov chain $\{\beta_1(s), \beta_2(s)\}$ generated by a systematic Gibbs sampler alternating between conditional sampling $[\beta_1 \mid \beta_2]$ and $[\beta_2 \mid \beta_1]$. Define $\theta_i = \left(\theta_i^{(1)}, \cdots, \theta_i^{(p_i)}\right)^\top = A_1 \beta_i$. Then, the evolution of $\{\theta_1(s), \theta_2(s)\}$ is equivalent to that of $\{\beta_1(s), \beta_2(s)\}$. If the first r columns of $X_i^\top A_i^\top$ are orthogonal to the rest $p_i - r$ columns*

$$[X_i^\top A_i^\top]_{1:n,k_1} \perp [X_i^\top A_i^\top]_{1:n,k_2}, \quad \forall k_1 \le r < k_2, \tag{3.12}$$

the evolutions of $\{\theta_1^{(1:r)}(s), \theta_2^{(1:r)}(s)\}$, $\{\theta_1^{((r+1):p_1)}(s)\}$ and $\{\theta_2^{((r+1):p_2)}(s)\}$ are independent.

Proof. We start by writing out the joint posterior

$$p(\beta \mid y, X) \propto \exp\left[-\tau_e \beta_1^\top X_1 X_2^\top \beta_2 - \frac{1}{2}\sum_{i=1}^{2} \beta_i^\top \left(\tau_e X_i X_i^\top + \tau_i \mathbb{I}_{p_i}\right)\beta_i\right] \tag{3.13}$$

$$= \exp\left[-\tau_e \left(\theta_1^{(1:r)}\right)^\top Q \theta_2^{(1:r)} - \frac{1}{2}\sum_{i=1}^{2} \theta_i^\top \left(\tau_e A_i^\top X_i X_i^\top A_i + \tau_i \mathbb{I}_{p_i}\right)\theta_i\right] \tag{3.14}$$

$$= p\left(\theta_1^{(1:r)}, \theta_2^{(1:r)} \mid y, X\right) \prod_{i=1}^{2} p\left(\theta_i^{((r+1):p_i)} \mid y, X\right), \tag{3.15}$$

where the last equality follows from the condition (3.12). Based on these identities,

$$\begin{aligned}
p\left(\theta_1^{((r+1):p_1)}(s+1) \mid y, X, \theta_2(s)\right) &= p\left(\theta_1^{((r+1):p_1)}(s+1) \mid y, X\right), \\
p\left(\theta_2^{((r+1):p_2)}(s+1) \mid y, X, \theta_1(s+1)\right) &= p\left(\theta_2^{((r+1):p_2)}(s+1) \mid y, X\right), \\
p\left(\theta_1^{(1:r)}(s+1) \mid y, X, \theta_2(s)\right) &= p\left(\theta_1^{(1:r)}(s+1) \mid y, X, \theta_2^{(1:r)}(s)\right), \\
p\left(\theta_2^{(1:r)}(s+1) \mid y, X, \theta_1(s+1)\right) &= p\left(\theta_2^{(1:r)}(s+1) \mid y, X, \theta_1^{(1:r)}(s+1)\right),
\end{aligned}$$

the conclusion of the theorem is thus proved. □

One implication of the result is that the multigrid decomposition developed for (1.1) is non-trivial in the sense that condition (3.12) must be imposed on the covariate matrix. Recall that we have written out the dummy variables X explicitly for (1.1), and thus verified this condition implicitly for the linear model form of (1.1).

Centering for linear models Model (3.10) with its priors can be rewritten as

$$y = X_2^\top \beta_2 + \epsilon, \quad \beta_2 \sim \mathcal{N}(M\beta_1, 1/\tau_2 \mathbb{I}_{p_2}), \quad \beta_1 \sim \mathcal{N}(0, 1/\tau_1 \mathbb{I}_{p_1}), \quad \epsilon \sim \mathcal{N}(0, 1/\tau_e \mathbb{I}_n), \tag{3.16}$$

to mimic the centering parametrization, where $M \in \mathbb{R}^{p_2 \times p_1}$ such that $X_1^\top = X_2^\top M$,[1] assuming that M exists.

