Bayesian Analysis (2021) **16**, Number 4, pp. 1309–1391

Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions (with Discussion)[*]

Giacomo Zanella[†] and Gareth Roberts[‡]

Abstract. We study the convergence properties of the Gibbs Sampler in the context of posterior distributions arising from Bayesian analysis of conditionally Gaussian hierarchical models. We develop a multigrid approach to derive analytic expressions for the convergence rates of the algorithm for various widely used model structures, including nested and crossed random effects. Our results apply to multilevel models with an arbitrary number of layers in the hierarchy, while most previous work was limited to the two-level nested case. The theoretical results provide explicit and easy-to-implement guidelines to optimize practical implementations of the Gibbs Sampler, such as indications on which parametrization to choose (e.g. centred and non-centred), which constraint to impose to guarantee statistical identifiability, and which parameters to monitor in the diagnostic process. Simulations suggest that the results are informative also in the context of non-Gaussian distributions and more general MCMC schemes, such as gradient-based ones.

Keywords: Gibbs Sampler, convergence rates, hierarchical models, multigrid decomposition, centred and non-centred parametrizations, statistical identifiability.

1 Introduction

Markov chain Monte Carlo (MCMC) is established as the computational workhorse of most Bayesian statistical analyses for complex models. For hierarchical models with conditionally conjugate priors, the Gibbs sampler (Gelfand and Smith, 1990; Smith and Roberts, 1993) remains one of the most natural algorithm of choice, thanks to its simplicity of implementation and low computational cost per iteration (thanks to conjugacy and conditional independence). Nonetheless, speed of convergence of the resulting Markov chain can be a major issue and can be highly sensitive to the model structure and the implementation details, such choice of parametrization (Hills and Smith, 1992; Gelfand et al., 1995) or identifiability constraints (Vines et al., 1996; Gelfand and Sahu, 1999; Xie and Carlin, 2006). This work provides a contribution towards gaining a quantitative understanding of the interaction between Bayesian hierarchical structures and the behaviour of MCMC algorithms, which lies at the heart of the practical success of Bayesian statistics.

While there is some previous work in the area (Roberts and Sahu, 1997; Meng and Van Dyk, 1997; Papaspiliopoulos et al., 2003; Jones and Hobert, 2004;

[*] In the corrected version, we updated and corrected the references Bass and Sahu (2016a) and Bass and Sahu (2016b), which became Bass and Sahu (2017) and Bass and Sahu (2019) in the new version.

[†] Department of Decision Sciences, BIDSA and IGIER; Bocconi University, Milan, Italy, giacomo.zanella@unibocconi.it

[‡] Department of Statistics, University of Warwick, Coventry, UK, gareth.o.roberts@warwick.ac.uk

© 2021 International Society for Bayesian Analysis https://doi.org/10.1214/20-BA1242

Papaspiliopoulos et al. 2007; Yu and Meng, 2011), current theoretical understanding of the interaction between Bayesian hierarchical models and MCMC convergence is still very limited, and almost nothing is known for models of hierarchical depth greater than two. The present paper offers a contribution towards such an understanding, focusing on theory for Gaussian hierarchical models and seeking quantitative results. In particular, we derive analytic expressions for the convergence rates of the Gibbs Sampler for various multilevel linear models and explore the dependence of these rates on the model structure, the choice of parametrization and the introduction of identifiability constraints. The theoretical results given in this paper extend and improve substantially on existing literature (Roberts and Sahu, 1997; Yu and Meng, 2011; Bass and Sahu, 2017; Gao and Owen, 2017) both in terms of generality of hierarchical structure and the availability of explicit rates. We also show by simulations that the understanding gained from the Gaussian case can be extrapolated to more general settings.

In general, the Gibbs sampler can be elegantly described in terms of orthogonal projections (Amit, 1991, 1996; Diaconis et al., 2010). While in principle this theory provides the tools to extract practical convergence information for Gibbs samplers in the context of multivariate Gaussian distributions, in order to apply it to practically used Bayesian multilevel models one needs detailed knowledge of the spectrum of non-trivial high-dimensional matrices, which has drastically limited its applicability to derive analytic results. In this paper we combine this general framework with a novel multigrid decomposition approach that allows us to focus on low-dimensional Markov chains and derive explicit analytic results concerning Gibbs sampler rates of convergence for multilevel linear models, such as nested and crossed random effect models with an arbitrary number of layers and/or factors.

Our results have various practical implications. First they can be readily used in the popular context of conditionally Gaussian models, where there exist unknown variances at various levels of the hierarchy (Gelman and Hill, 2006). In that case our results describe, for example, the optimal updating strategies for the hierarchical mean structure conditional on the variances, allowing to optimize the mean parametrization on the fly (Section 3.2), or the computationally optimal way of imposing statistically identifiability (Sections 4.2), and provide theoretically grounded indication of which parameters to monitor in the convergence diagnostic process (Section 2.1). Also, our results can be used as a building block to derive computational complexity statements about the Gibbs Sampler in the context of multilevel linear model (see e.g. Papaspiliopoulos et al., 2019 for work in that direction). Note that in the context of conditionally Gaussian models the entire Gaussian mean component could be updated in a single block, thus avoiding convergence issues related to single-site updates. However these block updates can in principle be computationally expensive (up to $O(n^3)$ cost in the dimension (n) of the Gaussian to be updated), while single-site updating schemes with provably bounded convergence rate can offer a more scalable alternative. For some class of models, sparse linear algebra methods can reduce the cost of the block update by exploiting sparsity in the posterior precision matrix, but the resulting computational cost depends on the model structure and can still be super-linear (see e.g. Section 4 for models leading to dense precision matrices and Papaspiliopoulos et al., 2019 for more discussion).

While impressive results are being obtained with black-box software implementation of Hamiltonian Monte Carlo (HMC) such as STAN (Carpenter et al., 2017), our results suggest that Gibbs Sampling schemes built on our methodological guidance can be substantially cheaper than gradient-based ones in the context of hierarchical models, leading to improved performances (Section 5). Moreover, our simulations show that the methodological results we develop in this paper are also helpful when fitting multilevel models with gradient-based schemes (Section 5.1) and lead to drastic improvements in efficiency also when using generic software, such as STAN.

Throughout the paper, we shall couch all our results in terms of L^2 rates of convergence. Specifically, let $(\boldsymbol{\beta}(s))_{s=1,2,\dots}$ be a Markov chain with stationary distribution π and transition operator defined by $P^s f(\boldsymbol{\beta}(0)) = \mathbb{E}[f(\boldsymbol{\beta}(s))|\boldsymbol{\beta}(0)]$. The *rate of convergence* $\rho(\boldsymbol{\beta}(s))$ associated to $(\boldsymbol{\beta}(s))_{s=1,2,\dots}$ is defined as the smallest number ρ such that for all $r > \rho$

$$\lim_{s\to\infty} \frac{\|P^s f - \mathbb{E}_\pi[f]\|_{L^2(\pi)}}{r^s} = 0 \qquad \forall f \in L^2(\pi)\,, \tag{1.1}$$

where $L^2(\pi)$ denotes the space of square π-integrable functions, $\|\cdot\|_{L^2(\pi)}$ is its associated L^2-norm and $\mathbb{E}_\pi[f] = \int f\, d\pi$ is the expectation of f with respect to π. The rate of convergence $\rho(\boldsymbol{\beta}(s))$ characterizes the speed at which $(\boldsymbol{\beta}(s))_{s=1,2,\dots}$ converges to its stationary distribution π, with a simple argument giving that if

$$T = \min\{s;\ \|P^s f - \mathbb{E}_\pi[f]\|_{L^2(\pi)} \leq \epsilon\},$$

then $T = \mathcal{O}\left(\frac{1}{-\log(\rho)}\right)$.

1.1 Paper overview and structure

Section 2 carefully introduces the 3-level hierarchical models we shall consider, and provides motivating simulations. Then in Section 3 we shall give a complete analysis for 3-level symmetric models (i.e. homogeneous variances and symmetric data structure). At the heart of the analysis is a multigrid decomposition of the Gibbs sampler into completely independent Markov chains describing different levels of hierarchical granularity, Theorem 1. Such multigrid decomposition simultaneously applies to every Gibbs sampler induced by all centred/non-centred parametrizations and is fundamentally a statistical property of the hierarchical models under consideration. Although multigrid ideas have already been used in methodological contexts to design improved MCMC schemes (Goodman and Sokal, 1989; Liu and Sabatti, 2000), to our knowledge they had never been used in theoretical contexts to study convergence rates. We demonstrate that the slowest of these independent chains is always that corresponding to the coarsest level, regardless of the value of the variance components and on the number of branches in the hierarchy, and thus derive explicit expressions for the rates of convergence.

In Section 4 we focus on crossed effect models, using again a multigrid decomposition approach to derive explicit convergence rates. The results show that in the context of crossed models, centred/non-centred reparametrizations are not sufficient to guarantee fast convergence of the resulting Gibbs Sampler. On the other hand, we show that the latter can be achieved by imposing stronger statistical identifiability through additional linear constraints and our theory provides indications on which constraints lead to faster convergence. Finally, a simulation study reported in Section 5 suggests that the analysis of the Gaussian case leads to useful guidance also in the case of non-Gaussian models for both the Gibbs Sampler and Hamiltonian Monte Carlo algorithms (Neal et al., 2011).

Section 6 considers 3-level non-symmetric hierarchical models, providing bounds on convergence rates based on comparisons with related symmetric models and discussing the use of *bespoke parametrizations*, where the choice of centred or non-centred parametrization in each branch of the hierarchy depends on branch-specific parameters.

Section 7 considers hierarchical models with arbitrary depth (≥ 4). Using an appropriate auxiliary random walk, whose evolution through the hierarchical tree is governed by the parameters' squared partial correlations, we are able to extend the multigrid analysis to general tree structures and some non-symmetric cases. We again demonstrate a fundamental multigrid decomposition in Theorem 9, where the coarsest level chain converges the slowest, and give explicit formulae for convergence rates.

2 Three level hierarchical linear models

The theoretical innovation in this paper is centred around an important case in which we can obtain explicit Gibbs sampler rates of convergence, and as a result study explicitly the effects of particular models, parametrization schemes and blocking strategies. We begin with a detailed study of the following three-level Gaussian linear model, providing a fairly complete understanding of the interaction between model structure and parametrization and the Gibbs Sampler convergence behaviour.

Model S3 (Symmetric 3-levels hierarchical model)**.** *Suppose*

$$y_{ijk} = \mu + a_i + b_{ij} + \epsilon_{ijk}, \tag{2.1}$$

where i, j and k run from 1 to I, J and K respectively and ϵ_{ijk} are iid normal random variables with mean 0 and variance σ_e^2. We employ the standard Bayesian model specification assuming $a_i \sim N(0, \sigma_a^2)$, $b_{ij} \sim N(0, \sigma_b^2)$ and a flat prior on μ.

For the theoretical analysis, we will consider the variance terms σ_a^2, σ_b^2 and σ_e^2 to be known (in contrast with the simulations where we generalize to the case of unknown variances). Defining $\mathbf{a} = (a_i)_i$, $\mathbf{b} = (b_{ij})_{i,j}$ and $\mathbf{y} = (y_{ijk})_{i,j,k}$, the Gibbs Sampler explores the posterior distribution $(\mu, \mathbf{a}, \mathbf{b})|\mathbf{y}$ by iteratively sampling from the full conditional distributions of μ, $\mathbf{a}$ and $\mathbf{b}$ as follows (see below for motivation of denoting such sampler as GS$(1,1)$).

Sampler GS$(1,1)$. *Initialize $\mu(0)$, $\mathbf{a}(0)$ and $\mathbf{b}(0)$ and then iterate*

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

where $p(\mu|\mathbf{a}, \mathbf{b}, \mathbf{y})$, $p(a_i|\mu, \mathbf{b}, \mathbf{y})$ and $p(b_{ij}|\mu, \mathbf{a}, \mathbf{y})$ are the full conditionals of Model S3 (see Zanella and Roberts (2021) for explicit expressions).

Given the conditional independence structure of the model, Sampler GS$(1,1)$ is equivalent to a blocked Gibbs sampler with components μ, $\mathbf{a}$ and $\mathbf{b}$, i.e. a scheme performing consecutive updates of $\mu|\mathbf{a},\mathbf{b}$, $\mathbf{a}|\mu,\mathbf{b}$ and $\mathbf{b}|\mu,\mathbf{a}$ at each iteration.

The parametrization $(\mu,\mathbf{a},\mathbf{b})$ induced by (2.1) is often referred to as *non-centred parametrization* (NCP) and it is contrasted with the *centred parametrization* (CP) obtained by replacing a_i and b_{ij} with $\gamma_i=\mu+a_i$ and $\eta_{ij}=\gamma_i+b_{ij}$ respectively. Under the centred parametrization $(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$ the model formulation becomes

$$y_{ijk}\sim N(\eta_{ij},\sigma_e^2),\qquad \eta_{ij}\sim N(\gamma_i,\sigma_b^2),\qquad \gamma_i\sim N(\mu,\sigma_a^2),\qquad p(\mu)\propto 1\,. \tag{2.2}$$

Figures 1b and 1a provide a graphical representation of the two parametrizations. In the $(\mu,\mathbf{a},\mathbf{b})$ case $(1,1)$ refers to the fact that both levels 1 and 2 use a non-centred parametrization, while in the $(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$ case $(0,0)$ indicates that both levels use a centred parametrization. The resulting Gibbs sampler for the centred parametrization is as follows.

Sampler GS$(0,0)$. *Initialize $\mu(0)$, $\boldsymbol{\gamma}(0)$ and $\boldsymbol{\eta}(0)$ and then iterate*

1. *Sample $\mu(s+1)$ from $p(\mu|\boldsymbol{\gamma}(s),\boldsymbol{\eta}(s),\mathbf{y})$;*
2. *Sample $\gamma_i(s+1)$ from $p(\gamma_i|\mu(s+1),\boldsymbol{\eta}(s),\mathbf{y})$ for all i;*
3. *Sample $\eta_{ij}(s+1)$ from $p(\eta_{ij}|\mu(s+1),\boldsymbol{\gamma}(s+1),\mathbf{y})$ for all i and j,*

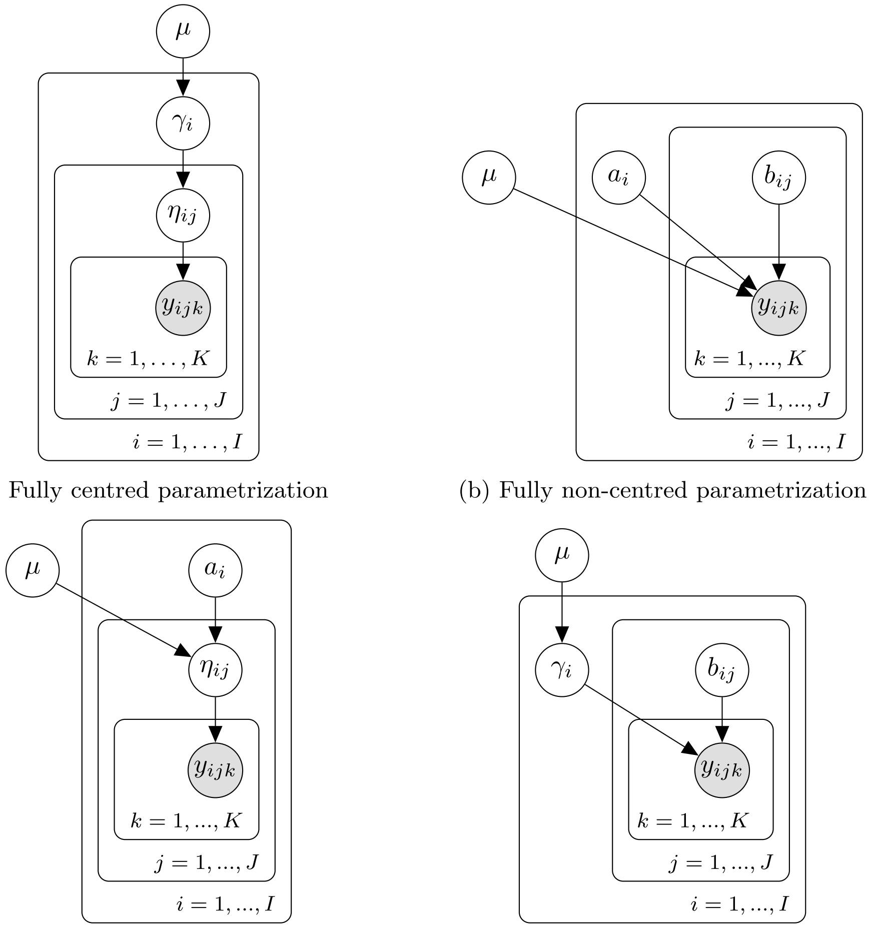
where $p(\mu|\boldsymbol{\gamma},\boldsymbol{\eta},\mathbf{y})$, $p(\gamma_i|\mu,\boldsymbol{\eta},\mathbf{y})$ and $p(\eta_{ij}|\mu,\boldsymbol{\gamma},\mathbf{y})$ are the full conditionals induced by (2.2) *(see supplementary material for explicit expressions).*

Together with the fully non-centred parametrization $(\mu,\mathbf{a},\mathbf{b})$ and the fully centred parametrization $(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$, one can also consider the mixed parametrizations given by $(\mu,\boldsymbol{\gamma},\mathbf{b})$ and $(\mu,\mathbf{a},\boldsymbol{\eta})$ and the corresponding Gibbs Sampler schemes $GS(0,1)$ and $GS(1,0)$. See Figures 1c and 1d for graphical representations.

2.1 Illustrative example

As an illustrative example, we simulated data from Model S3 with $I=J=100$, $K=5$, $\mu=0$, $\sigma_a=\sigma_e=10$ and $\sigma_b=10^{-0.5}$. This corresponds to a scenario of high level of noise in the measurements. We fit model S3 assuming the standard deviations $(\sigma_a,\sigma_b,\sigma_e)$ to be unknown and placing weakly informative priors, namely $\frac{1}{\sigma_a^2}$, $\frac{1}{\sigma_b^2}$ and $\frac{1}{\sigma_e^2}$ a priori distributed according to an Inverse Gamma distribution with shape and rate parameters equal to 0.01. We compare the efficiency of the Gibbs sampling schemes corresponding to the four different parametrizations, denoting them by $GS(1,1)$, $GS(0,0)$, $GS(0,1)$ and $GS(1,0)$, initializing the chains at true values of the parameters $(\mu,\mathbf{a},\mathbf{b})$ and $(\sigma_a,\sigma_b,\sigma_e)$. The more realistic case of starting the chains from randomly chosen states led to the same conclusions.

Rows 1-2 of Figure 2 plot the global mean μ and displays the potentially dramatic difference among mixing properties of the Gibbs Sampler under different parametrizations. Based on those, one would certainly exclude using $GS(1,1)$ and $GS(1,0)$ to fit this model, leaving $GS(0,0)$ and $GS(0,1)$ as possibly feasible algorithms. However, as an additional check, a cautious practitioner may also explore the mixing of the parame-



(a) Fully centred parametrization

(b) Fully non-centred parametrization

(c) Mixed parametrization: $(\mu, \mathbf{a}, \boldsymbol{\eta})$

(d) Mixed parametrization: $(\mu, \boldsymbol{\gamma}, \mathbf{b})$

Figure 1: Graphical representations of 3-levels hierarchical linear models under different parametrizations.

ters at the first level, namely $\mathbf{a}$ and $\boldsymbol{\gamma}$. Rows 3-4 of Figure 2 display the behaviour of the global averages of such parameters, namely $a. = \frac{\sum_i a_i}{I}$ and $\gamma. = \frac{\sum_i \gamma_i}{I}$, in the first 1000 iterations. Again, we see a dramatic difference induced by different parametrizations and, somehow surprisingly, despite having good mixing behaviour at level 0 (i.e. μ), $GS(0,0)$ displays very poor mixing behaviour at level 1 (i.e. $\boldsymbol{\gamma}$). It is then natural to explores also the mixing behaviour at level 2 and rows 5-6 of Figure 2 do so again by plotting the global averages $b.. = \frac{\sum_{ij} \beta_{ij}}{IJ}$ and $\eta.. = \frac{\sum_{ij} \eta_{ij}}{IJ}$. In this case $GS(1,1)$ and $GS(0,1)$ are the only chains mixing well. Based on Figure 2 it is natural to choose to fit the model using the sampler $GS(0,1)$ corresponding to the mixed parametrization $(\mu, \boldsymbol{\gamma}, \mathbf{b})$, as it is the only one mixing well at each of the three levels.

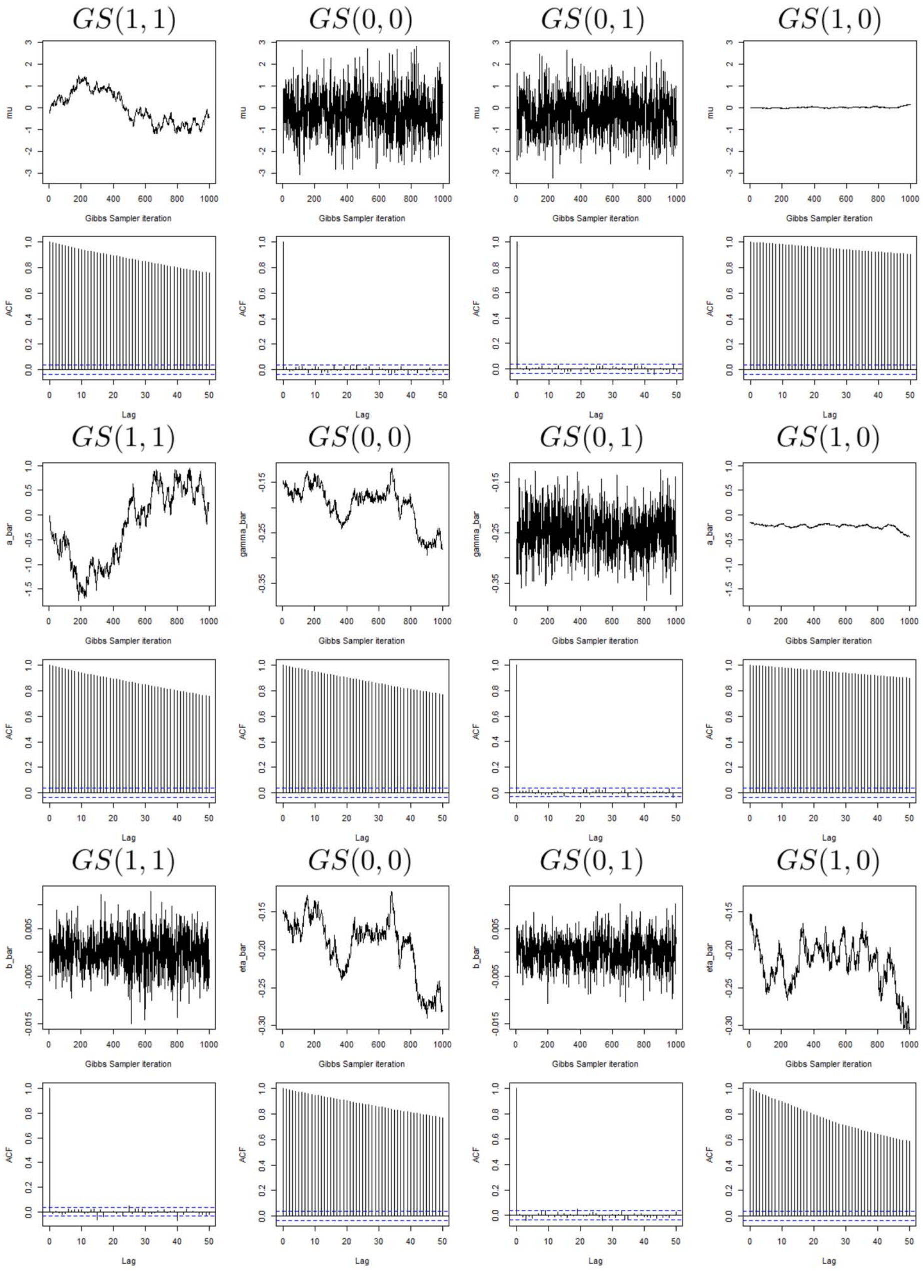


Figure 2: Mixing behaviour at level 0 (μ; rows 1-2), level 1 ($a.$ and $\gamma.$; rows 3-4) and level 2 ($b..$ and $\eta..$; rows 5-6); under four different parametrizations. For each level, the ranges of the y-axes are constant across parametrizations sharing the same parameters.

This simple example shows many typical issues arising when fitting Bayesian multilevel models and raises many questions. For example, one would like to know what are good parameters to use to diagnose convergence, in order to avoid misleading conclusions like the one suggested by rows 1-2 of Figure 2. In fact, while in two level model good mixing of the global hyperparameters such as μ typically indicates good global mixing, this is not true in other multi-level models. Indeed, it is legitimate to wonder whether diagnoses based only on the global means, as in Figure 2, are enough to deduce good mixing of the whole Markov chain, which in our example has more than 10^4 dimensions ($1 + I + IJ$ mean components and 3 precision components). Below we will show that for Model S3, mixing of the global means ensures mixing of the whole $(1 + I + IJ)$-dimensional mean components of the chain given the variances (see e.g. Corollary 1). Therefore it is enough to monitor the three global means and the three variances to ensure a reliable check of the chain mixing properties.

Even more crucially, it is desirable to have simple and theoretically grounded guidance in choosing a computationally efficient parametrization, given the huge impact it can have on computational performances. The theoretical analysis developed in the next section will provide useful guidance in this respect.

3 Multigrid decomposition for the three level hierarchical model

The basic ingredient of our analysis is the following multigrid decomposition. Consider the four possible parametrization of Model S3: $(\mu, \mathbf{a}, \mathbf{b})$, $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ and the mixed parametrizations $(\mu, \boldsymbol{\gamma}, \mathbf{b})$ and $(\mu, \mathbf{a}, \boldsymbol{\eta})$. In order to provide a unified treatment, regardless of the chosen parametrization, we denote the parameters used by $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ and the resulting Gibbs Sampler by $GS(\boldsymbol{\beta})$. For example, in the NCP case $\boldsymbol{\beta}^{(0)} = \mu$, $\boldsymbol{\beta}^{(1)} = \mathbf{a}$, $\boldsymbol{\beta}^{(2)} = \mathbf{b}$ and $GS(\boldsymbol{\beta})$ coincides with GS$(1, 1)$. First consider the map δ sending $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ to

$$\delta(\boldsymbol{\beta}) = \begin{pmatrix} \delta^{(0)}\boldsymbol{\beta} \\ \delta^{(1)}\boldsymbol{\beta} \\ \delta^{(2)}\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \delta^{(0)}\boldsymbol{\beta}^{(0)} \, , \, \delta^{(0)}\boldsymbol{\beta}^{(1)} \, , \, \delta^{(0)}\boldsymbol{\beta}^{(2)} \\ \delta^{(1)}\boldsymbol{\beta}^{(1)} \, , \, \delta^{(1)}\boldsymbol{\beta}^{(2)} \\ \delta^{(2)}\boldsymbol{\beta}^{(2)} \end{pmatrix} , \tag{3.1}$$

where, loosely speaking, $\delta^{(i)}\boldsymbol{\beta}$ represent the increments of $\boldsymbol{\beta}$ at the i-th coarseness level. More precisely

$$\delta^{(0)}\beta^{(0)} = \beta^{(0)} \, , \qquad \delta^{(0)}\boldsymbol{\beta}^{(1)} = \beta_{\cdot}^{(1)} \, , \qquad \delta^{(0)}\boldsymbol{\beta}^{(2)} = \beta_{\cdot\cdot}^{(2)} \, ,$$
$$\delta^{(1)}\boldsymbol{\beta}^{(1)} = \left(\beta_1^{(1)} - \beta_{\cdot}^{(1)}, \dots, \beta_I^{(1)} - \beta_{\cdot}^{(1)}\right), \, \delta^{(1)}\boldsymbol{\beta}^{(2)} = \left(\beta_{1\cdot}^{(2)} - \beta_{\cdot\cdot}^{(2)}, \dots, \beta_{I\cdot}^{(2)} - \beta_{\cdot\cdot}^{(2)}\right) ,$$
$$\delta^{(2)}\boldsymbol{\beta}^{(2)} = \left(\beta_{11}^{(2)} - \beta_{1\cdot}^{(2)}, \beta_{12}^{(2)} - \beta_{1\cdot}^{(2)}, \dots, \beta_{I(J-1)}^{(2)} - \beta_{I\cdot}^{(2)}, \beta_{IJ}^{(2)} - \beta_{I\cdot}^{(2)}\right) ,$$

where

$$\beta_{\cdot}^{(1)} = \frac{\sum_i \beta_i^{(1)}}{I} \, , \qquad \beta_{\cdot\cdot}^{(2)} = \frac{\sum_{i,j} \beta_{ij}^{(2)}}{IJ} \, , \qquad \beta_{i\cdot}^{(2)} = \frac{\sum_j \beta_{ij}^{(2)}}{J} \, .$$

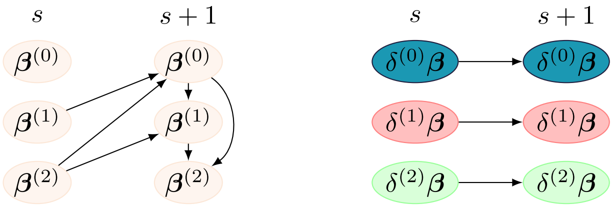

s
$s+1$
$\boldsymbol{\beta}^{(0)}$
$\boldsymbol{\beta}^{(1)}$
$\boldsymbol{\beta}^{(2)}$
s
$s+1$
$\delta^{(0)}\boldsymbol{\beta}$
$\delta^{(1)}\boldsymbol{\beta}$
$\delta^{(2)}\boldsymbol{\beta}$

Figure 3: Illustration of Theorem 1. Left: the transition from $\boldsymbol{\beta}(s)$ to $\boldsymbol{\beta}(s+1)$ in Sampler $GS(\boldsymbol{\beta})$ follows the structure of a Gibbs Sampler with 3 components. Right: the transition from $\delta\boldsymbol{\beta}(s)$ to $\delta\boldsymbol{\beta}(s+1)$ in Sampler $GS(\boldsymbol{\beta})$ follows the structure of three independent Markov chains.

It is easy to see that the map δ is a bijection between $\mathbb{R}^d$ and $\mathbb{R}^3 \times (\mathbb{R}^I)^* \times (\mathbb{R}^I)^* \times_{i=1}^I (\mathbb{R}^J)^*$, where $(\mathbb{R}^p)^* = \{(v_1, \dots, v_p) \in \mathbb{R}^p : \sum_{i=1}^p v_i = 0\}$. The dimensionality of $\delta\boldsymbol{\beta}$ equals that of $\boldsymbol{\beta}$, which is $1 + I + IJ$, because $\delta\boldsymbol{\beta}$ has $3 + 2I + IJ$ parameters and $2 + I$ constraints. The following theorem shows that the Markov chain induced by $GS(\boldsymbol{\beta})$ factorizes under the transformation δ, as illustrated in Figure 3.

Theorem 1 (Multigrid decomposition)**.** *Let $(\boldsymbol{\beta}(s))_{s=1}^\infty$ be a Markov chain on $\mathbb{R}^d$ evolving according to $GS(\boldsymbol{\beta})$. Then the functionals $(\delta^{(0)}\boldsymbol{\beta}(s))_{s=1}^\infty$, $(\delta^{(1)}\boldsymbol{\beta}(s))_{s=1}^\infty$ and $(\delta^{(2)}\boldsymbol{\beta}(s))_{s=1}^\infty$ evolve as three independent Markov chains.*

While the posterior independence of $\delta^{(0)}\boldsymbol{\beta}$, $\delta^{(1)}\boldsymbol{\beta}$ and $\delta^{(2)}\boldsymbol{\beta}$ is well-known, Theorem 1 shows the much stronger statement that the Markov chains induced by the Gibbs Sampler are independent.

Remark 1. *The three subspaces of $\mathbb{R}^d$ spanned by the vectors $\delta^{(0)}\boldsymbol{\beta}$, $\delta^{(1)}\boldsymbol{\beta}$ and $\delta^{(2)}\boldsymbol{\beta}$, respectively, do not depend on the choice of parametrization $\boldsymbol{\beta}$. For example, the four instances of $\delta^{(0)}\boldsymbol{\beta}$ – i.e. $(\mu, a_{\cdot}, b_{\cdot\cdot})$, $(\mu, a_{\cdot}, \eta_{\cdot\cdot})$, $(\mu, \gamma_{\cdot}, b_{\cdot\cdot})$ and $(\mu, \gamma_{\cdot}, \eta_{\cdot\cdot})$ – span the same subset of the parametric space of Model S3. In this sense, the multigrid decomposition of Theorem 1 factorizes the Gibbs Sampler into three independent chains operating on subspaces that depend only on the model under consideration and not on the particular parametrization being considered. Thus in a sense, the multigrid decomposition is intrinsic to the model, and not specific to the chosen parametrization.*

Theorem 1 provides a useful tool to analyse the Markov chain of interest, $\boldsymbol{\beta}(s)$. In fact the factorization into independent Markov chains implies that the rate of convergence of $\boldsymbol{\beta}(s)$ is simply given by the slowest rate of convergence among $\delta^{(0)}\boldsymbol{\beta}(s)$, $\delta^{(1)}\boldsymbol{\beta}(s)$ and $\delta^{(2)}\boldsymbol{\beta}(s)$. Interestingly, the slowest chain is always the chain at the highest level $\delta^{(0)}\boldsymbol{\beta}(s)$, regardless of the choice of parametrization and the values of $(I, J, K, \sigma_a, \sigma_b, \sigma_e)$.

Theorem 2 (Hierarchical ordering of convergence rates)**.** *Let $\delta^{(0)}\boldsymbol{\beta}(s)$, $\delta^{(1)}\boldsymbol{\beta}(s)$ and $\delta^{(2)}\boldsymbol{\beta}(s)$ be the Markov chains defined in Theorem 1. Then the associated convergence rates satisfy*

$$\rho(\delta^{(0)}\boldsymbol{\beta}(s)) \geq \rho(\delta^{(1)}\boldsymbol{\beta}(s)) \geq \rho(\delta^{(2)}\boldsymbol{\beta}(s)) = 0\,.$$

Theorems 1 and 2 imply that the rate of convergence of the global chain $\boldsymbol{\beta}(s)$ coincides with the one of the sub-chain $\delta^{(0)}\boldsymbol{\beta}(s)$ sampling the global means $(\beta^{(0)}, \beta^{(1)}_{\cdot}, \beta^{(2)}_{\cdot\cdot})$.

Corollary 1 (Rate of convergence of $GS(\boldsymbol{\beta})$)**.** *Given the notation of Theorem 1,*

$$\rho(\boldsymbol{\beta}(s)) = \rho(\delta^{(0)}\boldsymbol{\beta}(s))\,.$$

3.1 Explicit rates of convergence under different parametrizations

The multigrid decomposition developed in Section 3 permits a direct analysis on the convergence properties of $\boldsymbol{\beta}(s)$. Since this is a Gibbs Sampler targeting a multivariate Gaussian distributions, in principle it could be analysed using the tools developed in Amit (1996); Roberts and Sahu (1997); Khare et al. (2009). Applying these results requires a spectral decomposition of a matrix derived from Σ. However, given the high-dimensionality of $\boldsymbol{\beta}(s)$, which has $1+I+IJ$ parameters, it is hard to apply directly such results and in fact the convergence properties of $\boldsymbol{\beta}(s)$ have been studied heuristically or numerically in the literature (see e.g. Gelfand et al., 1995, Section 4 and Roberts and Sahu, 1997, Section 4.2). Circumventing these theoretical difficulties, Corollary 1 implies that it suffices to study the skeleton chain $\delta^{(0)}\boldsymbol{\beta}(s)$, which is a low-dimensional chain (namely 3-dimensional) amenable to direct analysis. Therefore we can derive analytic expressions for the rates of convergence for the Gibbs Sampler under different parametrizations.

Theorem 3. *Given an instance of Model S3, the rate of convergence of the four Gibbs Sampler schemes $GS(0,0)$, $GS(1,1)$, $GS(0,1)$ and $GS(1,0)$ are given by*

$$\rho_{00} = 1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_b^2}\,\frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2}\,, \qquad \rho_{10} = \max\left\{\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_b^2}, \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2}\right\}\,,$$

$$\rho_{01} = 1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_e^2}\,\frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2}\,, \qquad \rho_{11} = \max\left\{\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_e^2}, \frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2}\right\}\,,$$

where $\tilde{\sigma}_a^2 = \frac{\sigma_a^2}{I}$, $\tilde{\sigma}_b^2 = \frac{\sigma_b^2}{IJ}$ *and* $\tilde{\sigma}_e^2 = \frac{\sigma_e^2}{IJK}$.

Theorem 3 provides explicit and informative formulas regarding the interaction between choice of parametrization and resulting efficiency of the Gibbs Sampler for Model S3. Figure 4 summarizes graphically the dependence of the converge rates of different parametrizations from the values of the variances of various levels. Roughly speaking, the figure suggests that there is a partition of the hyperparameter space (corresponding to the white regions in each plot) such that in each region exactly one of the four parametrizations performs well.

Consider for example the illustrative example of Section 2.1. Applying Theorem 3 to such context we obtain that the L^2 rates of convergence (up to the third decimal digit) of the various Gibbs Samplers under consideration given $(I, J, K, \sigma_a, \sigma_b, \sigma_e) = (100, 100, 5, 10, 10^{-0.5}, 10)$ are

$$(\rho_{00}, \rho_{11}, \rho_{01}, \rho_{10}) = (0.995, 0.998, 0.007, 0.999)\,.$$

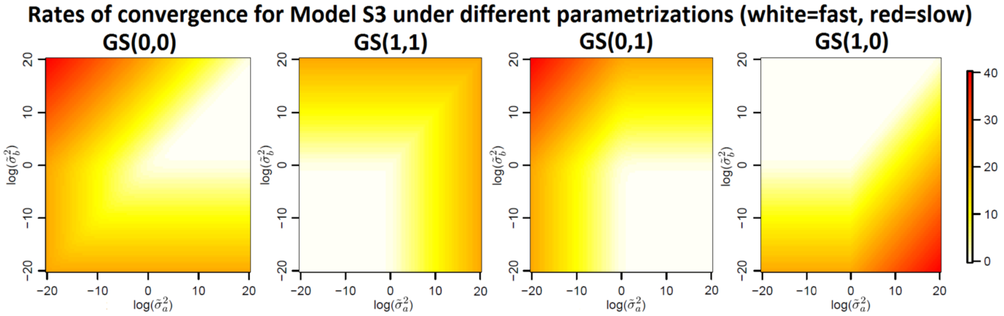


Figure 4: Plot of rates of convergence for three-levels Gaussian hierarchical models under different parametrizations. Color levels correspond to values of $\log(1-\rho)$, where ρ is the rate of convergence, as a function of $\log(\tilde{\sigma}^2_a)$ and $\log(\tilde{\sigma}^2_b)$ for fixed $\log(\tilde{\sigma}^2_e)=0$.

Recall that values of ρ close to 1 mean slow convergence, see (1.1) and discussion thereof. These numbers provide a quantitative and theoretically grounded description of the behaviour heuristically observed in Section 2.1 and can be easily used to optimize performance (see e.g. Section 3.2 below).

3.2 Conditionally optimal parametrization

We now consider the optimal parametrization (among the four possible choices $(\mu,\mathbf{a},\mathbf{b})$, $(\mu,\boldsymbol{\gamma},\mathbf{b})$, $(\mu,\mathbf{a},\boldsymbol{\eta})$ and $(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$) as a function of the normalized variance components $(\tilde{\sigma}^2_a,\tilde{\sigma}^2_b,\tilde{\sigma}^2_e)$. Using the formulae of Theorem 3 we obtain the following explicit answers.

Corollary 2 (Optimal parametrization for Model S3)**.** *The rate of convergence of the Gibbs Sampler targeting Model S3 is minimized by the following parametrization choice:*

- *use a centred parametrization $\boldsymbol{\eta}$ at the lowest level if and only if $\tilde{\sigma}^2_b \geq \tilde{\sigma}^2_e$,*
- *use a centred parametrization $\boldsymbol{\gamma}$ at the middle level if and only if $\tilde{\sigma}^2_a \geq \tilde{\sigma}^2_b + \tilde{\sigma}^2_e$.*

The resulting Gibbs Sampler has a rate of convergence ρ bounded above by $\frac{2}{3}$, with the equality $\rho = \frac{2}{3}$ holding if and only if $\tilde{\sigma}^2_a = \tilde{\sigma}^2_b + \tilde{\sigma}^2_e$ and $\tilde{\sigma}^2_b = \tilde{\sigma}^2_e$ (in which case all parametrizations are equivalent).

Table 1 provides a graphical representation of the decision process. This simple rule guarantees that the resulting Gibbs Sampler has a rate of converges smaller than $\frac{2}{3}$, thus ensuring high sampling efficiency for fixed variances. Table 1 implies that the choice of parametrization of a given level (i.e. whether it is computationally most efficient

	$\tilde{\sigma}^2_a \geq \tilde{\sigma}^2_b + \tilde{\sigma}^2_e$	$\tilde{\sigma}^2_a < \tilde{\sigma}^2_b + \tilde{\sigma}^2_e$
$\tilde{\sigma}^2_b \geq \tilde{\sigma}^2_e$	$(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$	$(\mu,\mathbf{a},\boldsymbol{\eta})$
$\tilde{\sigma}^2_b < \tilde{\sigma}^2_e$	$(\mu,\boldsymbol{\gamma},\mathbf{b})$	$(\mu,\mathbf{a},\mathbf{b})$

Table 1: Optimal parametrization as a function of the normalized variance components.

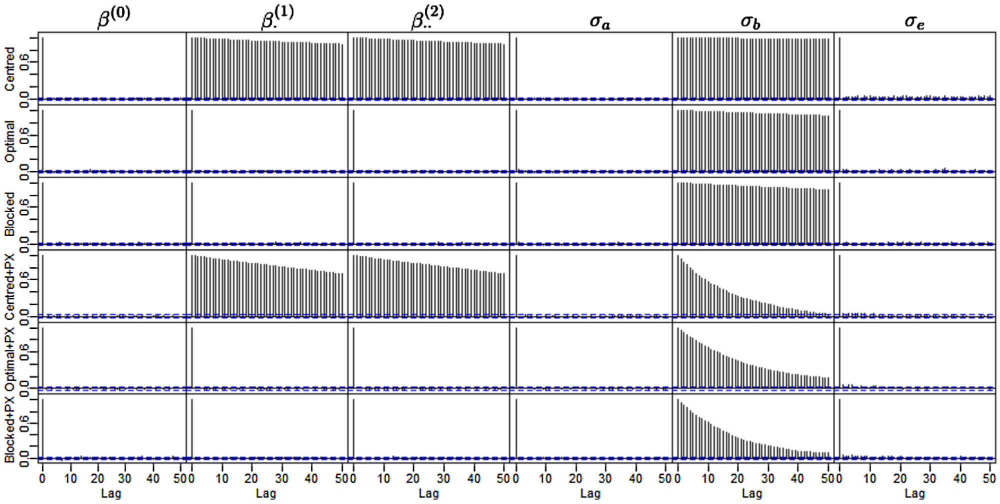


Figure 5: Autocorrelation functions for the global means $(\beta^{(0)}, \beta_{\cdot}^{(1)}, \beta_{\cdot\cdot}^{(2)})$ and the standard deviations $(\sigma_a, \sigma_b, \sigma_e)$, under the three updating schemes (Centred, Optimal and Blocked) described in Section 3.2. The combination with the parameter expansion methodology is denoted as "+PX".

to use a centred or non-centred parametrization) depends on the ratio between the normalized variance at the level under consideration and the sum of the normalized variances of the levels below. These results extend previous intuition for the two-level case (Papaspiliopoulos et al., 2003) to hierarchical models with three levels.

Corollary 2 allows for simple and effective strategies to guarantee high sampling efficiency in practical implementations of Gibbs Sampling for Model S3 in the case of unknown variances. Common implementations choose a fixed parametrization $\boldsymbol{\beta}$ of the Gaussian component, such as the fully centred parametrization $\boldsymbol{\beta} = (\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$, and alternate updating $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ with $GS(\boldsymbol{\beta})$ and $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$ with direct sampling (which is straightforward using the conditional independence of σ_a, σ_b and σ_e given $\boldsymbol{\beta}$). Given Corollary 2, instead, one can at each iteration choose the optimal parametrization $\boldsymbol{\beta}$ given $(\sigma_a, \sigma_b, \sigma_e)$ according to Table 1, with basically no additional computational cost compared to the cost of a Gibbs Sampling iteration. This ensures that the sampling step $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ will have a high efficiency, regardless of the values of $(I, J, K, \sigma_a, \sigma_b, \sigma_e)$.

We implement and illustrate this strategy in Figure 5, where we compare MCMC autocorrelation functions in the context of the illustrative example of Section 2.1, with unknown variances $(\sigma_a, \sigma_b, \sigma_e)$. We compare the following three schemes: the sampler updating $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})|(\sigma_a, \sigma_b, \sigma_e)$ with $GS(0, 0)$ and $(\sigma_a, \sigma_b, \sigma_e)|(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ exactly; the sampler choosing the optimal parametrization $\boldsymbol{\beta}$ according to Table 1 and then updating $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ with $GS(\boldsymbol{\beta})$ and $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$ exactly; the sampler updating both $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ and $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$ exactly (which can be implemented because the distribution of $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ is multivariate Gaussian). We call the three samplers "Centred", "Optimal" and "Blocked", respectively. The results are displayed in the first three rows of Figure 5 and show that the Optimal sampler reduces significantly the autocorrelation compared to Centred one, and achieves a mixing that is basically equivalent to the one of the Blocked sampler. The potential benefit of the Optimal sampler compared to the Blocked one is that the Gibbs update of $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ in the Optimal sampler only requires univariate updates and has a potentially lower computational cost compared to a full multivariate block update of $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$, which requires large matrix operations. While these matrix operations can be performed efficiently in the context of nested linear models (see e.g. Papaspiliopoulos and Zanella, 2017), their cost becomes significantly larger for example in the context of crossed random effect models (see Section 4 below and Papaspiliopoulos et al., 2019). Note that the similarity of performances between the Optimal and Blocked sampler is not surprising given our theoretical results above. In fact Corollary 2 guarantees that the sampler $GS(\boldsymbol{\beta})$ used in the Gibbs update have a rate of convergence upper bounded by $2/3$, which is well separated from 1. When such updates are nested within a larger sampler (e.g. the one updating $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ and $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$) the difference between and exact update of $\boldsymbol{\beta}$ and a Gibbs one with good rate of convergence can easily become negligible.

We then combined the three schemes described above with the parameter expansion (PX) methodology of Meng and Van Dyk (1999); Liu and Wu (1999), and denote the resulting schemes as Centred+PX, Optimal+PX and Blocked+PX. The PX methodology aims to avoid potential slow mixing due to strong dependencies between $\boldsymbol{\beta}$ and $(\sigma_a, \sigma_b, \sigma_e)$, and in this case it is successful in doing so for the slowly mixing parameter σ_b (see Figure 5, rows 4-6). The results suggest that the optimal choice of parametrization for $\boldsymbol{\beta}$ can be conveniently combined with the PX methodology, and that the two have complementary roles in speeding up the convergence of the Gibbs Sampler.

4 Multigrid decomposition for crossed effect models

The multigrid decomposition can be used to analyse non-nested models. In this section we focus on the following crossed effect model.