Now the posterior distribution is

$$p(\beta \mid y, X) \propto \exp\left[\tau_2 \beta_1^\top M^\top \beta_2 - \frac{1}{2}\beta_2^\top \left(\tau_e X_2 X_2^\top + \tau_2 \mathbb{I}_{p_2}\right) \beta_2\right] \tag{3.17}$$

$$\times \exp\left[-\frac{1}{2}\beta_1^\top \left(\tau_2 M^\top M + \tau_1 \mathbb{I}_{p_1}\right) \beta_1\right]. \tag{3.18}$$

Let the SVD of M be

$$M = B_1^\top Q B_2, \tag{3.19}$$

where $Q \in \mathbb{R}^r, r = \text{rank}(M)$. Again we denote the complement of B_i as A_i. Then we require the following condition

$$[X_2^\top A_2^\top]_{1:n,k_1} \perp [X_2^\top A_2^\top]_{1:n,k_2}, \quad \forall k_1 \le r < k_2. \tag{3.20}$$

to validate a similar multigrid decomposition. Again, this condition automatically holds for the two-level hierarchical model, but do not hold in general.

[1] For the simplest model (1.1), we actually use $M = \mathbf{1}_I$.

3.3 Thoughts and speculations

In both the non-centering and centering formulations, conditions (3.12) and (3.20) most likely do not hold for an arbitrary design matrix X. Thus, a multigrid decomposition similar to that of Z&R seems difficult to come by. Some natural questions arise: Does a useful multigrid decomposition exist for a general linear regression model in some other ways? If so, what would be a correct construction? If not, how can we gain more insights on the Gibbs sampler for a general Bayesian regression model (3.10)? Can we find a good matrix M so that the convergence rate of the Gibbs sampler corresponding to (3.17) is faster than that based on (3.13)? What if the Gibbs sampler has more than two components?

Besides the Gaussian prior we have studied here, many other prior distributions have been proposed to accommodate both sparsity and biases in coefficient estimations, including *spike-and-slab* priors (Mitchell and Beauchamp, 1988), *horseshoe* priors (Carvalho et al., 2010), *neuronized* priors (Shin and Liu, 2021), and so on. Can one extend Z&R's and our results to accommodate other priors that are more appropriate for high-dimensional problems? The Gaussian spike-and-slab prior may be a most likely solvable case?

4 Partial centering for improving convergence

4.1 Partial-centering for two-level models

Partial centering provides a continuous trade-off between centering and non-centering. With these parametrizations (e.g., centering, non-centering, partial centering) sharing almost the same mathematical formulation, can we derive the most efficient algorithm by optimizing over various parametrizations including not only parametrizations covered by Z&R, but also those dictating partial centering?

Inspired by an example in Liu and Wu (1999) to demonstrate the power of *parameter expansion*, Papaspiliopoulos and Roberts (2003) proposed the following *partial centering parametrization* in by introducing a constant $0 \le A \le 1$:

Model S2 (Symmetric two-level model with **partial centering** parametrization)**.** *Suppose*

$$y_{ij} \sim \mathcal{N}((1-A)\mu + a_i, \sigma_e^2), \quad a_i \sim \mathcal{N}(A\mu, \sigma_a^2), \quad i \in [1:I], j \in [1:J], \tag{4.1}$$

where $y_{ij}, \mu, a_i in \mathbb{R}$. *Same as before, a flat prior is imposed on* μ.

A similar standard Gibbs sampler as GS(0) and GS(1) can be easily implemented. With $A = 0$, (4.1) reduces to non-centering parametrization; whereas with $A = 1$, (4.1) reduces to centering parametrization. For a general A, Papaspiliopoulos and Roberts (2003) also offered the convergence rate of the standard Gibbs sampler as

$$\rho_A = \frac{\left(A\sigma_a^{-2} - (1-A)J\sigma_e^{-2}\right)^2}{\left(\sigma_a^{-2} + J\sigma_e^{-2}\right)\left(A^2\sigma_a^{-2} + (1-A)^2\sigma_e^{-2}\right)}. \tag{4.2}$$

One surprising fact is that $\rho_{A^*} = 0$ for $A^* = J\sigma_e^{-2}/(\sigma_a^{-2} + J\sigma_e^{-2})$, implying that we achieve exact sampling in one step via this optimal partial centering parameterization. Note that this A^* also results in the fact that μ and $\bar{a}$ are independent *a posteriori*.

4.2 Partial-centering for three-level models

It is of great interest to extend this flexible parametrization scheme to other models. We here provide an illustration via a slightly more complex model.

Model S3 (Symmetric three-level model with **partial centering** parametrization)**.** *With constants $A, B, C \in \mathbb{R}$, suppose*

$$\begin{aligned} y_{ijk} &\sim \mathcal{N}((1-A-C)\mu + (1-B)a_i + b_{ij}, \sigma_e^2), \\ b_{ij} &\sim \mathcal{N}(Ba_i + C\mu, \sigma_b^2), \quad a_i \sim \mathcal{N}(A\mu, \sigma_a^2), \end{aligned} \tag{4.3}$$

where $y_{ij}, \mu, a_i, \epsilon_{ij} \in \mathbb{R}$ and i, j, k range from 1 to I, J, K respectively. Same as before, a flat prior is imposed on μ.

Sampler GS (A, B, C)**.** *Initialize $\mu(0)$, $\boldsymbol{a}(0)$, $\boldsymbol{b}(0)$ and then iterate*

1. *Sample $\mu(s+1)$ from $p(\mu \mid \boldsymbol{a}(s), \boldsymbol{b}(s), \boldsymbol{y})$;*
2. *Sample $a_i(s+1)$ from $p(a_i \mid \mu(s+1), \boldsymbol{b}(s), \boldsymbol{y})$ for all i;*
3. *Sample $b_{ij}(s+1)$ from $p(b_{ij} \mid \mu(s+1), \boldsymbol{a}(s+1), \boldsymbol{y})$ for all i, j.*

If we select (A, B, C) from $\{0,1\}^2 \times \{0\}$, (4.3) reduces to the four parametrizations considered in Sections 2 and 3 of Z&R, respectively. Defining hierarchical models as trees, Section 7 of Z&R develop an abstract theory to deal with various parametrizations including the partial ones here, but they do not provide more insights for cases $(A, B, C) \notin \{0,1\}^2 \times \{0\}$. Let $\tau_a = I\sigma_a^{-2}, \tau_b = IJ\sigma_b^{-2}, \tau_e = IJK\sigma_e^{-2}$ be the rescaled precisions. We have the following result.

Theorem 4.1. *If $(\tau_b + \tau_e)^2\tau_a + \tau_b\tau_e(\tau_b - \tau_e) \neq 0$, the prescribed Gibbs sampler can achieve exact sampling in one step via suitable scalings of A, B, C.*

Proof. First, we define $\delta\boldsymbol{\beta} = \left(\delta^{(0)}\boldsymbol{\beta}, \delta^{(1)}\boldsymbol{\beta}, \delta^{(2)}\boldsymbol{\beta}\right)$ exactly the same as equation (3.1) in Z&R, where $\delta^{(0)}\boldsymbol{\beta} = \left(\mu, \bar{a}, \bar{b}\right), \bar{a} = \sum_i a_i/I, \bar{b} = \sum_{ij} b_{ij}/IJ$. Apply Theorem 9 in Z&R to conclude that $\{\delta^{(0)}\boldsymbol{\beta}\}, \{\delta^{(1)}\boldsymbol{\beta}\}, \{\delta^{(2)}\boldsymbol{\beta}\}$ evolve independently for the prescribed Gibbs sampler.

Then, applying Theorem 11 of Z&R, we derive the following ordering

$$\rho_{(A,B,C)} = \rho\left(\delta^{(0)}\boldsymbol{\beta}\right) \geq \rho\left(\delta^{(1)}\boldsymbol{\beta}\right) \geq \rho\left(\delta^{(2)}\boldsymbol{\beta}\right) = 0.$$

At last, we have to deal with the posterior distribution of $\delta^{(0)}\boldsymbol{\beta}$, which is a 3-dim Gaussian. The evolution of $\{\delta^{(0)}\boldsymbol{\beta}(t)\}$ is simply characterized by a systematic scan Gibbs

sampler, scanning according to $\mu \to \bar{a} \to \bar{b} \to \mu$. By Liu et al. (1995), to obtain the convergence rate of a systematic scan Gibbs sampler, it suffices to know about pairwise correlations

$$r_1 = \operatorname{corr}(\mu, \bar{a}) = \frac{BC\tau_b + A\tau_a - (1-A-C)(1-B)\tau_e}{\sqrt{C^2\tau_b + A^2\tau_a + (1-A-C)^2\tau_e}\sqrt{B^2\tau_b + \tau_a + (1-B)^2\tau_e}},$$
$$r_2 = \operatorname{corr}(\mu, \bar{b}) = \frac{C\tau_b - (1-A-C)\tau_e}{\sqrt{C^2\tau_b + A^2\tau_a + (1-A-C)^2\tau_e}\sqrt{\tau_b + \tau_e}},$$
$$r_3 = \operatorname{corr}(\bar{a}, \bar{b}) = \frac{B\tau_b - (1-B)\tau_e}{\sqrt{\tau_b + \tau_e}\sqrt{B^2\tau_b + \tau_a + (1-B)^2\tau_e}}.$$