Model Ck (k-factors crossed-effects model)**.**

$$y_{i_1\dots i_k} =\mu + a^{(1)}_{i_1} + \dots + a^{(k)}_{i_k} + \epsilon_{i_1\dots i_k} \qquad i_s = 1, \dots, n_s, \quad s = 1, \dots, k\,, \tag{4.1}$$

with $a^{(s)}_{i_s} \overset{iid}{\sim} N(0, 1/\tau_s)$ *for* $s \in \{1, \dots, k\}$, $\epsilon_{i_1\dots i_k} \overset{iid}{\sim} N(0, 1/\tau_e)$ *and* $p(\mu) \propto 1$. *We denote the number of observed datapoints by* $N = \prod_{s=1}^{k} n_s$.

Similarly to Sections 2 and 3, we use bold letters to denote the following vectors: $\boldsymbol{y} = (y_{i_1\dots i_k})_{i_1,\dots,i_k}$, $\boldsymbol{a}^{(s)} = (a^{(s)}_{i_s})_{i_s}$, $\boldsymbol{a} = (\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)})$ and $\boldsymbol{a}^{(-s)} = (\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(s-1)}, \boldsymbol{a}^{(s+1)}, \dots, \boldsymbol{a}^{(k)})$. The standard Gibbs Sampler to sample from the posterior distribution $\mathcal{L}(\mu, \boldsymbol{a}|\boldsymbol{y})$ of Model Ck is defined as follows.

Sampler GS-crossed. *At each iteration*

1. *sample* μ *from* $\mathcal{L}(\mu|\boldsymbol{a}, \boldsymbol{y})$,
2. *sample* $\boldsymbol{a}^{(s)}$ *from* $\mathcal{L}\left(\boldsymbol{a}^{(s)}|\mu, \boldsymbol{a}^{(-s)}, \boldsymbol{y}\right)$ *with* s *going from* 1 *to* k.

Model Ck and Sampler GS-crossed have recently been analysed in Papaspiliopoulos et al. (2019) using the multigrid decomposition approach developed in Section 3 of this paper to derive expressions for the convergence rate of Sampler GS-crossed. In particular, Papaspiliopoulos et al. (2019) considered the following linear functions of $\boldsymbol{a}$

$$\bar{a}^{(s)} = \frac{1}{n_s}\sum_{i=1}^{n_s} a_i^{(s)} \quad \text{and} \quad \delta\boldsymbol{a}^{(s)} = (\boldsymbol{a}^{(s)} - \bar{a}^{(s)})\,, \tag{4.2}$$

for each $s \in \{1, \dots, k\}$ and proved the following result.

Theorem 4 (Papaspiliopoulos et al. (2019))**.** *Let*

$$((\mu, \boldsymbol{a})(t))_{t=1}^{\infty} = \left(\mu(t), \boldsymbol{a}^{(1)}(t), \dots, \boldsymbol{a}^{(k)}(t)\right)_{t=1}^{\infty}$$

be the Markov chain generated by Sampler GS-crossed. Then the time-wise transformations $\left((\mu, \bar{a}^{(1)}, \dots, \bar{a}^{(k)})(t)\right)_{t=1}^{\infty}$ *and* $\left(\delta\boldsymbol{a}^{(1)}(t)\right)_{t=1}^{\infty}, \dots, \left(\delta\boldsymbol{a}^{(k)}(t)\right)_{t=1}^{\infty}$ *are* $(k+1)$ *independent Markov chains. Moreover, the rate of convergence of* $((\mu, \boldsymbol{a})(t))_{t=1}^{\infty}$ *is*

$$\rho = \max_{s\in\{1,\dots,k\}} \frac{N\tau_e}{N\tau_e + n_s\tau_s}\,. \tag{4.3}$$

Theorem 4 implies that the convergence properties of Sampler GS-crossed deteriorate as N increases because $\max_{s\in\{1,\dots,k\}}(N\tau_e)^{-1}(N\tau_e + n_s\tau_s)$ goes to 1 as $N \to \infty$. Motivated by this consideration, Papaspiliopoulos et al. (2019) propose a collapsed Gibbs Sampler that avoids such slowdown for increasing data size while preserving the same computational cost per iteration of Sampler GS-crossed. In the following two sections we extend the analysis of Model Ck performed in Papaspiliopoulos et al. (2019), focusing on the role of, respectively, reparametrizations and statistical identifiability.

4.1 Reparametrizations and crossed effects models

In the context of nested models, reparametrization techniques based on hierarchical centering offer a way to make the Gibbs Sampler robust to large datasets (see e.g. Corollary 2). We now show that this is not the case in the crossed effects context of Model Ck. In this section we focus on the case $k = 2$, which is a case often studied theoretically in the literature (see e.g. Gao and Owen (2017); Brown et al. (2018) for recent examples). In this case, hierarchical centering leads to four possible parametrizations defined as

$$(\mu, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}) = (\mu, \boldsymbol{a}^{(1)} + (1-\lambda_1)\mu, \boldsymbol{a}^{(2)} + (1-\lambda_2)\mu)\,, \quad \text{for } (\lambda_1, \lambda_2) \in \{0,1\}^2\,. \tag{4.4}$$

Each parametrization corresponds to a different Gibbs Sampler, which at each iteration updates μ from $\mathcal{L}(\mu|\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{y})$, $\boldsymbol{\beta}^{(1)}$ from $\mathcal{L}\big(\boldsymbol{\beta}^{(1)}|\mu, \boldsymbol{\beta}^{(2)}, \boldsymbol{y}\big)$, and $\boldsymbol{\beta}^{(2)}$ from $\mathcal{L}\big(\boldsymbol{\beta}^{(2)}|\mu, \boldsymbol{\beta}^{(1)}, \boldsymbol{y}\big)$. The following result characterizes the rate of convergence $\rho_{\lambda_1\lambda_2}$ of such Gibbs Samplers for all combinations $(\lambda_1, \lambda_2) \in \{0,1\}^2$.

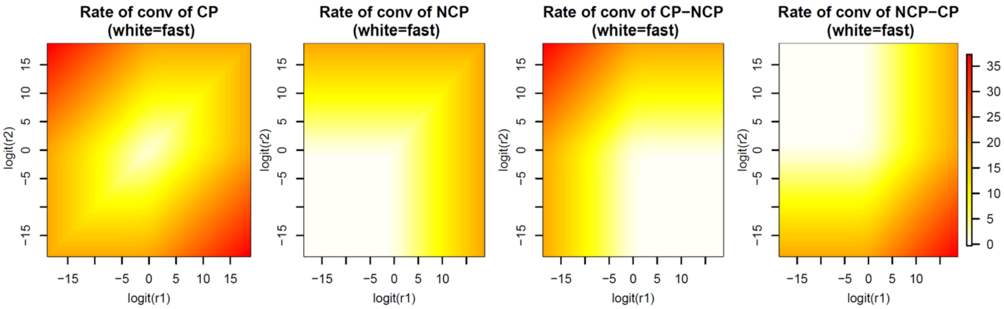


Figure 6: Plot of the rates of convergence in (4.5). Color levels correspond to values of $\log(1-\rho)$, where ρ is the rate of convergence (or its lower bound for ρ_{11}), as a function of $\log(r_1/(1-r_1))$ and $\log(r_2/(1-r_2))$.

Theorem 5. *Let* $r_1 = \frac{N\tau_e}{N\tau_e + n_1\tau_1}$ *and* $r_2 = \frac{N\tau_e}{N\tau_e + n_2\tau_2}$. *Then we have*

$$\begin{aligned} \rho_{11} &= \max\{r_1, r_2\}\,, & \rho_{10} &= 1 - r_2(1-r_1)\,, \\ \rho_{01} &= 1 - r_1(1-r_2)\,, & \rho_{00} &\geq 1 + r_1 r_2 - \min\{r_1, r_2\}\,. \end{aligned} \tag{4.5}$$

Figure 6 summarizes graphically the results of Theorem 5, showing the dependence of the converge rates on the choice of parametrization. The rate displayed in Figure 6 for the fully centred parametrization is the lower bound given in (4.5).

Theorem 5 implies that centering both factors (i.e. setting $\lambda_1 = \lambda_2 = 0$) is always computationally worse than any of the other parametrizations because $\rho_{00} \geq \max\{\rho_{11}, \rho_{01}, \rho_{10}\}$. On the other hand, the optimal choice of (λ_1, λ_2) among $(1,1)$, $(0,1)$ and $(1,0)$ depends on the specific values of r_1 and r_2. More precisely, the expressions in (4.5) imply that the convergence rate is minimized by centering the first factor (i.e. setting $\lambda_1 = 0$) if and only if $r_1 \geq (2-r_2)^{-1}$ and centering the second factor (i.e. setting $\lambda_2 = 0$) if and only if $r_2 \geq (2-r_1)^{-1}$. These results are in agreement with, for example, the empirical results obtained in Gelfand et al. (1996, Section 6) and Browne (2004).

Theorem 5 also implies that $\min\{\rho_{00}, \rho_{01}, \rho_{10}, \rho_{11}\} \to 1$ as $n_1, n_2 \to \infty$. Therefore, regardless of the parametrizations chosen, the convergence of Gibbs Samplers targeting Model Ck deteriorates as the number of factors n_1 and n_2 increases. This is in contrast with the nested case analysed in Section 3, where reparametrization techniques are successful in providing samplers with good convergence properties for all choices of hyperparameter values. In the next section we show that a more effective way to achieve good convergence properties is to impose stronger identifiability constraints.

4.2 Connections to statistical identifiability

The parameters $(\mu, \boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)})$ in Model Ck are not identifiable, in the sense that the mapping $(\mu, \boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)}) \to \mathcal{L}(\boldsymbol{y}|\mu, \boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)})$ is not injective. While this is not strictly speaking an issue for Bayesian inferences, one may wonder whether imposing identifiability on model parameters results in avoiding the degradation of mixing described in previous sections (see e.g. Vines et al., 1996; Gelfand and Sahu, 1999; Xie and Carlin, 2006; Kaufman et al., 2010; Vallejos et al., 2015 for related discussion and some examples in applications). We consider imposing identifiability by conditioning on some linear constraints, such as the commonly used choices of $a_1^{(s)} = 0$ or $\bar{a}^{(s)} = 0$. More generally, one can obtain identifiability for Model Ck by imposing a linear constraint $c_s = 0$ for each s from 1 to k, where $c_s = \sum_{j=1}^{n_s} w_j^{(s)} a_j^{(s)}$ is a linear combination of $(a_1^{(s)}, \dots, a_{n_s}^{(s)})$ weighted by some non-negative terms $(w_1^{(s)}, \dots, w_{n_s}^{(s)})$ satisfying $\sum_{j=1}^{n_s} w_j^{(s)} > 0$. Interestingly, one can exploit the multigrid decomposition (with minor modifications to adapt to the linear constraints, see Lemma 5.1 in the supplement for details) to derive the convergence rates of the resulting Gibbs Samplers for all choices of weights $(w_1^{(s)}, \dots, w_{n_s}^{(s)})$.

Theorem 6. *Consider Sampler GS-crossed conditioned on $c_s = 0$ for $s = 1, \dots, k$, meaning that μ gets updated from $\mathcal{L}(\mu | \boldsymbol{a}, \boldsymbol{y}, c_1 = \dots = c_k = 0)$ rather than $\mathcal{L}(\mu | \boldsymbol{a}, \boldsymbol{y})$, and $\boldsymbol{a}^{(s)}$ from $\mathcal{L}\left(\boldsymbol{a}^{(s)} | \mu, \boldsymbol{a}^{(-s)}, \boldsymbol{y}, c_1 = \dots = c_k = 0\right)$ rather than $\mathcal{L}\left(\boldsymbol{a}^{(s)} | \mu, \boldsymbol{a}^{(-s)}, \boldsymbol{y}\right)$. The rate of convergence of the resulting chain is*

$$\rho = \max_{s \in \{1, \dots, k\}} \left(\frac{N\tau_e (1 - q_s)}{N\tau_e + n_s \tau_s} \right) , \tag{4.6}$$

where $q_s = (\sum_{j=1}^{n_s} w_j^{(s)})^2 / (n_s \sum_{j=1}^{n_s} (w_j^{(s)})^2)$.

Comparing (4.6) with (4.3) we can see that, since $(1 - q_s) \in [0, 1)$, the rate of convergence always decreases after imposing the identifiability constraints $c_s = 0$ for $s = 1, \dots, k$. Thus, Theorem 6 implies that, in this context, imposing identifiability always improves the convergence properties of the Gibbs Sampler. To our knowledge, this is the first rigorous result of this kind in the Bayesian computation literature. On the other hand, the result also shows that imposing identifiability per se does not guarantee fast convergence. For example, Theorem 6 implies that the rate of convergence of Sampler GS-crossed conditioned on $a_1^{(s)} = 0$ for each $s \in \{1, \dots, k\}$ is given by

$$\rho = \max_{s \in \{1, \dots, k\}} \left(\frac{N\tau_e}{N\tau_e + n_s \tau_s} \frac{n_s - 1}{n_s} \right) ,$$

while the rate of convergence of Sampler GS-crossed conditioned on $\bar{a}^{(s)} = 0$ for each $s \in \{1, \dots, k\}$ equals 0, i.e. the sampler produces i.i.d. draws from the posterior distribution $\mathcal{L}(\mu, \boldsymbol{a} | \boldsymbol{y}, \bar{a}^{(1)} = \dots = \bar{a}^{(k)} = 0)$. While in both cases we observe an improvement over the original Gibbs Sampler in terms of convergence rates, the result shows that conditioning on $a_1^{(s)} = 0$ for each $s \in \{1, \dots, k\}$ leads to a convergence rate that can still go to 1 as the datasize increase. Interestingly (4.6) implies that the rate of convergence is minimized when q_s is maximized, which happens when the weights in the linear constraints are constant, for example $w_j^{(s)} = n_s^{-1}$ for all $s = 1, \dots, k$ and $j = 1, \dots, n_s$.

5 Beyond the Gaussian case: a Poisson example

The results of Section 4.2 provide guidance on the choice of which linear constraint to use to impose identifiability for models also beyond the Gaussian case. As an example, we consider the following crossed random effect model with Poisson likelihood, which is the simplest analogue of Model Ck in the context of count data.

Model CkP (Poisson crossed-effects model)**.**

$$y_{i_1\dots i_k} \sim Poisson(\mu\, a^{(1)}_{i_1}\cdots a^{(k)}_{i_k}) \qquad i_s = 1,\dots,n_s \; for\; s\in\{1,\dots,k\}, \tag{5.1}$$

with $a^{(s)}_{i_s} \overset{iid}{\sim} Gamma(\alpha_s,\beta_s)$ *for* $s=1,\dots,k$ *and* $\mu\sim Gamma(\alpha_\mu,\beta_\mu)$*.*

Consider sampling from the posterior distribution $\mathcal{L}(\mu,\boldsymbol{a}|\boldsymbol{y})$ of Model CkP using the standard Gibbs Sampler that, similarly to Sampler GS-crossed, at each iteration updates μ from $\mathcal{L}(\mu|\boldsymbol{a},\boldsymbol{y})$ and then $\boldsymbol{a}^{(s)}$ from $\mathcal{L}\left(\boldsymbol{a}^{(s)}|\mu,\boldsymbol{a}^{(-s)},\boldsymbol{y}\right)$ for $s=1,\dots,k$. Here $\boldsymbol{y}$, $\boldsymbol{a}$ and $\boldsymbol{a}^{(-s)}$ are defined as in the beginning of Section 4.

We explore the extent to which the conclusions drawn from Theorem 6 apply also to Model CkP by means of simulations. We consider the case $k=2$ with three different combinations of values of n_1 and n_2. The data $(y_{i_1i_2})$ are generated from the model with $(a^{(1)}_{i_1})$ and $(a^{(2)}_{i_2})$ sampled from their prior distributions and μ set to 1. For the prior hyperparameters we use $\alpha_1=\alpha_2=\alpha_\mu=2$ and $\beta_1=\beta_2=\beta_\mu=0.1$. We compare the standard Gibbs Sampler with no constraints, with the versions obtained by imposing the linear constraints $a^{(1)}_1=a^{(2)}_1=1$ and $\bar{a}^{(1)}=\bar{a}^{(2)}=1$, respectively, where $\bar{a}^{(1)}$ and $\bar{a}^{(2)}$ are defined as in (4.2). Table 2 reports the resulting effective same sizes (minimum and median across parameters). The results are consistent with the theoretical guidance offered by Theorem 6 since: (a) imposing identifiability always improves mixing of the samplers; (b) imposing constraints on $\bar{a}^{(1)}$ and $\bar{a}^{(2)}$ leads to faster convergence compared to imposing constraints on $a^{(1)}_1$ and $a^{(2)}_1$; (c) the difference in resulting efficiency between the two set of linear constraints increases with n_1 and n_2.

	ESS (min, median) $n_1=5$, $n_2=5$	ESS (min, median) $n_1=5$, $n_2=100$	ESS (min, median) $n_1=100$, $n_2=100$
Unconstrained	(16.4, 27.0)	(3.8, 32.8)	(3.0, 5.1)
$a^{(1)}_1=a^{(2)}_1=1$	(2798, 5716)	(341, 991)	(27.6, 126)
$\bar{a}^{(1)}=\bar{a}^{(2)}=1$	(49003, 50000)	(46597, 50000)	(46982, 50000)

Table 2: Effective sample sizes (ESS) for the standard unconstrained Gibbs Sampler for Model CkP, and for the two version where identifiability is obtained by imposing the constraints $a^{(1)}_1=a^{(2)}_1=1$ and $\bar{a}^{(1)}=\bar{a}^{(2)}=1$, respectively. ESS values correspond to 10^5 iterations of each algorithm, with the first half of the samples discarded as burn-in.

5.1 Comparison with Hamiltonian Monte Carlo

Finally, we also explore whether the results in Theorem 6 can be useful to guide the implementation of other MCMC schemes targeting Model CkP, such as Hamiltonian Monte

Carlo (HMC) (Neal et al., 2011) and the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) implemented in the widely used software STAN (Carpenter et al., 2017). We consider the same setting of the rightmost column of Table 2 where $n_1 = n_2 = 100$, comparing the Gibbs Sampler with HMC and NUTS (used their STAN implementation with default setting). Table 3 reports effective sample sizes (ESS), runtimes, number of leapfrog steps per iteration and ESS per leapfrog iteration for HMC/NUTS, ESS per sweep (update of μ, $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$) for Gibbs Sampling and ESS per unit of computation time for all schemes. Traceplots and autocorrelation functions are provided in the supplement. All simulations reported in Tables 2 and 3 were performed on the same desktop computer with 16 GB of RAM and an Intel core i7-7700 @ 3.60 GHz processor, using R (R Core Team, 2018). Effective sample sizes are estimated using the *coda* R package. The supplementary material provides the R code used to implement the Gibbs Samplers and the Stan code used to specify the models with linear constraints.

	ESS (min, median)	Runtime [s]	Leapfrog per iter.	min(ESS)/ n.leap	min(ESS)/ n.sweep	min(ESS)/ time [1/s]
HMC-v1	(1267, 2433)	13092	1412	9.0e-05	-	0.10
HMC-v2	(44.7, 2036)	1777	284	1.6e-05	-	0.03
HMC-v3	(1291, 6231)	1247	212	6.1e-04	-	1.04
NUTS-v1	(3.7, 39.7)	1901	325	1.1e-06	-	0.0019
NUTS-v2	(113, 355)	314	51	2.2e-04	-	0.36
NUTS-v3	(4977, 13341)	127	20	2.5e-02	-	39.1
Gibbs-v1	(4.6, 11.4)	0.74	-	-	4.5e-04	6.20
Gibbs-v2	(14.4, 130)	0.73	-	-	1.4e-03	19.7
Gibbs-v3	(4903, 5020)	0.82	-	-	4.9e-01	6016

Table 3: Comparison of HMC, NUTS and the Gibbs Sampler for Model CkP without linear constraints (v1) and with the linear constraints $a_1^{(1)} = a_1^{(2)} = 1$ (v2) or $\bar{a}^{(1)} = \bar{a}^{(2)} = 1$ (v3). ESS and runtimes of each algorithm refer to 10^4 iterations, with the first half of the samples discarded as burn-in. All numbers are averaged over 10 replicates.

First, Table 3 suggests that imposing identifiability through linear constraints helps significantly also gradient-based samplers such as HMC and NUTS, with a pattern in accordance with Theorem 6. The improvement involves both speeding up convergence (higher ESS per iteration) and reducing runtime, which is reduced because the number of required leapfrog steps per iteration (which is automatically tuned in STAN) gets lower for better identified and less correlated targets, as the one with the linear constraints. The only exception to this pattern is the fact that HMC-v1 (no constraints) is more efficient than HMC-v2 (constraint $a_1^{(1)} = a_1^{(2)} = 1$). This is mainly due to the very high number of leapfrog steps per iteration employed by HMC-v1 compared to HMC-v2, which ends up being beneficial in terms of efficiency despite the major increase in runtime. This phenomenon is related to the specific tuning procedures implemented in STAN and arguably does not contradict the general fact that improving identifiability by imposing linear constraints is beneficial for HMC and NUTS.

Second, Table 3 suggests that the Gibbs Sampler can be substantially more efficient than HMC and NUTS for random effect models such as Model CkP, thanks to a lower runtime and, arguably, a more direct use of conditional independence among random effects. We note that empirical runtimes can be highly dependent on software implementations and this could be unfavourable to a generic software implementation such as the STAN. In order to obtain an implementation-independent comparison of efficiency one should combine the ESS per leapfrog iteration and ESS per sweep values reported in Table 3 with theoretical considerations on the computational costs of such operations, which however are non-trivial and highly model-dependent. We leave a more detailed investigation of these aspects, both theoretical and computational, to future work.

Remark 2. *Interestingly, the multigrid decomposition can be applied also to Model CkP, with the appropriate modifications. In this case the Markov chain* $((\mu,\boldsymbol{a})(t))_{t=1}^{\infty}$ *induced by the Gibbs Sampler can be decomposed into* $(k+1)$ *independent Markov chains* $\big((\mu,\tilde{a}^{(1)},\dots,\tilde{a}^{(k)})(t)\big)_{t=1}^{\infty}$ *and* $\big(\tilde{\delta}\boldsymbol{a}^{(1)}(t)\big)_{t=1}^{\infty},\dots,\big(\tilde{\delta}\boldsymbol{a}^{(k)}(t)\big)_{t=1}^{\infty}$*, where* $\tilde{a}^{(s)}=\sum_{i_s}a_{i_s}^{(s)}$ *and* $\tilde{\delta}a_{i_s}^{(s)}=a_{i_s}^{(s)}/\tilde{a}^{(s)}$*. In this case the rate of convergence of the original chain coincides with the one of* $\big((\mu,\tilde{a}^{(1)},\dots,\tilde{a}^{(k)})(t)\big)_{t=1}^{\infty}$*, which evolves according to a* $(k+1)$*-dimensional Gibbs Sampler with full conditionals given by:*

$$\begin{aligned}\mu|\boldsymbol{y},\tilde{a} &\sim Gamma(\alpha_\mu+y.,\beta_\mu+\prod_{s=1}^{k}\tilde{a}^{(s)})\,,\\ \tilde{a}^{(s)}|\boldsymbol{y},\mu,\tilde{a}^{(-s)} &\sim Gamma(I\alpha_s+y.,\beta_s+\mu\prod_{\ell\neq s}^{k}\tilde{a}^{(\ell)}) \qquad \text{for } s\in\{1,\dots,k\}\,,\end{aligned} \tag{5.2}$$

where $y.=\sum_{i_1,\dots,i_k}y_{i_1\dots i_k}$*. We expect such a* $(k+1)$*-dimensional Gibbs Sampler to be potentially amenable to analysis using the framework of iterated random functions (Diaconis and Freedman, 1999), in order to obtain an upper bound on convergence rates (see e.g. Alsmeyer and Fuh, 2001, Theorem 2.1.(b)). We leave these extensions to future works and mention it in Section 8 as a possible avenue for future research directions.*

6 Non-symmetric hierarchical models

Section 3 describes how to optimize parametrization as a function of $(I,J,K,\sigma_a,\sigma_b,\sigma_e)$ for Model S3. In general, both the variance terms σ_b^2 and σ_e^2, and the number of branches J and K could depend on i and j. In this section we consider non-symmetric cases for two and three level hierarchical models. In these non-symmetric cases the computationally optimal strategy will involve centering some branches of the hierarchy and non-centering others: we will call these strategies *bespoke parametrizations*.