By Liu et al. (1995) and Roberts and Sahu (1997), we find that $\rho_{(A^*,B^*,C^*)} = 0$ for

$$A^* = \frac{\tau_b\tau_e(\tau_b - \tau_e)}{(\tau_b+\tau_e)^2\tau_a + \tau_b\tau_e(\tau_b-\tau_e)},\ B^* = \frac{\tau_e}{\tau_b+\tau_e},\ C^* = \frac{\tau_a\tau_e(\tau_b+\tau_e)}{(\tau_b+\tau_e)^2\tau_a + \tau_b\tau_e(\tau_b-\tau_e)},$$

due to vanishing correlations $r_1 = r_2 = r_3 = 0$. □

An analytical formula is available for the convergence rate of the standard Gibbs sampler $\mathrm{GS}(A, B, C)$ even for general A, B, C. But this general formula is a little complicated and out of the scope of this article. We believe that this formula may help us understand the experimental phase transitions depicted in Figure 4 of Z&R, and further enhance our understanding towards different parametrizations. A direct question is whether exact sampling in one step is possible for less symmetric 2, 3-level hierarchical models.

We end this section by raising more questions. Does the partial centering trick generalize to more complex structures with more confounding factors and deeper hierarchies? How do we develop partial centering for vector hierarchical models discussed in Section 2 to design a better Gibbs sampler? Can we go beyond Gaussian priors to perform it in other cases, like the Poisson example in Section 5 of Z&R?

5 Concluding remarks

Although Z&R's multigrid decomposition has little to do with the classical multigrid idea for both numerical PDEs and Monte Carlo simulations, their decomposition provides a key insight to the understanding of the convergence of Gibbs sampling for Bayesian hierarchical models. This insight naturally leads to a constructive strategy for designing better Gibbs sampling algorithms via reparametrization for such models. Our article centers on the possibilities of extending this decomposition strategy to more complex, yet structured, Bayesian models, and to include more options (e.g., parameter expansion) for algorithmic optimization. We specifically analyzed a few concrete examples, one in each direction. Our results are both encouraging and challenge-revealing. On one hand, we have obtained some analytical expressions of the convergence rates of various Gibbs samplers, from which we may derive an optimal parameterization; on

the other hand, we find that situations become much more complex and the optimal parameterization may not exist or computable in high-dimensional cases, such as vector hierarchical models and mixed effects models. In summary, we find that the decomposition framework established by Z&R is both elegant and practical, and that much future endeavor is warranted for exploring and exploiting their framework.

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1367

Gao, K. and Owen, A. B. (2019). "Estimation and inference for very large linear mixed effects models." *Statist. Sinica*. MR4260743. doi: https://doi.org/10.5705/ss.202018.0029. 1361

Goodman, J. and Sokal, A. D. (1989). "Multigrid Monte Carlo method. conceptual foundations." *Physical Review D*, 40(6): 2035. 1357

Liu, J. S. and Sabatti, C. (2000). "Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation." *Biometrika*, 87(2): 353–369. MR1782484. doi: https://doi.org/10.1093/biomet/87.2.353. 1357

Liu, J. S., Wong, W. H., and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes." *Biometrika*, 81(1): 27–40. MR1279653. doi: https://doi.org/10.1093/biomet/81.1.27. 1359

Liu, J. S., Wong, W. H., and Kong, A. (1995). "Covariance structure and convergence rate of the Gibbs sampler with various scans." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 157–169. MR1210432. 1369

Liu, J. S. and Wu, Y. N. (1999). "Parameter expansion for data augmentation." *Journal of the American Statistical Association*, 94(448): 1264–1274. MR1731488. doi: https://doi.org/10.2307/2669940. 1367

McCormick, S. F. (1987). *Multigrid methods*. SIAM. MR0972752. doi: https://doi.org/10.1137/1.9781611971057. 1357

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 1367

Papaspiliopoulos, O. and Roberts, G. O. (2003). "Non-centered parameterisations for hierarchical models and data augmentation." In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, volume 307. Oxford University Press, USA. MR2003180. 1357, 1360, 1367

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2020). "Scalable inference for crossed random effects models." *Biometrika*, 107(1): 25–40. MR4064138. doi: https://doi.org/10.1093/biomet/asz058. 1357, 1359

Roberts, G. O. and Sahu, S. K. (1997). "Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2): 291–317. MR1440584. doi: https://doi.org/10.1111/1467-9868.00070. 1359, 1369

Shin, M. and Liu, J. S. (2021). "Neuronized priors for Bayesian sparse linear regression." *Journal of the American Statistical Association*, (just-accepted): 1–43. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 1367

Zanella, G. and Roberts, G. (2021). "Multilevel linear models, Gibbs samplers and multigrid decompositions." *Bayesian Analysis*. 1357