Consider the following non-symmetric 2-levels model (which we describe in terms of precisions rather than variances for notational convenience).

Model NS2 (Non-symmetric 2-levels hierarchical model)**.** *Consider the following 2-levels model with centred parametrization*

$$p(\mu)\propto 1\,,\qquad \gamma_i\sim N(\mu,\tau_a^{-1})\,,\qquad y_{ij}\sim N(\gamma_i,\tau_{e,i}^{-1})\,,\qquad i=1,\dots,I;j=1,\dots,J_i\,,$$

where the precision components $(\tau_a,(\tau_{e,i})_i)$ *are assumed to be known.*

Papaspiliopoulos et al. (2003) studied the symmetric version of Model NS2, where $J_i = J$ and $\tau_{e,i} = \tau_e$ for all i and some fixed J and τ_e. They showed that the convergence rates induced by the centred and non-centred parametrizations respectively are

$$\rho_{CP} = \frac{\tau_a}{\tau_a + \tilde{\tau}_e} \quad \text{and} \quad \rho_{NCP} = \frac{\tilde{\tau}_e}{\tau_a + \tilde{\tau}_e}\,, \tag{6.1}$$

where $\tilde{\tau}_e = J\tau_e$. The following Theorem provides an extension to the general non-symmetric case. We consider Model NS2 with a bespoke parametrization $(\mu, \beta_1, \dots, \beta_I)$ defined by I indicators $(\lambda_1, \dots, \lambda_I) \in \{0,1\}^I$ as $\beta_i = \gamma_i - \lambda_i\mu$, meaning that λ_i equals 0 if component i is centred and 1 if it is non-centred.

Theorem 7. *The rate of convergence of the Gibbs Sampler targeting Model NS2 with bespoke parametrization given by* $(\lambda_1, \dots, \lambda_I) \in \{0,1\}^I$ *is*

$$\rho_{\lambda_1 \dots \lambda_I} = \frac{\sum_{i\,:\,\lambda_i=1} \tilde{\tau}_i \frac{\tilde{\tau}_i}{\tilde{\tau}_i + \tau_a} + \sum_{i\,:\,\lambda_i=0} \tau_a \frac{\tau_a}{\tilde{\tau}_i + \tau_a}}{\sum_{i\,:\,\lambda_i=1} \tilde{\tau}_i + \sum_{i\,:\,\lambda_i=0} \tau_a}\,, \tag{6.2}$$

where $\tilde{\tau}_i = J_i\tau_{e,i}$.

Equation (6.2) shows that in the non-symmetric case, the GS rate of convergence is given by a weighted average of the precision ratios $\frac{\tau_a}{\tilde{\tau}_i + \tau_a}$ and $\frac{\tilde{\tau}_i}{\tilde{\tau}_i + \tau_a}$ depending on whether each component is centred or not. This has clear analogies with the symmetric case in (6.1). The weights in the average of (6.2) are themselves functions of $(\lambda_1, \dots, \lambda_I)$, thus introducing dependence across components in terms of centering and the overall convergence rate. Nonetheless, the following corollary shows that even in the context of Model NS2, optimizing parametrization in each branch of the tree can be carried out independently following the same intuition of the symmetric case: for each i in $\{1, \dots, I\}$ use centred parametrization γ_i if and only if $\tau_a \leq J_i\tau_{e,i}$, otherwise use a non-centred parametrization $a_i = \gamma_i - \mu$.

Corollary 3. *Let* $\bar{\lambda}_i = \mathbb{1}(\tau_a > \tilde{\tau}_i)$ *for all* i *from 1 to* I. *Then*

$$\rho_{\bar{\lambda}_1 \dots \bar{\lambda}_I} \leq \rho_{\lambda_1 \dots \lambda_I} \qquad \textit{for any } (\lambda_1 \dots \lambda_I) \in \{0,1\}^I\,.$$

By (6.2), the strategy described in Corollary 3 ensures that $\rho_{\bar{\lambda}_1 \dots \bar{\lambda}_I} \leq 1/2$. This is the same upper bound one can obtain in the symmetric case (see (6.1) and Papaspiliopoulos et al. (2003)), meaning that in this case bespoke parametrizations are successful in dealing with the lack of symmetry.

Consider now the three-level non-symmetric case.

Model NS3 (Non-symmetric 3-levels hierarchical model)**.** *Consider a more general 3-levels model with centred parametrization*

$$\begin{aligned} p(\mu) &\propto 1 & & \\ \gamma_i &\sim N(\mu, \sigma_a^2) & & i = 1, \dots, I, \\ \eta_{ij} &\sim N(\gamma_i, \sigma_{b,i}^2) & & j = 1, \dots, J_i, \\ y_{ijk} &\sim N(\eta_{ij}, \sigma_{e,ij}^2) & & k = 1, \dots, K_{i,j}, \end{aligned}$$

where variance components are assumed to be known.

In this case the multigrid factorization of Theorem 1 does not apply directly to Model NS3, but nonetheless it can still be used to obtain upper bounds on the rates of convergence combining it with monotonicity properties of the spectral radius of non-negative matrices (see the supplement and Roberts and Sahu, 1997, Theorem 7 for details).

Theorem 8. *Given an instance of Model NS3 we define*

$$r_{a,b}^{(i)} = \frac{\sigma_a^2}{\sigma_a^2 + J_i^{-1}\sigma_b^2}, \quad \textit{and} \quad r_{e,b}^{(i)} = \frac{1}{J_i}\sum_{j=1}^{J_i} \frac{K_{ij}^{-1}\sigma_{e,ij}^2}{\sigma_{b,i}^2 + K_{ij}^{-1}\sigma_{e,ij}^2}\,.$$

If $r_{a,b}^{(i)} \geq r_{a,b}^{(i')} r_{e,b}^{(i')}$ for every $i, i' \in \{1, \dots, I\}$, then the rate of convergence of the Gibbs Sampler with centred parametrization $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ satisfies

$$\rho \;\leq\; 1 - \frac{1}{I}\sum_{i=1}^{I} r_{a,b}^{(i)} + \max_{i=1,\dots,I} r_{a,b}^{(i)} r_{e,b}^{(i)}\,.$$

The results of Theorem 8 suggest that as the number of data points increase the efficiency of the Gibbs sampler with centred parametrization increases. In fact, as K_{ij} increases the assumptions of Theorem 8 are eventually satisfied and the bound on the convergence rate goes to 0 as J_i and K_{ij} increase. Theorem 8 provides rigorous theoretical support and characterization of the well known fact that the centred parametrization is to be preferred in contexts of large and informative datasets (Gelfand et al., 1995; Papaspiliopoulos et al., 2003). We note that the convergence rate for the Gibbs Sampler targeting Model NS3 is not easily tractable, and that deriving analytic expressions for the optimal bespoke parametrization in this context is still an open problem.

7 Hierarchical linear models with arbitrary number of levels

Here we consider Gaussian hierarchical models with k levels for arbitrary k. We refer to the highest level of the hierarchy (i.e. the one furthest away from the data) as level 0, down to level $k-1$ the lowest level (i.e. closest to the data). The 3 level model of Section 3 is a special case of the theory developed here where $k = 3$.

7.1 Model formulation

In order to allow for more generality and keep the notation concise, in this section we will use a graphical models notation. In particular T will denote a finite tree with k levels and root $t_0 \in T$. For each node $t \in T$ we will denote by $pa(t)$ the unique parent of t and by $ch(t)$ the collection of children of t. Moreover we write $s \preceq t$ and $s \succeq t$ if s is respectively an ancestor or a descendant of t (with s and t possibly being equal) while $s \prec t$ and $s \succ t$ denote the same notions with the additional condition of $s \neq t$. For each node $t \in T$ we denote by $\ell(t)$ the level of node t (i.e. its distance from t_0). For each

$d \in \{0, \dots, k-1\}$ we denote by $T_d = \{t \in T : \ell(t) = d\}$ the collection of nodes at level d. For example we have $T_0 = \{t_0\}$ and $T = \cup_{d=0}^{k-1} T_d$. Noisy observations will occur only at leaf nodes. The collection of leaf nodes is denoted as $T_L = \{t \in T : ch(t) = \emptyset\}$. For simplicity we assume that all leaf nodes are at level $k-1$, i.e. $T_L = T_{k-1}$, although this assumption could be easily relaxed allowing some branches to be longer than others.

Model NSk (k-levels hierarchical model)**.** *Suppose that we have a hierarchy described by a tree T with observations occurring at leaf nodes T_L. We assume the following hierarchical model*

$$y_t^{(i)} \sim N(\gamma_t, 1/\tau_t^{(e)}) \qquad t \in T_L\,, \tag{7.1}$$

$$\gamma_t \sim N(\gamma_{pa(t)}, 1/\tau_t) \qquad t \in T \backslash t_0, \tag{7.2}$$

where $i \in \{1, \dots, n_t\}$ with n_t being the number of observed data at leaf node t, $(\tau_t)_{t \in T\backslash t_0}$ and $(\tau_t^{(e)})_{t \in T_L}$ are known precision components and all normal random variables are sampled independently. Following the standard Bayesian model specification we assume a flat prior on γ_{t_0}.

We are interested in sampling from the posterior distribution of $\boldsymbol{\gamma}_T = (\gamma_t)_{t \in T}$ given some observations $\mathbf{y} = (y_t)_{t \in T_L}$. The centred parametrization $\boldsymbol{\gamma}_T$ of Model NSk leads to the following Gibbs Sampler.

Sampler GS($\boldsymbol{\gamma}_T$). *Initialize $\boldsymbol{\gamma}_T(0)$ and then iterate the following kernel:*

For $d = 0, \dots, k-1$, sample $\gamma_t(s+1)$ from $p(\gamma_t | \boldsymbol{\gamma}_{T_{d-1}}(s+1), \boldsymbol{\gamma}_{T_{d+1}}(s), \mathbf{y})$ for all $t \in T_d$, where $p(\gamma_t | \boldsymbol{\gamma}_{T_{d-1}}, \boldsymbol{\gamma}_{T_{d+1}}, \mathbf{y}) = p(\gamma_t | \boldsymbol{\gamma}_{T\backslash t}, \mathbf{y})$ is the full conditional distribution of γ_t given by Model NSk. When d equals 0 or $k-1$ the levels $\boldsymbol{\gamma}_{T_{d-1}}$ and $\boldsymbol{\gamma}_{T_{d+1}}$ have to be replaced by empty sets in the conditioning.

7.2 Non-centering and hierarchical reparametrizations

Model NSk expresses Gaussian hierarchical models using a centred parametrization. The corresponding non-centred version is given by the following example.

Example 1 (Fully non-centred parametrization)**.** *Under the same setting as Model NSk, define*

$$y_t^{(i)} \sim N\Big(\sum_{r \preceq t} \alpha_r, 1/\tau_t^{(e)}\Big) \qquad t \in T_L,$$

$$\alpha_t \sim N(0, 1/\tau_t) \qquad t \in T \backslash t_0,$$

and assume a flat prior on α_{t_0}.

The non-centred parametrization $\boldsymbol{\alpha}_T$ can be obtained as a linear transformation of the centred version $\boldsymbol{\gamma}_T$ of Model NSk. More generally, we will consider the class of parametrizations that can be obtained by reparametrizing $\boldsymbol{\gamma}_T$ as follows.

Definition 1 (Hierarchical reparametrizations)**.** *Let $\boldsymbol{\gamma}_T = (\gamma_t)_{t\in T}$ be a random vector with elements indexed by a tree T. We say that $\boldsymbol{\beta}_T = (\beta_t)_{t\in T}$ is a hierarchical (linear) reparametrization of $\boldsymbol{\gamma}_T$ if*

$$\beta_t = \sum_{r \preceq t} \lambda_{tr}\gamma_r \qquad t \in T, \tag{7.3}$$

for some real-valued coefficients $\Lambda = (\lambda_{tr})_{r\preceq t, t\in T}$ satisfying $\lambda_{tt} \neq 0$ for all $t \in T$. We denote (7.3) by $\boldsymbol{\beta}_T = \Lambda\boldsymbol{\gamma}_T$.

Using terminology from Papaspiliopoulos et al. (2003), we refer to the family of hierarchical reparametrizations of $\boldsymbol{\gamma}_T = (\gamma_t)_{t\in T}$ as *partially non-centred parametrizations* (PNCP) of Model NSk. Note that (7.3) does not span the space of all linear transformations of $\boldsymbol{\gamma}_T$. In fact $\Lambda = (\lambda_{tr})_{r\preceq t, t\in T}$ can be thought as a $|T| \times |T|$ matrix $\Lambda = (\lambda_{tr})_{r,t\in T}$ inducing a linear transformation of $\boldsymbol{\gamma}_T$ with the additional sparsity constraint of being zero on all elements λ_{tr} such that $r \npreceq t$. The following Lemma shows that the definition of the class of PNCP does not depend on the starting parametrization used to formulate Model NSk. For example, one could equivalently define the class of PNCP of Model NSk as the collection of hierarchical reparametrization of the non-centred parametrization $\boldsymbol{\alpha}_T$ of Example 1.

Lemma 1. *If $\boldsymbol{\beta}_T$ is a hierarchical reparametrization of $\boldsymbol{\gamma}_T$, then also $\boldsymbol{\gamma}_T$ is a hierarchical reparametrization of $\boldsymbol{\beta}_T$.*

As for the 3-levels case we are interested in assessing the computational efficiency of the different Gibbs Sampling schemes arising from different PNCP's. For each PNCP $\boldsymbol{\beta}_T$ the corresponding Gibbs Sampler scheme $GS(\boldsymbol{\beta}_T)$ is defined analogously to $GS(\boldsymbol{\gamma}_T)$.

Sampler GS($\boldsymbol{\beta}_T$). *Initialize $\boldsymbol{\beta}_T(0)$ and then iterate the following kernel:*

For $d = 0, \ldots, k-1$, sample $\beta_t(s+1)$ from $p(\beta_t|(\boldsymbol{\beta}_{T_p}(s+1))_{0\leq p<d}, (\boldsymbol{\beta}_{T_p}(s))_{d<p\leq k-1}, \mathbf{y})$ for all $t \in T_d$, where $p(\beta_t|(\boldsymbol{\beta}_{T_p})_{0\leq p<d}, (\boldsymbol{\beta}_{T_p})_{d<p\leq k-1}, \mathbf{y}) = p(\beta_t|\boldsymbol{\beta}_{T\setminus t}, \mathbf{y})$ is the full conditional distribution of β_t given by Model NSk.

Sampler GS($\boldsymbol{\beta}_T$) is easy to implement because the family of PNCP preserves the hierarchical structure of Model NSk. In fact, any PNCP of Model NSk exhibits the following conditional independence structure:

$$\beta_r \perp \beta_t | \boldsymbol{\beta}_{T\setminus\{r,t\}} \text{ unless } r \preceq t \text{ or } t \preceq r\,. \tag{H}$$

Note that this is a weaker condition than the one holding for the centred parametrization $\boldsymbol{\gamma}_T$. In the latter case, the conditional independence graph corresponds exactly to the tree T, meaning that if $r \neq t$

$$\gamma_r \perp \gamma_t | \boldsymbol{\gamma}_{T\setminus\{r,t\}} \text{ unless } r = pa(t) \text{ or } t = pa(r)\,. \tag{T}$$

Despite being weaker than (T), condition (H) still guarantees that all parameters at the same level are conditionally independent (thus simplifying the update of $\boldsymbol{\beta}_{T_d}|\boldsymbol{\beta}_{T\setminus T_d}$) and that the full conditional distribution of each β_t depends only on the descendants or ancestors of t. The following Lemma and Corollary provide a simple way to check that any PNCP of Model NSk satisfies (H).

Lemma 2. *Property (H) is closed under hierarchical re-parametrizations, meaning that if $\boldsymbol{\beta}_T$ satisfies (H) then any hierarchical re-parametrization of $\boldsymbol{\beta}_T$ satisfies (H) too.*

Corollary 4. *Any PNCP $\boldsymbol{\beta}_T$ of Model NSk satisfies (H).*

7.3 Symmetry assumption

To provide a full analysis of the convergence properties of Sampler GS($\boldsymbol{\beta}_T$) we need a symmetry assumption that we now define. Let ρ_{tr} denote the partial correlation $Corr\left(\beta_t, \beta_r \middle| \boldsymbol{\beta}_{T\setminus\{t,r\}}\right)$, namely

$$\rho_{tr} = -\frac{Q_{tr}}{\sqrt{Q_{tt}Q_{rr}}} \qquad\qquad t \neq r\,,$$

and $\rho_{tt} = 1$ for all t. Here Q is the precision matrix of $\boldsymbol{\beta}_T$. Let $\mathbf{X} = (X_\ell)_{\ell=0}^{k-1}$ be a random walk going through T from root to leaves as follows: $X_0 = t_0$ almost surely and then, for $\ell \in \{0, \dots, k-2\}$

$$P(X_{\ell+1} = t \,|\, X_\ell = r) = \frac{\rho_{tr}^2}{\sum_{t' \in ch(r)} \rho_{t'r}^2} \mathbb{1}(t \in ch(r))\,. \tag{7.4}$$

Equation (7.4) implies that at each step $\mathbf{X}$ jumps from the current state r to one of its children $t \in ch(r)$ choosing t proportionally to the squared partial correlation between β_r and β_t. Since $\ell(X_d) = d$ almost surely for all $d \in \{0, \dots, k-1\}$ we can use the following simplified notation: for any t and r in T we use $P(t)$, $P(t|r)$ and $P(t \cap r)$ to denote respectively $P(X_{\ell(t)} = t)$, $P(X_{\ell(t)} = t \,|\, X_{\ell(r)} = r)$ and $P(X_{\ell(t)} = t \cap X_{\ell(r)} = r)$.

Given the above definitions, we define the following symmetry condition: there exist a $k \times k$ symmetric matrix $C = (c_{dp})_{d,p=0}^{k-1}$ such that

$$\rho_{tr} = c_{\ell(r)\ell(t)} \sqrt{P(t|r)} \qquad\qquad r \preceq t\,, \tag{S}$$

and $\rho_{tr} = 0$ if $r \npreceq t$ and $t \npreceq r$. Note that ρ_{tr} is invariant to coordinate-wise rescaling of $\boldsymbol{\beta}_T$ and therefore both property (S) and the transition kernel of $\mathbf{X}$ are invariant to rescalings. Therefore we can consider, without loss of generality, the following rescaled version of $\boldsymbol{\beta}_T$ defined by

$$\tilde{\beta}_t = \beta_t \sqrt{\frac{Q_{tt}}{P(t)}} \qquad\qquad t \in T\,. \tag{7.5}$$

Condition (S) can then be written in terms of the precision matrix of $\tilde{\boldsymbol{\beta}}_T = (\tilde{\beta}_t)_{t \in T}$ as

$$\tilde{Q}_{tt} = P(t) \quad \text{and} \quad -\tilde{Q}_{tr} = c_{\ell(t)\ell(r)} P(t \cap r) \quad \text{for } t \neq r\,. \tag{$\widetilde{\text{S}}$}$$

The rescaled version $\tilde{\boldsymbol{\beta}}_T$ will be helpful to design an appropriate multigrid decomposition of $\boldsymbol{\beta}_T$. Also, property ($\widetilde{\text{S}}$) is closed under symmetric hierarchical parametrizations.

Definition 2 (Symmetric hierarchical reparametrizations)**.** *We say that $\boldsymbol{\beta}_T = \Lambda \boldsymbol{\alpha}_T$ is a symmetric hierarchical reparametrization of $\boldsymbol{\alpha}_T$ if the coefficients of $\Lambda = (\lambda_{tr})_{r \preceq t, t \in T}$ depend only on the levels of r and t in the hierarchy T.*

Lemma 3. *Property ($\widetilde{\mathrm{S}}$) is closed under symmetric hierarchical reparametrizations, meaning that if $\tilde{\boldsymbol{\beta}}_T$ satisfies ($\widetilde{\mathrm{S}}$) then any symmetric hierarchical reparametrization of $\tilde{\boldsymbol{\beta}}_T$ satisfies ($\widetilde{\mathrm{S}}$) too.*

Various special cases of Model NSk satisfy assumption (S). For example, we now consider three cases: a fully symmetric case (both the tree T and the variances $(\tau_t)_{t \in T}$ are symmetric), a weakly symmetric case (non-symmetric tree and symmetric variances) and a non-symmetric case (both tree and variances non-symmetric).

Model Sk (Symmetric k-levels hierarchical model)**.** *Consider the k-level Gaussian Hierarchical model where the observed data are generated from*

$$y_{i_1,\dots,i_{k-1},j} \sim N(\gamma^{(k-1)}_{i_1,\dots,i_{k-1}}, 1/\tau_e) \qquad (i_1,\dots,i_{k-1},j) \in [I_1] \times \cdots \times [I_{k-1}] \times [J]\,,$$

where $[N] = \{1,\dots,N\}$ for any positive integer N. The parameters have the following hierarchical structure: for each level d from 1 to $k-1$

$$\gamma^{(d)}_{i_1,\dots,i_d} \sim N(\gamma^{(d-1)}_{i_1,\dots,i_{d-1}}, 1/\tau_d) \qquad\qquad (i_1,\dots,i_d) \in [I_1] \times \cdots \times [I_d]\,.$$

Here $(\tau_1,\dots,\tau_{k-1},\tau_e)$ are known precisions and the root parameter $\gamma^{(0)}$ is given a flat prior $p(\gamma^{(0)}) \propto 1$. For each $d \in \{1,\dots,k-1\}$ the positive integer I_d represents the number of branches from level $d-1$ to level d.

It is easy to see that the posterior distribution of Model Sk, conditioned on the observed data $\mathbf{y} = (y_{i_1,\dots,i_{k-1},j})_{i_1,\dots,i_{k-1},j}$, satisfies (S). In this case the random walk $\mathbf{X}$ defined by (7.4) coincides with the natural random walk going through T.

Example 2 (Weakly symmetric case)**.** *Another special case of Model NSk satisfying (S) is given by the case of a general tree T and precision terms defined as $\tau_t = \frac{\tau_{\ell(t)}}{\prod_{s \prec t} |ch(s)|}$ for all $t \in T$ and $\tau^{(e)}_t = \frac{\tau_e}{n_t \prod_{s \prec t} |ch(s)|}$, where $(\tau_1,\dots,\tau_k,\tau_e) \in \mathbb{R}^{k+1}_+$ are level-specific precision terms. This is an extension of Model Sk where the lack of symmetry of T is compensated by appropriate variance terms. Condition (S) can be checked by evaluating the partial correlations $(\rho_{tr})_{t,r \in T}$ of the resulting vector $\boldsymbol{\gamma}_T$.*

Example 3 (Non-symmetric cases)**.** *In both Model Sk and Example 2 the auxiliary Markov chain $\mathbf{X}$ defined in (7.4) follows a natural random walk, in the sense that at each time the chain chooses the next state uniformly at random among children nodes. However, condition (S) is also satisfied by non-symmetric cases where $\mathbf{X}$ is not a natural random walk. In particular any instance of Model NSk such that*

$$\sum_{r \in ch(t)} \rho^2_{tr} \;=\; c_{\ell(t)} \qquad \textit{for all } t \in T \backslash T_L\,, \tag{S*}$$

for some $(k-1)$-dimensional vector $(c_0,\dots,c_{k-2})$ induces a posterior distribution satisfying (S). In fact, in the context of Model NSk conditions (S) and (S) are equivalent (this can be derived from (T) and (7.4)).*

The cases previously considered are expressed in terms of centred parametrization, meaning that as all the instances of Model NSk they satisfy (T). Nevertheless Lemma 3 shows that any symmetric hierarchical reparametrization of a vector satisfying ($\widetilde{\mathrm{S}}$) still satisfies ($\widetilde{\mathrm{S}}$). This implies, for example, that the fully non-centred version of Model Sk and any mixed strategy where some level is centred and some is not centred, still satisfies ($\widetilde{\mathrm{S}}$) (after rescaling). Moreover, note that the exact analysis we will now provide for the Gibbs sampler on models satisfying ($\widetilde{\mathrm{S}}$) can be used to provide bound on general cases that do not satisfy ($\widetilde{\mathrm{S}}$) (see for example Theorem 8).

7.4 Multigrid decomposition

We now show how to use the multigrid decomposition to analyse the Gibbs Sampler for random vectors $\boldsymbol{\beta}_T$ satisfying (H) and (S). Our aim is to provide a transformation of $\boldsymbol{\beta}_T$ that factorizes the Gibbs Sampler Markov Chain into independent and more tractable sub-chains. Similarly to Section 3 in the following we will often denote $\boldsymbol{\beta}_{T_d} = (\beta_t)_{t\in T_d}$ by $\boldsymbol{\beta}^{(d)}$. We proceed in two steps, first introducing the averaging operators $\phi^{(p)}$ and then the residual operators $\delta^{(p)}$. For any $p \leq d$ the averaging operator $\phi^{(p)} : \mathbb{R}^{T_d} \to \mathbb{R}^{T_p}$ is defined as

$$\phi_r^{(p)}\boldsymbol{\beta}^{(d)} = \mathbb{E}[\beta_{X_d}|\boldsymbol{\beta}_T, X_p = r] = \sum_{t\in T_d} \beta_t P(t|r) \qquad r \in T_p\,, \tag{7.6}$$

where $\mathbf{X} = (X_\ell)_{\ell=0}^{k-1}$ is the Markov chain defined by (7.4). Loosely speaking $\phi^{(p)}\boldsymbol{\beta}^{(d)} = \mathbb{E}[\beta_{X_d}|\boldsymbol{\beta}_T, X_p]$ can be interpreted as the averages of $\boldsymbol{\beta}^{(d)}$ at the coarseness corresponding to the p-th level of the hierarchy. In particular $\phi^{(d)}\boldsymbol{\beta}^{(d)} = \boldsymbol{\beta}^{(d)}$ and $\phi_{t_0}^{(0)}\boldsymbol{\beta}^{(d)} = \mathbb{E}[\beta_{X_d}|\boldsymbol{\beta}_T]$.

Example 4 (Averaging operators in the symmetric case)**.** *Let $\boldsymbol{\beta}_T = \boldsymbol{\gamma}_T$ be given by Model Sk. Then*

$$\phi_r^{(p)}\boldsymbol{\beta}^{(d)} = \frac{1}{\prod_{\ell=p+1}^d I_\ell}\left(\sum_{t\in T_d\,:\,t\succeq r} \beta_t\right) \qquad r \in T_p\,.$$

Given ϕ, we define the residual operators $\delta^{(p)} : \mathbb{R}^{T_d} \to \mathbb{R}^{T_p}$ as $\delta^{(p)} = (\delta_r^{(p)})_{r\in T_p}$ with $\delta_r^{(p)} : \mathbb{R}^{T_d} \to \mathbb{R}$ defined as

$$\delta_r^{(p)}\boldsymbol{\beta}^{(d)} = \phi_r^{(p)}\boldsymbol{\beta}^{(d)} - \phi_{pa(r)}^{(p-1)}\boldsymbol{\beta}^{(d)} \qquad r \in T_p\,, \tag{7.7}$$

for $1 \leq p \leq d \leq k-1$ and $\delta^{(0)}\boldsymbol{\beta}^{(d)} = \phi^{(0)}\boldsymbol{\beta}^{(d)}$ for $0 = p \leq d \leq k-1$. Analogously to the 3-level case of Section 3, under suitable assumptions the residual operators $\delta^{(p)}$ decompose the Markov chain generated by the Gibbs Sampler into independent sub-chains. To prove the result we first need the following lemma.

Lemma 4 (p-residuals interact only with p-residuals)**.** *Let $\boldsymbol{\beta}_T$ be a Gaussian random vector satisfying* (H) *and* ($\widetilde{\mathrm{S}}$)*. Then for any p and d with $0 \leq p \leq d \leq k-1$, for all*

$r \in T_p$ we have the identity

$$\mathbb{E}[\delta_r^{(p)}\boldsymbol{\beta}^{(d)}|\boldsymbol{\beta}\backslash\boldsymbol{\beta}^{(d)}] - \mathbb{E}[\delta_r^{(p)}\boldsymbol{\beta}^{(d)}] = \sum_{\ell\in\{p,\dots,k-1\}\backslash d} c_{d\ell}\left(\delta_r^{(p)}\boldsymbol{\beta}^{(\ell)} - \mathbb{E}[\delta_r^{(p)}\boldsymbol{\beta}^{(\ell)}]\right).$$

Theorem 9 (Multigrid decomposition for k levels)**.** *Let $(\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ be a Markov chain evolving according to GS($\boldsymbol{\beta}_T$) with $\boldsymbol{\beta}_T$ satisfying (H) and ($\widetilde{\text{S}}$). Then the functionals $(\delta^{(0)}\boldsymbol{\beta}(s))_s, \dots, (\delta^{(k-1)}\boldsymbol{\beta}(s))_s$ are k independent Markov chains. Moreover, each chain $\delta^{(p)}\boldsymbol{\beta}(s) = (\delta^{(p)}\boldsymbol{\beta}^{(d)}(s))_{d=p}^{k-1}$ evolves according to the following blocked Gibbs sampler scheme with $(k-p)$ components: for d going from p to $k-1$ sample*

$$\delta^{(p)}\boldsymbol{\beta}^{(d)}(s+1) \sim \mathcal{L}\left(\delta^{(p)}\boldsymbol{\beta}^{(d)}|(\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s+1))_{p\leq\ell<d}, (\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s))_{d<\ell\leq k-1}\right), \tag{7.8}$$

where $\mathcal{L}(X|Y)$ denotes the conditional distribution of X given Y.

Theorem 9 implies that running a Gibbs sampler $(\boldsymbol{\beta}(s))_s$ targeting distributions satisfying (H) is equivalent to running k independent blocked Gibbs Samplers, one for each level of coarseness, from $\delta^{(0)}\boldsymbol{\beta}(s)$ to $\delta^{(k-1)}\boldsymbol{\beta}(s)$.

Corollary 5. *Let $\boldsymbol{\beta}_T$ satisfy (H) and ($\widetilde{\text{S}}$). Then the rate of convergence of GS($\boldsymbol{\beta}_T$) is given by $\max\{\rho_0, \dots, \rho_{k-1}\}$ where for each $p \in \{0, \dots, k-1\}$, ρ_p is the rate of convergence of $(\delta^{(p)}\boldsymbol{\beta}(s))_s$.*

7.5 Hierarchical ordering of rates

Combining the results in Roberts and Sahu (1997, Section 2.2) with the multigrid decomposition, we can characterize the rates of convergence of the k independent Markov chains described above as follows.

Theorem 10. *The rate of convergence of $(\delta^{(p)}\boldsymbol{\beta}(s))_s$ is given by the largest modulus eigenvalue of $(\mathbb{I}_{k-p} - L)^{-1}U$. Here $\mathbb{I}_{k-p}$ is the $(k-p)$ dimensional identity matrix, while L and U are, respectively, the strictly lower and strictly upper triangular part of $(c_{d\ell})_{d,\ell=p}^{k-1}$, with $c_{d\ell}$ given by ($\widetilde{\text{S}}$).*

Theorem 10 implies that the convergence properties of the k independent Markov chains are closely related. In particular, from the rates of convergence point of view, the k Markov chains updating $\delta^{(p)}\boldsymbol{\beta}$ for $p = 0, \dots, k-1$ behave as Gibbs samplers targeting a decreasing number of dimensions (from k down to 1) of a single k-dimensional Gaussian distribution with precision matrix given by $-C$, where $C = (c_{d\ell})_{d,\ell=p}^{k-1}$ is given by ($\widetilde{\text{S}}$). This suggests that the convergence properties of the sub-chains will typically improve from that of $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ to that $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$ and that the rate of convergence of $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ will typically determine the rate of the whole sampler GS($\boldsymbol{\beta}_T$). In particular, in the centred parametrization case we can use the well-known Cauchy interlacing theorem (see e.g. Bhatia, 2013) to show that the rate of convergence is monotonically non-increasing from $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ to $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$.

Theorem 11 (Ordering of rates for centred parametrization)**.** *Let γ be a Gaussian vector satisfying* (T) *and* ($\widetilde{\mathrm{S}}$) *and let $(\gamma(s))_{s\in\mathbb{N}}$ be the corresponding Markov chain evolving according to $GS(\gamma_T)$. Then the convergence rates of the k independent Markov chains $(\delta^{(0)}\gamma(s))_s, \dots, (\delta^{(k-1)}\gamma(s))_s$ satisfy*

$$\rho(\delta^{(0)}\gamma(s)) \geq \rho(\delta^{(1)}\gamma(s)) \geq \cdots \geq \rho(\delta^{(k-1)}\gamma(s)) = 0\,. \tag{7.9}$$

In Theorem 11 we needed the additional assumption (T) to prove (7.9). The reason is that, while in most cases the convergence rates of a deterministic-scan Gibbs Sampler targeting a n-th dimensional Gaussian distribution improves if one of the coordinates is conditioned to a fixed value and the sampler targets only the remaining $(n-1)$ coordinates, this is not true in general. Example 2 of Roberts and Sahu (1997) provides a counter-example (see also Whittaker, 1990, page 319). In Roberts and Sahu (1997), this example was used a counter-example regarding blocking strategies, it also works in the present context. We note that, if one were to consider a random scan version of the Gibbs Sampler, the reversibility of the induced Markov chains would allow us to prove the ordering result in Theorem 11 with no need to assume (T). We leave this as a direction of future research and briefly mention it in Section 8.

Theorem 11 implies the following corollary.

Corollary 6. *Let γ be a Gaussian vector satisfying* (T) *and* ($\widetilde{\mathrm{S}}$)*. Then the rate of convergence of $GS(\gamma_T)$ is given by the largest squared eigenvalue of the k-dimensional matrix $C - \mathbb{I}_k$, where $C = (c_{d\ell})_{d,\ell=0}^{k-1}$ is defined by* ($\widetilde{\mathrm{S}}$) *and $\mathbb{I}_k$ is the k-dimensional matrix.*

In the special case of Model Sk it is easy to deduce the following result.

Corollary 7. *The rate of convergence of $GS(\gamma_T)$ targeting Model Sk is given by the largest squared eigenvalue of the k-dimensional matrix*

$$\begin{pmatrix} 0 & r_1 & & & \\ (1-r_2) & 0 & r_2 & & \\ & \dots & \dots & \dots & \\ & & (1-r_{k-2}) & 0 & r_{k-2} \\ & & & (1-r_{k-1}) & 0 \end{pmatrix},$$

where $r_\ell = \frac{I_\ell \tau_\ell}{\tau_{\ell-1} + I_\ell \tau_\ell}$ with $(\tau_1, \dots, \tau_{k-1})$ and $(I_1, \dots, I_{k-1})$ given by Model Sk, $\tau_0 = 0$, $\tau_k = \tau_e$ and $I_k = J$.

7.6 Example: rates of convergence for 4-level models

The results developed in Sections 7.4 and 7.5 allow the analysis of hierarchical models with an arbitrary number of levels. For example we could consider 4-level extensions of Model S3.

Model S4 (Symmetric 4-levels hierarchical model)**.** *Suppose*

$$y_{ijk\ell} = \mu + a_i + b_{ij} + c_{ijk} + \epsilon_{ijk\ell}, \tag{7.10}$$

where i, j, k and ℓ run from 1 to I, J, K and L respectively and $\epsilon_{ijk\ell}$ are iid normal random variables with mean 0 and variance σ_e^2. We employ a standard Bayesian model specification with $a_i \sim N(0, \sigma_a^2)$, $b_{ij} \sim N(0, \sigma_b^2)$, $c_{ijk} \sim N(0, \sigma_c^2)$ and a flat prior on μ.

In order to fit Model S4 with a Gibbs Sampler like GS($\boldsymbol{\beta}_T$), one could consider centering or non-centering each of the three levels $(a_i)_i$, $(b_{ij})_{ij}$ and $(c_{ijk})_{ijk}$. Let $(\lambda_1, \lambda_2, \lambda_3) \in \{0,1\}^3$ be the non-centering indicators associated to the resulting in $8 = 2^3$ combinations. Here $\lambda_d = 1$ indicates that the d-th level is non-centred while $\lambda_d = 0$ indicates that it is centred. The corresponding rates of convergence $\rho_{(\lambda_1,\lambda_2,\lambda_3)}$ can then be expressed in terms of the following normalized variance ratios

$$r_{i,j} = \frac{\tilde{\sigma}_i^2}{\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2} \qquad\qquad i,j \in \{1,2,3,4\}\,,$$

where $\tilde{\sigma}_1^2 = \frac{\sigma_a^2}{I}$, $\tilde{\sigma}_2^2 = \frac{\sigma_b^2}{IJ}$, $\tilde{\sigma}_3^2 = \frac{\sigma_c^2}{IJK}$ and $\tilde{\sigma}_4^2 = \frac{\sigma_e^2}{IJKL}$. If $\lambda_1 = 1$ (i.e. using the non-centred parametrization $(a_i)_i$ at level 1) the rates are

$$\rho_{111} = \max\{r_{1,4}, r_{2,4}, r_{3,4}\}\,, \qquad \rho_{110} = \max\{r_{1,3}, r_{2,3}, r_{4,3}\}\,,$$
$$\rho_{100} = \max\{r_{1,2}, 1 - r_{2,3}r_{3,4}\}\,, \qquad \rho_{101} = \max\{r_{1,2}, 1 - r_{2,4}r_{4,3}\}\,.$$

When $\lambda_1 = 0$ the expressions for the convergence rates are still explicit, but slightly more involved and are reported in Section 3.1 of the supplementary material. These rates can be derived from Corollary 5 and Theorem 10. It is worth noting that also in this 4-level case the skeleton chain $\delta^{(0)}\boldsymbol{\beta}$ is always the slowest chain for all centred and non-centred parametrizations (which can be checked by computing the rates of convergence of $\delta^{(1)}\boldsymbol{\beta}$, $\delta^{(2)}\boldsymbol{\beta}$ and $\delta^{(3)}\boldsymbol{\beta}$ using Theorem 10 and comparing those to the ones of $\delta^{(0)}\boldsymbol{\beta}$), even if for the general k-level case we were able to prove this fact only for the fully-centred parametrization (Theorem 11). The expressions given here can be easily used to derive conditionally optimal parametrizations for Model S4 given the rescaled variance components $(\tilde{\sigma}_i^2)_{i=1}^4$. For example, choosing whether to center or not each level by comparing the level-specific rescaled variances with the sum of the rescaled variances of the lower levels like in Section 3.2 leads to rates of convergence upper bounded by $\frac{3}{4}$.

8 Conclusions and future work

In this work we studied the convergence properties of the Gibbs Sampler algorithm in the context of Gaussian multilevel models. To do so we developed a novel analytic approach based on multigrid decompositions that allows the factorization of the Markov chain of interest into independent and easier to analyse sub-chains. This decomposition enables us to evaluate explicitly the L^2-rate of convergence in various models of interest. The results offer a detailed insight into the interaction between multilevel structures (e.g. nested and crossed) and the Gibbs Sampler and provide guidance on the choice of the computationally optimal parametrizations or linear constraints, which can potentially be relevant also beyond the Gaussian case (see e.g. Section 5), and indication of which parameters to monitor in the convergence diagnostic process (see Theorem 2 and discussion at the end of Section 2.1). Since the first preprint version of this paper, the multigrid decomposition developed in this paper has already found other practical applications. In particular Papaspiliopoulos et al. (2019) have successfully exploited it to analyse the computational complexity of the Gibbs Sampler in the context of crossed random effect models (see also Gao and Owen, 2017) and to design an algorithmic modification with linear computational complexity.

Together with explicit formulas for L^2-rates of convergences, the multigrid decomposition we developed in this paper provides a simple and intuitive theoretical characterizations of practical behaviors commonly observed in practice when fitting hierarchical models with MCMC, such as slower mixing for hyper-parameters at higher levels (see Theorems 2 and 11), algorithmic scalability with width of the hierarchy but not with height (e.g. Theorem 3 and Corollary 7) and good performances of centred parametrization in data-rich contexts (Theorem 8). The results presented in this paper provide a first step towards providing quantitative understanding of the behavior of MCMC algorithms (even beyond the Gibbs Sampler) in the extremely popular context of Bayesian hierarchical and multilevel models.

The present work could be extended in many directions. For example, it would be interesting to extend the results for non-symmetric cases, either by generalizing the bounds of Theorem 8 or by weakening the symmetry assumption in (S). In terms of classes of models considered, a natural and important extension would be to consider the multivariate case (where each parameter γ_t is a multivariate random vector) and the regression case. We expect many results developed in this work to extend to the multivariate and regression case, even if in that context the role played by non-symmetric cases will be more crucial. Another important class of models that would be worth approaching with methodologies analogous to the ones developed here are models based on Gaussian processes commonly used, for example, in spatial statistics (see e.g. Bass and Sahu, 2019).

An important and ambitious aim would be to extend the results to other tractable distributions within the exponential family beyond the Gaussian case. A starting point for this could be to analyse Model CkP as mentioned in Remark 2. Also, many non-Gaussian hierarchical models can be well-approximated by Gaussian ones for sufficiently large data sets, so that it is reasonable to conjecture that the qualitative conclusions (at least) of our study might remain valid when extrapolated to non-Gaussian models, rather like the analysis given in Sahu and Roberts (1999). A detailed study of this question is left for future work.

We have concentrated in this paper on deterministic samplers. However, explicit rates of convergence of random scan samplers are also available in the Gaussian case as described in Amit (1996) and extended in Roberts and Sahu (1997). Deterministic and random scan samplers can sometimes differ substantially in their convergence properties, see for example Roberts and Rosenthal (2015), although no general theory for this phenomenon is well-understood, so that the insights of this work could be particularly useful in this direction. Also, in the random scan case the reversibility of the induced Markov chains would allow us to apply the Cauchy interlacing theorem under weaker assumptions than Theorem 11 and thus prove orderings results for general hierarchical parametrizations $\boldsymbol{\beta}_T = (\beta_t)_{t \in T}$.

While this work is focused on L^2-rates of convergence, the same approach could be used to derive bounds on the distance (e.g. total variation or Wasserstein) between the distribution of the Markov chain at a given iteration and the target distribution (see e.g. Amit, 1996, Roberts and Sahu, 1997, (15) and Khare et al., 2009, Section 4.4). Such a formulation would be interesting to extend the recent growth in literature on providing rigorous characterizations of the computational complexity of Bayesian hierarchical linear models, see for example Rajaratnam and Sparks (2015); Roberts and Rosenthal (2016); Johndrow et al. (2015). In order to provide full characterizations, however, the case of unknown variances should be considered (see e.g. Jones and Hobert, 2004 for the two level case).

Supplementary Material

Proofs and additional material on the simulations (DOI: 10.1214/20-BA1242SUPP; .pdf). The supplementary material contains the proofs of the theoretical results presented in the paper, as well as additional formulas and material related to the simulations.

References

Alsmeyer, G. and Fuh, C.-D. (2001). "Limit theorems for iterated random functions by regenerative methods." *Stochastic processes and their applications*, 96(1): 123–142. MR1856683. doi: https://doi.org/10.1016/S0304-4149(01)00104-1. 1329

Amit, Y. (1991). "On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions." *J. Multivariate Anal.*, 38(1): 82–99. MR1128938. doi: https://doi.org/10.1016/0047-259X(91)90033-X. 1310

Amit, Y. (1996). "Convergence properties of the Gibbs sampler for perturbations of Gaussians." *Ann. Statist.*, 24(1): 122–140. MR1389883. doi: https://doi.org/10.1214/aos/1033066202. 1310, 1319, 1340

Bass, M. R. and Sahu, S. K. (2017). "A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC." *Statistics and Computing*, 27(6): 1491–1512. MR3687322. doi: https://doi.org/10.1007/s11222-016-9700-z. 1310

Bass, M. R. and Sahu, S. K. (2019). "Dynamically Updated Spatially Varying Parameterizations of Hierarchical Bayesian Models for Spatial Data." *Journal of Computational and Graphical Statistics*, 28(1): 105–116. MR3939375. doi: https://doi.org/10.1080/10618600.2018.1482761. 1340

Bhatia, R. (2013). *Matrix analysis*, volume 169. Springer Science & Business Media. MR1477662. doi: https://doi.org/10.1007/978-1-4612-0653-8. 1337

Brown, L. D., Mukherjee, G., Weinstein, A., et al. (2018). "Empirical Bayes estimates for a two-way cross-classified model." *The Annals of Statistics*, 46(4): 1693–1720. MR3819114. doi: https://doi.org/10.1214/17-AOS1599. 1324

Browne, W. J. (2004). "An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models." *Multilevel modelling newsletter*, 16(1): 13–25. 1325

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic programming language." *Journal of statistical software*, 76(1). 1311, 1328

Diaconis, P. and Freedman, D. (1999). "Iterated random functions." *SIAM review*, 41(1): 45–76. MR1669737. doi: https://doi.org/10.1137/S0036144598338446. 1329

Diaconis, P., Khare, K., and Saloff-Coste, L. (2010). "Stochastic alternating projections." *Illinois J. Math.*, 54(3): 963–979. MR2928343. 1310

Gao, K. and Owen, A. (2017). "Efficient moment calculations for variance components in large unbalanced crossed random effects models." *Electronic Journal of Statistics*, 11(1): 1235–1296. MR3635913. doi: https://doi.org/10.1214/17-EJS1236. 1310, 1324, 1339

Gelfand, A. E. and Sahu, S. K. (1999). "Identifiability, improper priors, and Gibbs sampling for generalized linear models." *Journal of the American Statistical Association*, 94(445): 247–253. MR1689229. doi: https://doi.org/10.2307/2669699. 1309, 1326

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). "Efficient parametrisations for normal linear mixed models." *Biometrika*, 82(3): 479–488. MR1366275. doi: https://doi.org/10.1093/biomet/82.3.479. 1309, 1319, 1331

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). "Efficient parametrizations for generalized linear mixed models (with discussion)." MR1425405. 1325

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American statistical association*, 85(410): 398–409. MR1141740. 1309

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. 1310

Goodman, J. and Sokal, A. D. (1989). "Multigrid Monte Carlo method. Conceptual foundations." *Physical Review D*, 40(6): 2035. 1311

Hills, S. E. and Smith, A. F. (1992). "Parameterization issues in Bayesian inference." *Bayesian statistics*, 4: 227–246. MR1380279. 1309

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15(1): 1593–1623. MR3214779. 1328

Johndrow, J. E., Mattingly, J. C., Mukherjee, S., and Dunson, D. (2015). "Approximations of Markov Chains and Bayesian Inference." *arXiv preprint arXiv:1508.03387*. 1341

Jones, G. L. and Hobert, J. P. (2004). "Sufficient burn-in for Gibbs samplers for a hierarchical random effects model." *The Annals of Statistics*, 32(2): 784–817. MR2060178. doi: https://doi.org/10.1214/009053604000000184. 1309, 1341

Kaufman, C. G., Sain, S. R., et al. (2010). "Bayesian functional ANOVA modeling using Gaussian process prior distributions." *Bayesian Analysis*, 5(1): 123–149. MR2596438. doi: https://doi.org/10.1214/10-BA505. 1326

Khare, K., Zhou, H., et al. (2009). "Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions." *The Annals of Applied Probability*, 19(2): 737–777. MR2521887. doi: https://doi.org/10.1214/08-AAP562. 1319, 1340

Liu, J. S. and Sabatti, C. (2000). "Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation." *Biometrika*, 353–369. MR1782484. doi: https://doi.org/10.1093/biomet/87.2.353. 1311

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. 1323

Meng, X.-L. and Van Dyk, D. (1997). "The EM Algorithm – An Old Folk-song Sung to a Fast New Tune." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3): 511–567. MR1452025. doi: https://doi.org/10.1111/1467-9868.00082. 1309

Meng, X.-L. and Van Dyk, D. A. (1999). "Seeking efficient data augmentation schemes via conditional and marginal augmentation." *Biometrika*, 86(2): 301–320. MR1705351. doi: https://doi.org/10.1093/biomet/86.2.301. 1323

Neal, R. M., et al. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov chain Monte Carlo*, 2(11): 2. MR2858447. 1312, 1328

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2003). "Non-centered parameterizations for hierarchical models and data augmentation (with discussion)." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 7*, 307–326. New York: Oxford University Press. MR2003180. 1309, 1322, 1330, 1331, 1333

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 59–73. MR2408661. doi: https://doi.org/10.1214/088342307000000014. 1309, 1310

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2019). "Scalable inference for crossed random effects models." *Biometrika*, 107(1): 25–40. MR4064138. doi: https://doi.org/10.1093/biomet/asz058. 1310, 1323, 1324, 1339

Papaspiliopoulos, O. and Zanella, G. (2017). "A note on MCMC for nested multilevel regression models via belief propagation." *arXiv preprint arXiv:1704.06064*. 1323

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 1328

Rajaratnam, B. and Sparks, D. (2015). "MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains." *arXiv preprint arXiv:1508.00947*. 1341

Roberts, G. O. and Rosenthal, J. S. (2015). "Surprising convergence properties of some simple Gibbs samplers under various scans." *International Journal of Statistics and Probability*, 5(1): 51–60. 1340

Roberts, G. O. and Rosenthal, J. S. (2016). "Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits." *Journal of Applied Probability*, 53(02): 410–420. MR3514287. doi: https://doi.org/10.1017/jpr.2016.9. 1341

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. 1309, 1310, 1319, 1331, 1337, 1338, 1340

Sahu, S. K. and Roberts, G. O. (1999). "On convergence of the EM algorithm and the Gibbs sampler." *Statistics and Computing*, 9(1): 55–64. 1340

Smith, A. F. and Roberts, G. O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–23. MR1210421. 1309

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). "BASiCS: Bayesian analysis of single-cell sequencing data." *PLoS computational biology*, 11(6): e1004333. 1326

Vines, S., Gilks, W., and Wild, P. (1996). "Fitting Bayesian multiple random effects models." *Statistics and Computing*, 6(4): 337–346. 1309, 1326

Whittaker, J. (1990). "Graphical models in applied multivariate analysis." MR1112133. 1338

Xie, Y. and Carlin, B. P. (2006). "Measures of Bayesian learning and identifiability in hierarchical models." *Journal of Statistical Planning and Inference*, 136(10): 3458–3477. MR2256284. doi: https://doi.org/10.1016/j.jspi.2005.04.003. 1309, 1326

Yu, Y. and Meng, X.-L. (2011). "To center or not to center: That is not the question – An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency." *Journal of Computational and Graphical Statistics*, 20(3): 531–570. MR2878987. doi: https://doi.org/10.1198/jcgs.2011.203main. 1310

Zanella, G. and Roberts, G. (2021). "Supplementary Material of "Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1242SUPP. 1313

Acknowledgments

The authors are grateful for stimulating discussions with Omiros Papaspiliopoulos and Art Owen. GZ was supported in part by an EPSRC Doctoral Prize fellowship, by the European Research Council (ERC) through StG "N-BNP" 306406 and by MIUR through the PRIN Project 2015SNS29B. GOR acknowledges support from EPSRC through grants EP/K014463/1 (i-Like), EP/D002060/1 (CRiSM), EP/R034710/1 (CoSInES), and EP/R018561/1 (Bayes for Health).

Invited Discussion

Quan Zhou[*] and Shuang Zhou[†]

We congratulate Zanella and Roberts (2021) on their seminal work on the convergence properties of blocked Gibbs samplers for Gaussian hierarchical models. The article develops an ingenious multigrid decomposition technique which can decompose the Gibbs sampler for any symmetric k-level hierarchical model into k independent Markov chains, providing profound insights and theoretical guidance on how to choose the parameterization and updating strategy in Gibbs sampling. Further, when the variance parameters are known, this method can be used to quickly find the closed-form expression for the sampler's convergence rate. In this note, we focus on the multigrid decomposition theory for nested hierarchical models and refer readers to Papaspiliopoulos et al. (2020, 2021) for more results on the multigrid decomposition for crossed effect models. We first consider the theory developed in Section 7 of Zanella and Roberts (2021) from a linear algebraic perspective, which offers an alternative and potentially more general approach to proving a key result in their paper (see Remark 2). Next, we perform a numerical experiment which shows that the insights obtained in Zanella and Roberts (2021) may also be applied to hierarchical linear mixed models.

1 Multigrid decomposition and block diagonalization of B-matrices

1.1 Preliminaries

We start by reviewing Theorem 1 of Roberts and Sahu (1997), which characterizes the convergence rate of a blocked Gibbs sampler targeting a multivariate normal distribution. Consider a parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_{k-1}) \in \mathbb{R}^m$, where $m = \sum_{d=0}^{k-1} m_d$ and $\boldsymbol{\beta}_d \in \mathbb{R}^{m_d}$ denotes the d-th component block of $\boldsymbol{\beta}$. Suppose $\boldsymbol{\beta}$ follows a non-degenerate normal distribution $N(\boldsymbol{u}, \boldsymbol{\Sigma})$, and let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. We partition $\boldsymbol{u}$ and $\boldsymbol{Q}$ by $\boldsymbol{u} = (\boldsymbol{u}_0, \dots, \boldsymbol{u}_{k-1})$ and $\boldsymbol{Q} = (\boldsymbol{Q}_{d,d'})_{0 \le d,d' \le k-1}$, where $\boldsymbol{u}_d \in \mathbb{R}^{m_d}$ and $\boldsymbol{Q}_{d,d'} \in \mathbb{R}^{m_d \times m_{d'}}$.[1] The full conditional distribution of $\boldsymbol{\beta}_d$ is

$$\boldsymbol{\beta}_d \mid \boldsymbol{\beta}_{-d} \sim N\big((\boldsymbol{A}\boldsymbol{\beta})_d + ((\boldsymbol{I}_m - \boldsymbol{A})\boldsymbol{u})_d,\ \boldsymbol{Q}_{d,d}^{-1}\big), \tag{1}$$

where the "A-matrix" is given by $\boldsymbol{A} = \boldsymbol{I}_m - \mathrm{diag}(\boldsymbol{Q}_{0,0}^{-1}, \dots, \boldsymbol{Q}_{k-1,k-1}^{-1})\boldsymbol{Q}$. Note that all diagonal blocks of $\boldsymbol{A}$ are zero matrices. Write $\boldsymbol{A} = \mathbb{L}(\boldsymbol{A}) + \mathbb{U}(\boldsymbol{A})$, where $\mathbb{L}(\boldsymbol{A})$ (resp. $\mathbb{U}(\boldsymbol{A})$) contains the block lower (resp. upper) triangular part of $\boldsymbol{A}$. Define the "B-matrix" by

$$\boldsymbol{B} = (\boldsymbol{I}_m - \mathbb{L}(\boldsymbol{A}))^{-1}\mathbb{U}(\boldsymbol{A}). \tag{2}$$

[*] Department of Statistics, Texas A&M University, quan@stat.tamu.edu

[†] School of Mathematical and Statistical Sciences, Arizona State University, szhou98@asu.edu

[1] We always use $\boldsymbol{\beta}_d$ (or $\boldsymbol{u}_d$) to denote a subvector and $\boldsymbol{Q}_{d,d'}$ to denote a submatrix. An entry of $\boldsymbol{\beta}$ is denoted by β_t, where β is not in bold. Note our $\boldsymbol{\beta}_d$ denotes the same quantity as $\boldsymbol{\beta}^{(d)}$ in Zanella and Roberts (2021, Section 7).

For the deterministic sweep Gibbs sampler (called DUGS in Roberts and Sahu (1997)) which updates $\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_{k-1}$ successively in each iteration, its dynamics can be described by

$$\boldsymbol{\beta}(s+1) \sim N(\boldsymbol{B}\boldsymbol{\beta}(s) + (\boldsymbol{I}_m - \boldsymbol{B})\boldsymbol{u}, \boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathrm{T}}), \tag{3}$$

where $\boldsymbol{\beta}(s)$ denotes the s-th sample of the vector $\boldsymbol{\beta}$. Roberts and Sahu (1997) proved that the rate of convergence (as defined in Equation (1) of Zanella and Roberts (2021)) of DUGS is given by $\rho(\boldsymbol{B})$, the spectral radius of $\boldsymbol{B}$. The following observation is almost immediate.

Lemma 1. *Consider the DUGS sampler given in (3). Let $\boldsymbol{\Delta}$ be an $n \times m$ matrix with rank m, and $\boldsymbol{\Delta}^-$ be its left inverse such that $\boldsymbol{\Delta}^-\boldsymbol{\Delta} = \boldsymbol{I}_m$. Then, the induced Markov chain $((\boldsymbol{\Delta}\boldsymbol{\beta})(s))_{s\geq 1}$ has the same rate of convergence as $(\boldsymbol{\beta}(s))_{s\geq 1}$.*

Proof. Observe that the B-matrix for $((\boldsymbol{\Delta}\boldsymbol{\beta})(s))_{s\geq 1}$ is given by $\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}^-$, which has the same spectral radius as $\boldsymbol{\Delta}^-\boldsymbol{\Delta}\boldsymbol{B} = \boldsymbol{B}$. □

1.2 Multigrid decomposition

Consider a collection of matrices $\{\boldsymbol{\Delta}^{(p,d)}\}_{0\leq p,d\leq k-1}$, where $\boldsymbol{\Delta}^{(p,d)} \in \mathbb{R}^{m_p \times m_d}$, such that $\boldsymbol{\beta}_d \mapsto ((\boldsymbol{\Delta}^{(0,d)}\boldsymbol{\beta}_d)^{\mathrm{T}}, \dots, (\boldsymbol{\Delta}^{(k-1,d)}\boldsymbol{\beta}_d)^{\mathrm{T}})^{\mathrm{T}}$ is an injective linear transformation from $\mathbb{R}^{m_d}$ to $\mathbb{R}^m$. Then, let $\boldsymbol{\Delta}$ be the $km \times m$ matrix defined by

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}^{(0)} \\ \vdots \\ \boldsymbol{\Delta}^{(k-1)} \end{bmatrix}, \quad \text{where } \boldsymbol{\Delta}^{(p)} = \operatorname{diag}(\boldsymbol{\Delta}^{(p,0)}, \dots, \boldsymbol{\Delta}^{(p,k-1)}) \in \mathbb{R}^{km_p \times m}. \tag{4}$$

Remark 1. The multigrid decomposition for hierarchical models given in Zanella and Roberts (2021, Section 7) corresponds to defining $\boldsymbol{\Delta}^{(p,d)}$ such that $\delta^{(p)}\boldsymbol{\beta}_d = \boldsymbol{\Delta}^{(p,d)}\boldsymbol{\beta}_d$ where $\delta^{(p)}$ is the residual operator defined in Section 7.4 therein (by their Equations (7.6) and (7.7), one can see that entries of $\boldsymbol{\Delta}^{(p,d)}$ are determined by the transition probabilities of the auxiliary random walk introduced in Section 7.3). In this construction, $\boldsymbol{\Delta}^{(p,d)} = \boldsymbol{0}$ if $p > d$. Lemma 4 of Zanella and Roberts (2021) proves that, under certain conditions on the precision matrix $\boldsymbol{Q}$, $\boldsymbol{\Delta}^{(p,d)}(\boldsymbol{A}\boldsymbol{\beta})_d$ only depends on $\boldsymbol{\Delta}^{(p)}\boldsymbol{\beta}$, which implies that $\boldsymbol{\Delta}\boldsymbol{A}\boldsymbol{\Delta}^-$ is block diagonal. We prove below that this further implies $\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}^-$ is also block diagonal.

Lemma 2. *Let $\boldsymbol{\Delta}$ be as defined by (4), and $\boldsymbol{B}$ be as given in (2). Suppose that $\boldsymbol{\Delta}\boldsymbol{A}\boldsymbol{\Delta}^- = \operatorname{diag}(\tilde{\boldsymbol{A}}^{(0)}, \dots, \tilde{\boldsymbol{A}}^{(k-1)})$, where $\tilde{\boldsymbol{A}}^{(d)}$ has dimension $km_d \times km_d$. Then, $\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}^-$ is also block diagonal, and the d-th diagonal block is $(\boldsymbol{I}_{km_d} - \mathbb{L}(\tilde{\boldsymbol{A}}^{(d)}))^{-1}\mathbb{U}(\tilde{\boldsymbol{A}}^{(d)})$.*

Proof. We first set some notation to describe the block structure of $\boldsymbol{\Delta}$. We treat each $\boldsymbol{\Delta}^{(p)}$ as a block matrix with $k \times k$ components, and $\boldsymbol{\Delta}$ as a block matrix with $k^2 \times k$ components. For each $i \in \{0, 1, \dots, k^2-1\}$, define $v(i) = \lfloor i/k \rfloor$ and $z(i) = i - kv(i)$. The component block $\boldsymbol{\Delta}_{i,j}$ has dimension $m_{v(i)} \times m_j$, and by construction $\boldsymbol{\Delta}_{i,j} = \boldsymbol{0}$ unless $j = z(i)$ (in which case $\boldsymbol{\Delta}_{i,j} = \boldsymbol{\Delta}^{(v(i),z(i))}$). The left inverse $\boldsymbol{\Delta}^-$ has the same "block structure" as $\mathbf{\Delta}^{\mathrm{T}}$, and thus we can use an analogous notation to denote its component blocks. Treating $\mathbf{\Delta A \Delta}^{-}$ as a block matrix with $k^2 \times k^2$ components, we find that

$$(\mathbf{\Delta A \Delta}^{-})_{i,j} = \sum_{0 \le d,p \le k-1} \mathbf{\Delta}_{i,d} \boldsymbol{A}_{d,p} \mathbf{\Delta}^{-}_{p,j} = \mathbf{\Delta}_{i,z(i)} \boldsymbol{A}_{z(i),z(j)} \mathbf{\Delta}^{-}_{z(j),j}.$$

The assumption on $\mathbf{\Delta A \Delta}^{-}$ means that $(\mathbf{\Delta A \Delta}^{-})_{i,j} = \mathbf{0}$ whenever $v(i) \neq v(j)$. Hence, by the definition of the function z, if $(\mathbf{\Delta A \Delta}^{-})_{i,j}$ is nonzero and $i < j$, then $z(i) < z(j)$. In other words, the upper triangular blocks of $\boldsymbol{A}$ are only involved in the calculation of the upper triangular blocks of $\mathbf{\Delta A \Delta}^{-}$, which leads to

$$\mathbb{L}(\mathbf{\Delta A \Delta}^{-}) = \mathbf{\Delta}\mathbb{L}(\boldsymbol{A})\mathbf{\Delta}^{-} = \mathrm{diag}(\mathbb{L}(\tilde{\boldsymbol{A}}^{(0)}), \dots, \mathbb{L}(\tilde{\boldsymbol{A}}^{(k-1)})).$$

By writing $\mathbf{\Delta}\mathbb{L}(\boldsymbol{A})^2\mathbf{\Delta}^{-} = (\mathbf{\Delta}\mathbb{L}(\boldsymbol{A})\mathbf{\Delta}^{-})^2$, we see that analogous identities hold for $\mathbb{U}(\boldsymbol{A})^n$ and $\mathbb{L}(\boldsymbol{A})^n$ for any integer $n \ge 0$. Apply Neumann series to get $\boldsymbol{B} = (\boldsymbol{I} - \mathbb{L}(\boldsymbol{A}))^{-1}\mathbb{U}(\boldsymbol{A}) = \sum_{n=0}^{\infty} \mathbb{L}(\boldsymbol{A})^n \mathbb{U}(\boldsymbol{A})$. It then follows that

$$\begin{aligned}\mathbf{\Delta B \Delta}^{-} &= \sum_{n=0}^{\infty} (\mathbf{\Delta}\mathbb{L}(\boldsymbol{A})^n\mathbf{\Delta}^{-})(\mathbf{\Delta}\mathbb{U}(\boldsymbol{A})\mathbf{\Delta}^{-}) \\ &= \sum_{n=0}^{\infty} \mathrm{diag}\left\{\mathbb{L}(\tilde{\boldsymbol{A}}^{(0)})^n\mathbb{U}(\tilde{\boldsymbol{A}}^{(0)}), \dots, \mathbb{L}(\tilde{\boldsymbol{A}}^{(k-1)})^n\mathbb{U}(\tilde{\boldsymbol{A}}^{(k-1)})\right\}.\end{aligned}$$

Apply Neumann series again to conclude the proof. □

Remark 2. In our proof of Lemma 2, the exact definition of $\{\mathbf{\Delta}^{(p,d)}\}_{p,d}$ is irrelevant. If $\mathbf{\Delta}$ is indeed constructed by the multigrid decomposition scheme described in Remark 1, the conclusion of Theorem 9 of Zanella and Roberts (2021) implies that $\mathbf{\Delta B \Delta}^{-}$ must be block diagonal since $((\mathbf{\Delta}^{(0)}\boldsymbol{\beta})(s))_{s\ge1}, \dots, ((\mathbf{\Delta}^{(k-1)}\boldsymbol{\beta})(s))_{s\ge1}$ are independent Markov chains. But by Lemma 1, the independence among these chains is unnecessary for the purpose of evaluating $\rho(\boldsymbol{B})$: all we need is some $\mathbf{\Delta}$ such that the eigenvalues of $\mathbf{\Delta B \Delta}^{-}$ are easy to evaluate. This observation motivates us to prove Lemma 2 by only assuming the block diagonal structure of $\mathbf{\Delta A \Delta}^{-}$.

The proof of Theorem 10 of Zanella and Roberts (2021) reveals that, under their assumptions, the spectrum of the matrix $\tilde{\boldsymbol{A}}^{(p)}$ in Lemma 2 is determined by some matrix $\boldsymbol{C}^{(p)}$ with dimension $(k-p) \times (k-p)$. For symmetric hierarchical models, the entries of $\boldsymbol{C}^{(p)}$ can be easily determined from the data likelihood, and one can find $\rho(\boldsymbol{B})$ by evaluating the eigenvalues of $(\boldsymbol{I} - \mathbb{L}(\boldsymbol{C}^{(p)}))^{-1}\mathbb{U}(\boldsymbol{C}^{(p)})$ for each p. Further, by Theorem 11 of Zanella and Roberts (2021), if the centred parameterization is used, we have $\rho(\boldsymbol{B}) = \rho\{(\boldsymbol{I} - \mathbb{L}(\boldsymbol{C}^{(0)}))^{-1}\mathbb{U}(\boldsymbol{C}^{(0)})\}$, which means that we only need to find the eigenvalues of one $k \times k$ matrix.

1.3 Extensions

The above analysis implies that the multigrid decomposition can also be used to study the rate of convergence of other blocked Gibbs samplers considered in Roberts and Sahu (1997), e.g. REGS (reversible version of DUGS) and RSGS (random sweep Gibbs sampler). By Theorem 2 of Roberts and Sahu (1997), the convergence rate of RSGS depends on the largest eigenvalue of $\boldsymbol{A}$, denoted by $\lambda(\boldsymbol{A})$, which can be found by calculating the eigenvalues of $\boldsymbol{C}^{(p)}$. For example, for Model S3 with centered parameterization, it can be shown that $\lambda(\boldsymbol{A})^2 = \rho(\boldsymbol{B})$; see also Theorem 5 of Roberts and Sahu (1997). For REGS, the rate of convergence is given by $\rho((\boldsymbol{I}-\mathbb{L}(\boldsymbol{A}))^{-1}\mathbb{U}(\boldsymbol{A})(\boldsymbol{I}-\mathbb{U}(\boldsymbol{A}))^{-1}\mathbb{L}(\boldsymbol{A}))$, which again can be found by replacing $\boldsymbol{A}$ with $\boldsymbol{C}^{(p)}$. For Model S3 with centered parameterization, a closed-form formula for the convergence rate of REGS can be obtained (e.g. by using `Mathematica`), but we find the expression too unwieldy to be included here.

Another potential direction for future research is to study how to construct the matrix $\boldsymbol{\Delta}$ under more general settings. Zanella and Roberts (2021) essentially impose two conditions on the precision matrix $\boldsymbol{Q}$, both being very natural for Bayesian hierarchical models. First, the conditional independence of all coordinates of $\boldsymbol{\beta}$ can be described by a tree (see Section 7.1). Second, $\boldsymbol{Q}$ satisfies a “symmetric” condition, which makes the evaluation of the matrix $\boldsymbol{C}^{(p)}$ convenient (see Section 7.3). It might be interesting to consider whether the first one can be replaced by assuming $\boldsymbol{Q}$ factorizes over a decomposable graph (and then the second condition needs to be modified accordingly).

2 A numerical study on linear mixed models

In Section 3 of Zanella and Roberts (2021), the authors apply their general theory to a three-level nested model and derive the exact convergence rate of the Gibbs sampler under four different parameterizations, which can be seen as a theoretical justification of the rules commonly used in the parameterization of hierarchical models. Consider their Corollary 2. Borrowing the terminology of Bernardo et al. (2003, Section 2.1), we may think of $(\tilde{\sigma}_b^2+\tilde{\sigma}_e^2)^{-1}$ as the observed information and $1/\tilde{\sigma}_b^2$ as the augmented information for each γ_i (in a two-level model with global mean γ_i). Similarly, at the middle level, $(\tilde{\sigma}_a^2+\tilde{\sigma}_b^2+\tilde{\sigma}_e^2)^{-1}$ is the observed information for μ and $1/\tilde{\sigma}_a^2$ is the augmented. Hence, Corollary 2 of Zanella and Roberts (2021) generalizes the existing results for two-level models: whether to choose a centered parameterization (CP) at a lower level depends on the ratio of observed and augmented information for the parameter. CP works well when the data are very informative about the parameter regardless of the dependence between the parameter and latent variables, while non-centered parameterization is preferred when the latent variables are poorly identified, in which case the strong correlation between the parameter and latent layer may cause slow convergence of Gibbs samplers (Papaspiliopoulos et al., 2007). To investigate whether such principles can be applied to more complex models, below we numerically study how the mixing behavior of the Gibbs sampler depends on parameterization in hierarchical linear mixed models.

For simplicity, we work with a two-level linear mixed model; more extensions can be found in Sahu (1994); Gelfand et al. (1995). Suppose we have

$$Y_{ij} = \mu + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \beta_i + \varepsilon_{ij}, \quad j = 1,\dots,J, \quad i = 1,\dots,I,$$

where μ denotes the global effect, $\boldsymbol{\alpha}$ is a $r \times 1$ coefficient vector, and $\{\beta_i : i = 1,\dots,I\}$ denotes the random effects at the lower level. Assume additive Gaussian errors $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim}$

	σ_a^2	σ_b^2	σ_e^2	CP	PNCP	NCP
Case I	0.1	0.1	10	0.909	0.923	0.157
Case II	10	0.1	0.1	0.763	0.977	0.976
Case III	0.1	10	0.1	0.019	0.969	0.999

Table 1: Rates of convergences of three Gibbs Sampler schemes for cases I, II, III.

$N(0, \sigma_e^2)$ for $i = 1, \dots, I$ and $j = 1, \dots, J$. Now define $\rho_i = \mu + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \beta_i$ and $\eta_i = \mu + \beta_i$. We consider three parameterizations.

- (CP) Centered parameterization with $(\mu, \boldsymbol{\alpha}, \rho)$:

$$Y_{ij} \mid \rho_i \sim N(\rho_i, \sigma_e^2), \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$
$$\rho_i \mid \mu, \boldsymbol{\alpha} \overset{\text{i.i.d.}}{\sim} N(\mu + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha}, \sigma_b^2), \quad \boldsymbol{\alpha} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma}_a), \quad p(\mu) \propto 1.$$

- (PNCP) Partially non-centered parameterization with $(\mu, \boldsymbol{\alpha}, \eta)$:

$$Y_{ij} = \eta_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$
$$\eta_i \mid \mu \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma_b^2), \quad \boldsymbol{\alpha} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma}_a), \quad p(\mu) \propto 1.$$

- (NCP) Non-centered parameterization with $(\mu, \boldsymbol{\alpha}, \beta)$:

$$Y_{ij} = \mu + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \beta_i + \varepsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$
$$\beta_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_b^2), \quad \boldsymbol{\alpha} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma}_a), \quad p(\mu) \propto 1.$$

We denote $\boldsymbol{\Sigma}_a = \mathrm{diag}(\sigma_a^2, \dots, \sigma_a^2)$ and let $\boldsymbol{X}$ be a design matrix with i-th row being $\boldsymbol{x}_i^{\mathrm{T}}$. For simplicity of exposition, we assume variance parameters $\sigma_a^2, \sigma_b^2, \sigma_e^2$ are known. In numerical studies, we let $r = 2$, $I = 5$ and $J = 10$. Set $\boldsymbol{x}_i^{\mathrm{T}} = (x_i, x_i^2)^{\mathrm{T}}$ with $x_i = (i-1)/4$, and the true parameters by $\mu^* = 0$, $\boldsymbol{\alpha}^* = (-0.5, 1)^{\mathrm{T}}$, and $\boldsymbol{\beta}^* = (0.2, -0.2, 0.1, -0.1, 0)^{\mathrm{T}}$. We consider three settings of the variance parameters and numerically calculate $\rho(\boldsymbol{B})$ for the corresponding B-matrices; see Table 1. For each case we run $10,000$ iterations of Gibbs samplers, discarding the first $5,000$ as burn-in. To gauge the mixing behavior of Gibbs samplers under considered updating schemes, denote the globally averaged parameters with different parameterizations by $\overline{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha} = (1/I)\sum_{i=1}^{I} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha}$, $\bar{\rho} = (1/I)\sum_{i=1}^{I} \rho_i$, $\bar{\eta} = (1/I)\sum_{i=1}^{I} \eta_i$ and $\bar{\beta} = (1/I)\sum_{i=1}^{I} \beta_i$. The autocorrelation functions for the globally averaged parameters are displayed in Figures 1, 2, 3. The numerical results are consistent with the theoretical insights obtained in Theorem 3 and Corollary 2 of Zanella and Roberts (2021). Clearly, NCP yields the fastest rate of convergence in case I where $\sigma_a^2/I, \sigma_b^2/I$ are much smaller than $\sigma_e^2/(IJ)$, similarly to the findings in Theorem 3 of Zanella and Roberts (2021). On the contrary, CP provides the best mixing of Gibbs samplers when the data are more informative about the parameters; see case II and case III. In this numerical study, PNCP yields somewhat mediocre results for all cases, suggesting that this parameterization may not be an effective complement to CP and NCP. But a more delicate parameterization may be constructed such that it adapts to the amount of information carried with the data; see Papaspiliopoulos et al. (2007).

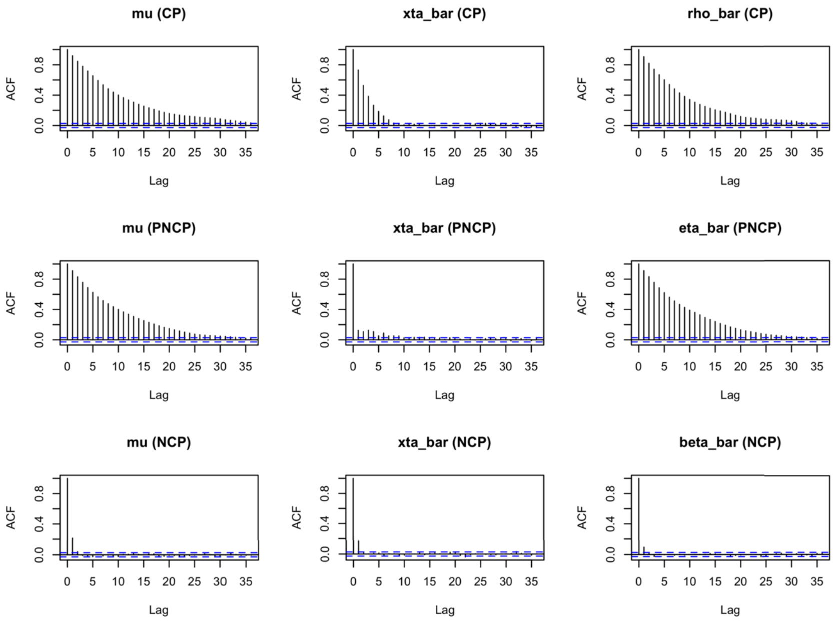


Figure 1: Autocorrelation functions for globally averaged parameters $\mu, \overline{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha}, \bar{\rho}, \bar{\eta}, \bar{\beta}$ for case I.

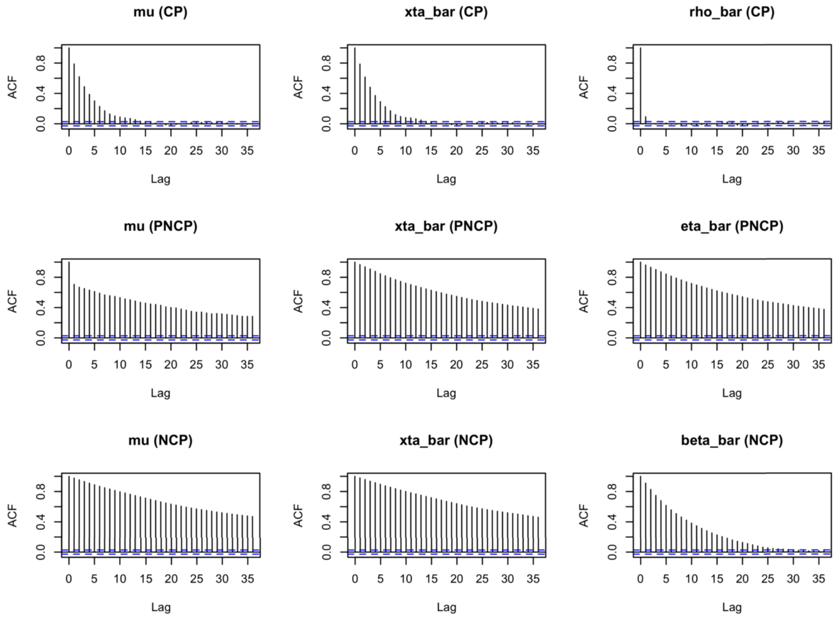


Figure 2: Autocorrelation functions for globally averaged parameters for case II.

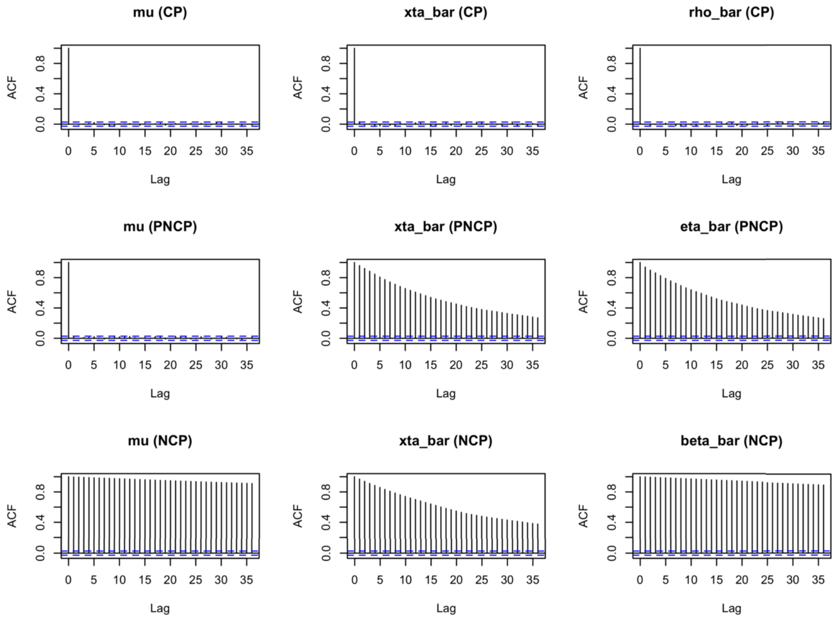


Figure 3: Autocorrelation functions for globally averaged parameters for case III.

References

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (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. 1347

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). "Efficient parametrisations for normal linear mixed models." *Biometrika*, 82(3): 479–488. MR1366275. doi: https://doi.org/10.1093/biomet/82.3.479. 1347

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 59–73. MR2408661. doi: https://doi.org/10.1214/088342307000000014. 1347, 1348

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. 1344

Papaspiliopoulos, O., Stumpf-Fétizon, T., and Zanella, G. (2021). "Scalable computation for Bayesian hierarchical models." *arXiv preprint arXiv:2103.10875*. 1344

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. 1344, 1345, 1346, 1347

Sahu, S. K. (1994). *Strategies for efficient implementation of MCMC algorithms*. University of Connecticut. MR2690821. 1347

Zanella, G. and Roberts, G. (2021). "Multilevel linear models, Gibbs samplers and multigrid decompositions." *Bayesian Analysis*. 1344, 1345, 1346, 1347, 1348

Invited Discussion

James M. Flegal*

A key requirement in a successful Markov chain Monte Carlo (MCMC) simulation is finding a sampler that mixes well. Finding such a sampler is likely to be the most challenging part of the process. To this end, I congratulate Drs. Zanella and Roberts for developing a theoretically framework to analyze Gibbs samplers for a variety of widely used model structures, including nested and crossed random effects. Moreover, they provide explicit recommendations for improving practical implementations, which are theoretically justified and practically useful as illustrated via simulation.

1 Primary contribution

The practical success of a Bayesian analysis requires a Markov chain that converges quickly to obtain effective MCMC simulation results in a finite amount of time. This paper provides important theoretical and practical contributions for Gibbs samplers for various multilevel linear models.

Consider the symmetric three-level Gaussian linear model, denoted **Model S3**, defined as

$$y_{ijk} = \mu + a_i + b_{ij} + \epsilon_{ijk}, \tag{1}$$

where i, j, and k run from 1 to I, J, and K, respectively and ϵ_{ijk} are iid Normal random variables with mean 0 and variance σ_e^2. The standard Bayesian model specification is assumed with $a_i \sim \mathrm{N}(0, \sigma_a^2)$, $b_{ij} \sim \mathrm{N}(0, \sigma_b^2)$, and a flat prior on μ. Drs. Zanella and Roberts investigate convergence rates of different Gibbs samplers for Model S3. First they consider a fully non-centered parametrization using (μ, a, b), which is denoted GS$(1,1)$. Alternatively, a fully centered parametrization denoted GS$(0,0)$ is obtained by replacing a_i and b_{ij} with $\gamma_i = \mu + a_i$ and $\eta_{ij} = \gamma_i + b_{ij}$, respectively. Two mixed parameterizations using (μ, γ, b) and (μ, a, η) are also considered. Theorem 3 gives explicit formulas for the efficiency of the Gibbs samplers, which are functions of $(\sigma_a^2, \sigma_b^2, \sigma_e^2)$. Recommended parameterizations guarantee the L^2 rate of convergence is less than $2/3$ and hence a quickly mixing chain is available for any values of $(I, J, K, \sigma_a^2, \sigma_b^2, \sigma_e^2)$!

When the variances $(\sigma_a^2, \sigma_b^2, \sigma_e^2)$ are unknown, these results can be used to guide a simple and effective strategy ensuring high sampling efficiency. In short, an optimal parametrization can be utilized at each iteration based on the current variances. For example, if the value of $(\sigma_a^2, \sigma_b^2, \sigma_e^2)$ indicates the fully non-centered parametrization using (μ, a, b) is optimal, then the Gibbs sampler would update using the full conditionals $(\mu, a, b) | (\sigma_a^2, \sigma_b^2, \sigma_e^2)$. Such a strategy can be implemented in conjunction with parameter expansion methodology creating a bespoke sampling algorithm that speeds up the convergence even further.

*Department of Statistics, University of California, Riverside, jflegal@ucr.edu

2 Extensions and limitations

The rigorous theory developed applies more generally than Model S3 at (1). Specifically, the paper provides a multigrid decomposition of a k-level Gibbs sampler into k independent sub-chains, which describe the different levels of the hierarchy. Then the overall rate of convergence can be obtained by an analysis on the sub-chains. In many cases, the slowest sub-chain corresponds to the coarsest level of the hierarchy and dictates the overall rate of convergence. An interesting observation is that such a multigrid decomposition stems from the model under consideration, and not the chosen parameterization.

The multigrid decomposition can also be used to analyze a crossed effect model where they obtain convergence properties for the four possible parameterizations (when $k = 2$). Unlike the symmetric three-level Gaussian linear model, the convergence rate of the Gibbs sampler can go to 1 as the number of factors increases and hence the reparameterization techniques alone are not enough to ensure good convergence properties. Imposing identifiability improves convergence for this model, but the convergence rate can still deteriorate as the amount of data increases. Drs. Zanella and Roberts illustrate, via simulation, that insights gained from the Gaussian case can be applied successfully in more general settings.

Potential extensions include considering non-symmetric cases or multivariate cases, models with a linear mean part or based on Gaussian processes, and other tractable distributions beyond the Gaussian case. Another interesting direction is to consider convergence rates for random scan, as opposed to deterministic, Gibbs samplers. Any of these would be excellent future contributions to the literature on MCMC convergence rates. Unfortunately, multigrid decomposition techniques may not be appropriate in some of these problems since they factorize the Gibbs sampler into independent sub-chains implying a reparameterization with some level of posterior independence.

Finding optimal parameterizations in practice may be challenging and somewhat unrealistic as model complexity or data size increases. It may be preferable to explore reparameterizations via a random scan Gibbs sampler instead. Convergence rate problems also become more complicated when sampling from modern high-dimensional posteriors (Rajaratnam and Sparks, 2015; Qin and Hobert, 2019; Duan et al., 2018).

3 MCMC output analysis

Consider simulated data from Model S3 at (1) with $I = J = 100$, $K = 5$, $\mu = 0$, $\sigma_a = 2$, $\sigma_b = 1/2$, and $\sigma_e = 1$. The standard Bayesian model specification described previously is assumed. Under these settings, Theorem 3 shows the L^2 rates of convergence of the various Gibbs samplers are

$$(\rho_{00}, \rho_{11}, \rho_{01}, \rho_{10}) = (0.4448, 0.9995, 0.5558, 0.9994).$$

Hence the fully centered parametrization and fully non-centered parametrization represent the best and worst options, respectively. Based on 5000 iterations, Figure 1 shows

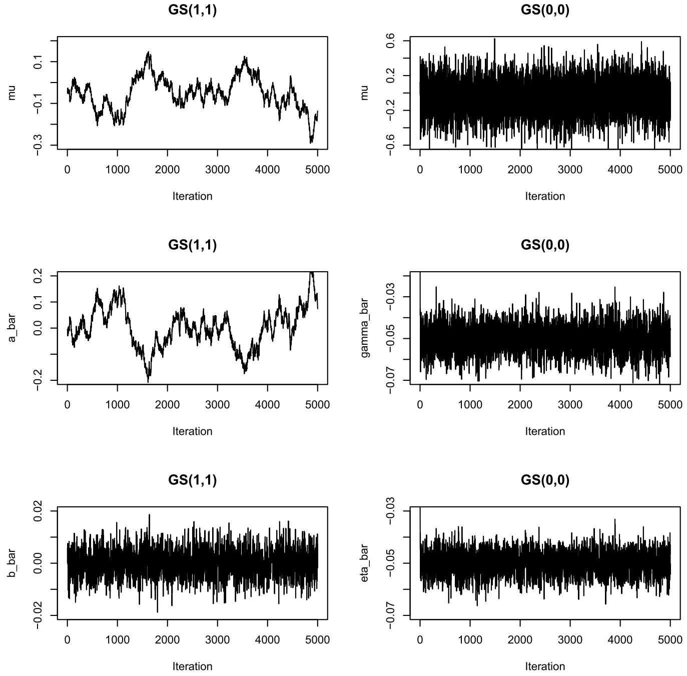


Figure 1: Trace plots of global averages based on 5000 iterations.

trace plots of the global averages $(\mu, a_{\cdot}, b_{\cdot\cdot})$ and $(\mu, \gamma_{\cdot}, \eta_{\cdot\cdot})$ for $\mathrm{GS}(1,1)$ and $\mathrm{GS}(0,0)$, respectively. Levels 0 and 1 are mixing quite poorly for the fully non-centered parametrization $\mathrm{GS}(1,1)$, while the mixing is excellent for all three levels of the fully centered parametrization $\mathrm{GS}(0,0)$.

Having identified $\mathrm{GS}(0,0)$ as the optimal sampler, the rest of my discussion identifies some additional practical issues. A primary challenge is the fact that there are $1 + I + IJ = 10101$ parameters. It is impossible to examine trace plots, as in Figure 1, for all parameters and unclear that looking only at the global means is sufficient. (Drs. Zanella and Roberts raise this issue as well.) In short, we want to determine a subset of parameters, from our initial 5000 iterations, that can be used to monitor behaviour of the chain.

One potential solution is to estimate the effective sample size (ESS) for each parameter, which can be calculated in a couple seconds using the `mcmcse` `R` package (Flegal et al., 2021). Figure 2 plots histograms of the ESS for the 100 level 1 parameters a_i and γ_i (top row) and the 10000 level 2 parameters b_{ij} and η_{ij} (bottom row). Red vertical lines in each histogram show the location of an ESS for the corresponding global mean. For a_i, b_{ij}, and η_{ij}, there is significant (anecdotal) evidence that monitoring the global means is sufficient since the ESSs are substantially smaller. For γ_i, the global mean ESS lies in the middle of the histogram and hence it is unclear if a practitioner can rely on only monitoring the global mean. Plots of this type could be more valuable in the

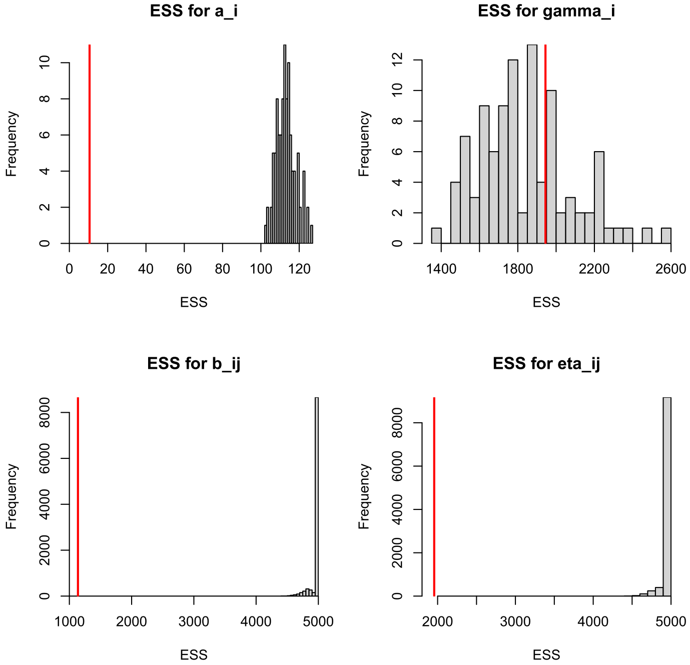

ESS for a_i
ESS for gamma_i
ESS for b_ij
ESS for eta_ij

Figure 2: Histograms of estimated ESSs based on 5000 iterations.

pressense of correlation between parameters within the Gibbs sampler blocks. It is also worth noting that reparameterizations will likely change ESSs.

Another common challenge for high-dimensional Markov chain simulations can be memory allocation. Roughly speaking a $r \times c$ double-precision matrix requires $rc8/10^9$ gigabytes of memory, so storing the initial 5000 iterations requires about 0.4 gigabytes of memory. Then using a long run to calculate standard errors to ensure reliable inference or terminating the simulation via sequential stopping rules (Vats et al., 2019) requires some creativity with regard to memory. As opposed to storing the entire chain, one option is to only store parameters of scientific interest and parameters necessary to monitor chain behaviour. Storage constraints can also be overcome by estimating the limiting covariance of the Monte Carlo estimators via recursive or low-cost batch means estimators (Chan and Yau, 2017; Zhu et al., 2021; Gong and Flegal, 2016).

I close by reiterating my congratulations to Drs. Zanella and Roberts for their interesting and notable paper, that hopefully will motivate future contributions in this area. I also want to thank Dr. Guindani for the opportunity to participate in this discussion.

References

Chan, K. W. and Yau, C. Y. (2017). "Automatic optimal batch size selection for recursive estimators of time-average covariance matrix." *Journal of the American Sta-*

tistical Association, 112: 1076–1089. MR3735361. doi: https://doi.org/10.1080/01621459.2016.1189337. 1355

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018). "Scaling up data augmentation MCMC via calibration." *The Journal of Machine Learning Research*, 19(1): 2575–2608. MR3899766. 1353

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., and Maji, U. (2021). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, and Kanpur, India. R package version 1.5-0. 1354

Gong, L. and Flegal, J. M. (2016). "A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 25: 684–700. MR3533633. doi: https://doi.org/10.1080/10618600.2015.1044092. 1355

Qin, Q. and Hobert, J. P. (2019). "Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression." *The Annals of Statistics*, 47(4): 2320–2347. MR3953453. doi: https://doi.org/10.1214/18-AOS1749. 1353

Rajaratnam, B. and Sparks, D. (2015). "MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains." *arXiv preprint arXiv:1508.00947*. 1353

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106: 321–337. MR3949306. doi: https://doi.org/10.1093/biomet/asz002. 1355

Zhu, W., Chen, X., and Wu, W. B. (2021). "Online Covariance Matrix Estimation in Stochastic Gradient Descent." *Journal of the American Statistical Association*, 1–30. 1355

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

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

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

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 $i = 1, \dots, I$ 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 $\rho_0 \leq \rho_1$, 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

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

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

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

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 $0 \leq A \leq 1$:

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

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

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 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

Contributed Discussion

Arnab Hazra[*] and Raphaël Huser[†]

It was our pleasure to read this paper and to get the opportunity to discuss it. Studying the convergence and mixing of Markov chain Monte Carlo (MCMC) chains is often neglected. Here, the authors raise this point and obtain some theoretical results about the convergence of the Gibbs sampler for multilevel conditionally hierarchical Gaussian models using multigrid decomposition. The authors also go beyond the Gaussian case and describe an example of a Poisson crossed-effects model. Importantly, the authors also discuss two different types of parametrizations for the same models, the so-called non-centered and centered parametrizations (NCP and CP, respectively).

Here, we focus on studying the convergence properties of the Gibbs sampler under different parametrizations used in some recent papers on spatial geostatistics and spatial extreme-value analysis using hierarchical Gaussian processes (GP), where independent (temporal) replications are available. Instead of focusing on analytic expressions, we focus on simulations. In a purely spatial setting, Bass and Sahu (2017) studied the convergence rates under different choices of the spatial correlation structures for a GP.

First, we consider a simple spatial Gaussian process model (Banerjee et al., 2003, Chapter 5) defined as

$$Y_t(\boldsymbol{s}) = \mu + \varepsilon_t(\boldsymbol{s}) + \eta_t(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2,\ t = 1, \ldots, T, \tag{1}$$

where μ denotes the global mean, $\varepsilon_t(\cdot)$ are independent and identically distributed (IID) zero-mean GPs with spatial covariance $\mathrm{Cov}\{\varepsilon_t(\boldsymbol{s}_1), \varepsilon_t(\boldsymbol{s}_2)\} = r\exp\{-d(\boldsymbol{s}_1, \boldsymbol{s}_2)/\phi\}$, $r, \phi > 0$, with $d(\boldsymbol{s}_1, \boldsymbol{s}_2)$ denoting the Euclidean distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, and $\eta_t(\boldsymbol{s}) \overset{\mathrm{IID}}{\sim} \mathrm{Normal}(0, 1-r)$. We simulate $T = 100$ replications at $N = 121$ uniform spatial grid locations $\{(i,j) : i,j = 0, 0.1, \ldots, 1\}$. True parameter choices are $\mu = 5$, $\phi = 0.2$ and $r = 0.9$. Here, conjugate priors for ϕ and r are not known and hence, to stick to Gibbs sampling, we prefer to treat these parameters as known and choose a weakly informative prior $\mu \sim \mathrm{Normal}(0, 100^2)$. Let $\boldsymbol{X}_t = [X_t(\boldsymbol{s}_1), \ldots, X_t(\boldsymbol{s}_N)]'$ be the generic notation for the spatial vectors and $\boldsymbol{\Sigma}_\phi$ be the correlation matrix obtained from $\mathrm{Cov}\{\varepsilon_t(\boldsymbol{s}_i), \varepsilon_t(\boldsymbol{s}_j)\}, i,j = 1, \ldots, N$. We fit (1) under NCP and CP. In NCP, we treat the levels as $\boldsymbol{Y}_t \overset{\mathrm{Indep}}{\sim} \mathrm{Normal}(\mu\boldsymbol{1} + \boldsymbol{\varepsilon}_t, (1-r)\boldsymbol{I}_N)$, $\boldsymbol{\varepsilon}_t \overset{\mathrm{IID}}{\sim} \mathrm{Normal}(\boldsymbol{0}, r\boldsymbol{\Sigma}_\phi)$, and $\mu \sim \mathrm{Normal}(0, 100^2)$. In CP, the levels are modified as $\boldsymbol{Y}_t \overset{\mathrm{Indep}}{\sim} \mathrm{Normal}(\tilde{\boldsymbol{\varepsilon}}_t, (1-r)\boldsymbol{I}_N)$, $\tilde{\boldsymbol{\varepsilon}}_t \overset{\mathrm{IID}}{\sim} \mathrm{Normal}(\mu\boldsymbol{1}, r\boldsymbol{\Sigma}_\phi)$, and $\mu \sim \mathrm{Normal}(0, 100^2)$. We study the trace plots and the autocorrelation function (ACF) plots of μ and $\bar{\varepsilon} = (NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T \varepsilon_t(\boldsymbol{s}_i)$ under NCP,

[*] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, arnab.hazra@kaust.edu.sa, ORCID: https://orcid.org/0000-0002-3647-2697

[†] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, raphael.huser@kaust.edu.sa, ORCID: https://orcid.org/0000-0002-1228-2071

and of μ and $\bar{\tilde{\varepsilon}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\varepsilon}_t(\boldsymbol{s}_i)$ under CP, and observe good mixing for CP while the mixing is poor for NCP. This corroborates the results for the two-layer models mentioned in the paper. The corresponding effective sample sizes (ESS) are presented in Table 1. Theoretical results follow from Bass and Sahu (2017).

GPs have been criticized for modeling spatial extremes, due to their light tails and inability to capture strong tail dependence. To extend this class, while retaining the computational attractiveness of GPs, several authors have proposed some location and/or scale mixture models (e.g., Huser et al., 2017; Morris et al., 2017; Krupskii et al., 2018; Hazra et al., 2020) and Huser and Wadsworth (2020) reviewed some of them. Here, we focus on some models which allow Gibbs sampling for the higher level random effects.

We next consider a simple location-mixture model (Krupskii et al., 2018), which can be shown to possess upper tail dependence,

$$Y_t(\boldsymbol{s}) = E_t + \varepsilon_t(\boldsymbol{s}) + \eta_t(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2,\ t = 1, \ldots, T, \tag{2}$$

where $E_t \overset{\text{IID}}{\sim}$ Exponential(λ) and the specifications for $\varepsilon_t(\cdot)$ and $\eta_t(\cdot)$ are the same as in (1); here, $\lambda > 0$ is the rate parameter. We choose a weakly informative prior $\lambda \sim$ Gamma$(0.01, 0.01)$. The model has three layers and allows Gibbs sampling for the unknown parameters and the latent variables. The simulation design is the same as before and we choose the true value $\lambda = 1$. We fit (2) under NCP and CP. In NCP, we treat the levels as $\boldsymbol{Y}_t \overset{\text{Indep}}{\sim}$ Normal$(E_t\boldsymbol{1}+\boldsymbol{\varepsilon}_t, (1-r)\boldsymbol{I}_N)$, $\boldsymbol{\varepsilon}_t \overset{\text{IID}}{\sim}$ Normal$(\boldsymbol{0}, r\boldsymbol{\Sigma}_\phi)$, $E_t \overset{\text{IID}}{\sim}$ Exponential(λ), and $\lambda \sim$ Gamma$(0.01, 0.01)$. In CP, we replace the first two levels of NCP by $\boldsymbol{Y}_t \overset{\text{Indep}}{\sim}$ Normal$(\tilde{\boldsymbol{\varepsilon}}_t, (1-r)\boldsymbol{I}_N)$ and $\tilde{\boldsymbol{\varepsilon}}_t \overset{\text{IID}}{\sim}$ Normal$(E_t\boldsymbol{1}, r\boldsymbol{\Sigma}_\phi)$, respectively. The trace plots and ACF plots of $\bar{\varepsilon}$ and $\bar{\tilde{\varepsilon}}$ show a similar pattern as that for (1). The trace plots and ACF plots of $\bar{E} = T^{-1}\sum_{t=1}^{T}E_t$ and λ show good mixing under CP while it is poor for $\bar{E}$ under NCP. The corresponding ESS are presented in Table 1.

Finally, we consider a scale-mixture model (Morris et al., 2017; Hazra et al., 2020)

$$Y_t(\boldsymbol{s}) = \sqrt{b\tau_t}\{\varepsilon_t(\boldsymbol{s}) + \eta_t(\boldsymbol{s})\}, \quad \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2,\ t = 1, \ldots, T, \tag{3}$$

where $\tau_t \overset{\text{IID}}{\sim}$ Inverse-gamma$(a/2, a/2)$ and the other terms are as before. We choose the prior $a \sim$ Discrete-uniform$(0.1, 0.2, \ldots, 50)$ similar to Hazra et al. (2020) and Hazra and Huser (2021), and a flat prior for b over $\mathbb{R}_+$. While different representations of the same model are possible, not all of them allow Gibbs sampling for τ_t, and thus, we skip them here. The model (3) has three layers and allows Gibbs sampling for the unknown parameters (probability proportional to size sampling for a) and the latent variables. The simulation design is the same as before and we choose the true values $a = 5$ and $b = 1$. We fit (3) under non-scaled and scaled parametrizations (NSP and SP, respectively). In NSP, we treat the levels as $\boldsymbol{Y}_t \overset{\text{Indep}}{\sim}$ Normal$(\tilde{\boldsymbol{\varepsilon}}_t, b(1-r)\tau_t\boldsymbol{I}_N)$, $\tilde{\boldsymbol{\varepsilon}}_t \overset{\text{Indep}}{\sim}$ Normal$(\boldsymbol{0}, br\tau_t\boldsymbol{\Sigma}_\phi)$, $\tau_t \overset{\text{IID}}{\sim}$ Inverse-gamma$(a/2, a/2)$, and the priors for a and b. In SP, we replace the first three levels of NSP by $\boldsymbol{Y}_t \overset{\text{Indep}}{\sim}$ Normal$(\tilde{\boldsymbol{\varepsilon}}_t^*, (1-r)\tilde{\tau}_t\boldsymbol{I}_N)$, $\tilde{\boldsymbol{\varepsilon}}_t^* \overset{\text{Indep}}{\sim}$ Normal$(\boldsymbol{0}, r\tilde{\tau}_t\boldsymbol{\Sigma}_\phi)$, and $\tilde{\tau}_t \overset{\text{IID}}{\sim}$ Inverse-gamma$(a/2, ab/2)$, respectively. The trace

Parametrization	GP $(\mu, \bar{\varepsilon}/\bar{\tilde{\varepsilon}})$	Location-mixture	Scale-mixture
		$(\bar{\varepsilon}/\bar{\tilde{\varepsilon}}, \bar{E}, \lambda)$	$(\bar{\tilde{\varepsilon}}/\bar{\tilde{\varepsilon}}^*, \bar{\tau}/\bar{\tilde{\tau}}, a, b)$
NCP/NSP	(48,47)	(55, 53, 902)	(10^4, 95, 6940, 94)
CP/SP	(10^4, 10^4)	(10^4, 8531, 8896)	(10^4, 3256, 7292, 8146)

Table 1: Effective sample sizes (ESS) for the Gibbs samplers for models (1), (2), and (3), under different parametrizations. ESS values correspond to 10^4 iterations, starting from true parameter choices.

plots and ACF plots of $\bar{\tilde{\varepsilon}}$ and $\bar{\tilde{\varepsilon}}^*$ (notation as before) and a show good mixing under both NSP and SP. The trace plots and ACF plots of $\bar{\tau} = T^{-1}\sum_{t=1}^{T}\tau_t$, $\bar{\tilde{\tau}} = T^{-1}\sum_{t=1}^{T}\tilde{\tau}_t$, and b show a good mixing behavior under SP (after thinning by keeping one per four/five samples, for $\bar{\tilde{\tau}}$) but not under NSP. The corresponding ESS are presented in Table 1. Specifying $\tau_t \sim \text{Inverse-gamma}(a/2, b/2)$ as in Morris et al. (2017) show long-range dependence in the ACF plots and using $\tau_t \sim \text{Inverse-gamma}(a/2, ab/2)$, as in Hazra et al. (2020), is recommended.

Overall, through simulation studies, we have illustrated the mixing of a Gibbs sampler under different parametrizations for some popular models for spatial geostatistics and spatial extremes in the light of multigrid decomposition proposed in this paper, which would help practitioners to design MCMC algorithms effectively. Theoretical derivations of the convergence rates in these spatial settings is a possible future endeavor.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC. MR3362184. 1372

Bass, M. R. and Sahu, S. K. (2017). "A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC." *Statistics and Computing*, 27(6): 1491–1512. MR3687322. doi: https://doi.org/10.1007/s11222-016-9700-z. 1372, 1373

Hazra, A. and Huser, R. (2021). "Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model." *The Annals of Applied Statistics*, 15(2): 572–596. MR4298957. doi: https://doi.org/10.1214/20-aoas1418. 1373

Hazra, A., Reich, B. J., and Staicu, A.-M. (2020). "A multivariate spatial skew-t process for joint modeling of extreme precipitation indexes." *Environmetrics*, 31(3): e2602. MR4098481. doi: https://doi.org/10.1002/env.2602. 1373, 1374

Huser, R., Opitz, T., and Thibaud, E. (2017). "Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures." *Spatial Statistics*, 21: 166–186. MR3692183. doi: https://doi.org/10.1016/j.spasta.2017.06.004. 1373

Huser, R. and Wadsworth, J. L. (2020). "Advances in statistical modeling of spatial extremes." *Wiley Interdisciplinary Reviews: Computational Statistics*, e1537. 1373

Krupskii, P., Huser, R., and Genton, M. G. (2018). "Factor copula models for replicated spatial data." *Journal of the American Statistical Association*, 113(521): 467–479. MR3803479. doi: https://doi.org/10.1080/01621459.2016.1261712. 1373

Morris, S. A., Reich, B. J., Thibaud, E., and Cooley, D. (2017). "A space-time skew-t model for threshold exceedances." *Biometrics*, 73(3): 749–758. MR3713109. doi: https://doi.org/10.1111/biom.12644. 1373, 1374

Contributed Discussion[*]

Christian P. Robert[†]

Congratulations to the authors, for this paper that examines in great details the fine convergence properties of several Gibbs versions of the same hierarchical posterior for an analysis of variance (ANOVA) type linear model. Although this may sound like an old-timer opinion, I find it most enjoyable to have Gibbs sampling back on track as a Markov chain Monte Carlo (MCMC) technique whose convergence properties can be properly assessed (Hobert and Geyer, 1998; Meng and van Dyk, 1999; Hobert, 2000; Hobert and Marchev, 2008). Also, even after all these years, it is always a surprise for me to realise that different versions of Gibbs samplings may hugely differ in convergence properties (Roberts and Sahu, 1997; Meng and van Dyk, 1999).

At first, intuitively, I thought the options (1,0) and (0,1) in the parameterisations of the hierarchical linear model should have been similarly performing. But I then realised the symmetry was missing, as one is "more" hierarchical than the other. While the results in the paper exhibiting a theoretical ordering of these choices are truly impressive in their precision and generality, I would suggest pursuing first an random exploration of the various parameterisations in order to handle cases where an analytical ordering proves impossible. It would most likely produce a superior performance, as hinted at by Figure 4. (This alternative happens to be briefly mentioned in the Conclusion section.) The notion of choosing the optimal parameterisation at each step is indeed somewhat unrealistic in that the optimality zones exhibited in Figure 4 are most likely unreachable in a more general model than the Gaussian ANOVA model. This is the more likely with a high number of parameters, parameterisations, and recombinations in the model (Section 7).

An idle and related question is whether or not an extension can be considered, namely to a more general hierarchical model where recentring is not feasible because of the non-linear nature of the parameters.

As noted above, Theorem 1 is both quite impressive and wide ranging. It also reminds me of the interweaving properties (Liu et al., 1994; Yu and Meng, 2011) and data augmentation versions of the early-day Gibbs. More to the point and to the current era, it offers more possibilities for coupling, parallelism, and increasing convergence, as well as for fighting against dimension curses.

> *"in this context, imposing identifiability always improves the convergence properties of the Gibbs Sampler."*

Another idle thought of mine is to wonder whether or not there is a limited number of reparameterisations to be exploited for building Gibbs samplers. I would imagine that

[*] This work was partly supported by a PaRis AI Research InstitutE (prAIrie) from the Agence Nationale de la Recherche (ANR-19-P3IA-0001).

[†] CEREMADE, Université Paris Dauphine PSL, University of Warwick, and CREST, xian@ceremade.dauphine.fr

by creating unidentifiable decompositions of (some) parameters, e.g., $\mu = \mu_1 + \mu_2 + \cdots$, one could unrestrictedly multiply the number of parameterisations. In contrast with imposing hard identifiability constraints as in Section 4.2, my intuition was that this "desidentification" would increase the mixing behaviour of the extended chain, but this somewhat clashes with the above (rigorous) statement from the authors.

The paper also opens the prospect of different possible implementations of Hamiltonian Monte Carlo (HMC) depending on different parameterisations (and different Hamiltonian functions, see, e.g., Thin et al. 2021; Mongwe et al. 2021) and whether or not the impact of these choices has been studied for HMC (which may be linked with Remark 2 in the paper).

References

Hobert, J. P. (2000). "Hierarchical models: a current computational perspective." *Journal of the American Statistical Association*, 95: 1312–1316. MR1825284. doi: `https://doi.org/10.2307/2669778`. 1376

Hobert, J. P. and Geyer, C. J. (1998). "Geometric Ergodicity of Gibbs and Block Gibbs Samplers for a Hierarchical Random Effects Model." *Journal of Multivariate Analysis*, 67: 414–430. MR1659196. doi: `https://doi.org/10.1006/jmva.1998.1778`. 1376

Hobert, J. P. and Marchev, D. (2008). "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms." *Annals of Statistics*, 36: 532–554. MR2396806. doi: `https://doi.org/10.1214/009053607000000569`. 1376

Liu, J., Wong, W., and Kong, A. (1994). "Covariance structure of the Gibbs sampler with application to the comparison of estimators and augmentation schemes." *Biometrika*, 81: 27–40. MR1279653. doi: `https://doi.org/10.1093/biomet/81.1.27`. 1376

Meng, X. and van Dyk, D. (1999). "Seeking efficient data augmentation schemes via conditional and marginal augmentation." *Biometrika*, 86: 301–320. MR1705351. doi: `https://doi.org/10.1093/biomet/86.2.301`. 1376

Mongwe, W., Mbuvha, R., and Marwala, T. (2021). "Quantum-Inspired Magnetic Hamiltonian Monte Carlo." *PLoS ONE*, 16(10): e0258277. 1377

Roberts, G. and Sahu, S. K. (1997). "Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler." *Journal of the Royal Statistical Society. Series B*, 59: 291–317. MR1440584. doi: `https://doi.org/10.1111/1467-9868.00070`. 1376

Thin, A., Janati, Y., Le Corff, S., Ollion, C., Doucet, A., Durmus, A., Moulines, E., and Robert, C. P. (2021). "NEO: Non Equilibrium Sampling on the Orbit of a Deterministic Transform." *arXiv*, 2103.10943. 1377

Yu, Y. and Meng, X.-L. (2011). "To Center or Not to Center: That Is Not the Question–An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency." *Journal of Computational and Graphical Statistics*, 20(3): 531–570. MR2878987. doi: `https://doi.org/10.1198/jcgs.2011.203main`. 1376

Contributed Discussion

Kaoru Irie[*] and Shonosuke Sugasawa[†]

The authors addressed the well-known yet challenging problem about (non)centering of parameters and its effect on the posterior computation and derived the convergence rates of the Gibbs sampler explicitly under the Gaussian hierarchical location models. We are delighted to see the theoretical results provided in the article, expecting them to be utilized by many Bayesian practitioners. We hope the series of studies will follow to cover the more structured models for which the effects of centering have been discussed (e.g., Kastner and Frühwirth-Schnatter 2014).

The convergence rates (Theorem 3 for Model S3), or the inequality conditions for the optimal parametrization (Corollary 2), is a promising result to offer guidance for practitioners in choosing the best parametrization. A gap to be filled is that the convergence rates provided in the article are conditional on the variances of the data generating process, $(\sigma_a, \sigma_b, \sigma_e)$, which are, of course, unknown to statisticians. Thus, our question is; how can we utilize these results in finding the best parametrization in practice? Below are our thoughts with the focus on Model S3.

One approach to the choice of parametrization is to plug in the point estimates of the variance parameters, $(\hat{\sigma}_a, \hat{\sigma}_b, \hat{\sigma}_e)$, in the convergence rates or the inequality conditions before running a Gibbs sampler. The moment estimator is an example that can be computed fast even with large I and J. This naive approach is expected to be successful if the posteriors of the variance parameters are highly concentrated, as in the examples considered in the article. In practice, however, it is also possible that the variance posteriors are diffused and two or more parametrizations are nearly tied.

Another approach is to switch from one parametrization to another at each iteration of a Gibbs sampler based on the sampled variance parameters. In doing so, after sampling $(\sigma_a, \sigma_b, \sigma_e)$ at each iteration, we check the inequalities of Corollary 2 with the latest samples plugged-in, decide the best parametrization, reparametrize the sampled location parameters accordingly if necessary, and move on to the next sampling step. This heuristic approach makes the parametrization adaptive, but the resulting stationary distribution might be different from the original posterior distribution.

A formal approach to the use of multiple parametrizations in a Gibbs sampler is known as the ancillary-sufficiency interweaving strategy (ASIS). Specifically, for Model S3, the component-wise interweaving strategy can be applied to the sampling steps of $\boldsymbol{\gamma}$ and μ (Yu and Meng 2011, Section 2.5). For example, when sampling γ_i at an iteration of the Gibbs sampler, we first sample it under the $GS(1,1)$ parametrization, set $b_{ij} = \eta_{ij} - \gamma_i$, re-sample γ_i under the $GS(1,0)$ parametrization, and then rebuild η_{ij} as $\gamma_i + b_{ij}$. A similar step is added to the sampling of μ, using the $GS(0,0)$ parametrization.

[*] Faculty of Economics, The University of Tokyo, irie@e.u-tokyo.ac.jp

[†] Center for Spatial Information Science, The University of Tokyo, sugasawa@csis.u-tokyo.ac.jp

We examine the ideas listed above in a simulation study. We follow the same setting in Section 2.1, except that we use smaller sample size: $I = J = 20$ and $K = 5$. We also consider different values of σ_b and σ_e in some scenarios, as summarized in Table 1 with the convergence rates that are computed using Theorem 3. When σ_b is modified to $10/\sqrt{K} + 0.2$, there are two promising parametrizations that have similar convergence rates.

In this setting, we employ the posterior sampling by four ways of choosing parametrization: using the parametrization with the best convergence rate (BEST), utilizing the inequality conditions based on the moment estimators of variances (MM), changing parametrizations adaptively during the iterative sampling based on the latest variance parameters sampled (ADPT), and using multiple parametrizations at every iteration (ASIS). For each sampler, we generate 1000 posterior samples after the 1000 burn-in samples, and computed the effective sample sizes (ESS) of μ, $a.$ ($\gamma.$), and $b..$ ($\eta..$).

This procedure is replicated 500 times. The averaged values of ESS are reported in Table 2. When $\sigma_b = 10^{-0.5}$ and the posteriors of the variances are relatively concentrated, the MM method can choose the optimal parametrization in most cases, showing the same efficiency as in the optimal BEST method. In contrast, when $\sigma_b = 10/\sqrt{K}+0.2$, it becomes more challenging to find the optimal parametrization prior to the main analysis, as seen in the disproved ESS of the MM method. This observation could explain the reason that the ADPT method can outperform the analysis based on the optimal choice of a single parametrization, as it implicitly quantifies the uncertainty about the variance parameters and switches between the two parametrizations. Similarly, the ASIS method provides the better ESS in most scenarios, except for μ when $\sigma_b = \sigma_e = 10^{-0.5}$.

	$\sigma_a = 10, \sigma_e = 10$				$\sigma_a = 10, \sigma_e = 10^{-0.5}$			
σ_b	GS(1,1)	GS(0,0)	GS(1,0)	GS(0,1)	GS(1,1)	GS(0,0)	GS(1,0)	GS(0,1)
$10^{-0.5}$	0.990	0.995	1.000	0.015	0.091	0.995	0.995	0.910
$\frac{10}{\sqrt{K}} + 0.2$	0.990	0.484	0.989	0.527	0.522	0.956	0.478	0.957

Table 1: Theoretical convergence rate of each simulation scenario.

		$\sigma_a = 10, \sigma_e = 10$				$\sigma_a = 10, \sigma_e = 10^{-0.5}$			
σ_b		BEST	MM	ADPT	ASIS	BEST	MM	ADPT	ASIS
$10^{-0.5}$	μ	971	971	972	975	855	855	875	32
	$a.$ ($\gamma.$)	938	938	979	984	709	709	779	848
	$b..$ ($\eta..$)	875	875	862	925	888	886	873	932
$\frac{10}{\sqrt{K}} + 0.2$	μ	954	919	825	976	333	347	366	382
	$a.$ ($\gamma.$)	317	116	991	981	817	660	700	859
	$b..$ ($\eta..$)	316	115	749	976	382	367	716	979

Table 2: Effective sample sizes (ESS) of μ, $a.$ (or $\gamma.$), and $b..$ (or $\eta..$) of 1000 posterior samples generated by four methods, being averaged over 500 Monte Carlo replications.

References

Yu, Y. & Meng, X. L. (2011). To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, **20**, 531–570. MR2878987. doi: https://doi.org/10.1198/jcgs.2011.203main. 1378

Kastner, G. & Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics and Data Analysis*, **76**, 408–423. MR3209449. doi: https://doi.org/10.1016/j.csda.2013.01.002. 1378

Contributed Discussion

Sam Power[*] and Andi Wang[†]

We would like to congratulate the authors on this impressive piece of work, which provides mathematically elegant constructions and theoretical results which are of immediate practical relevance to the MCMC research community and practitioners. The authors do an admirable job of laying the groundwork for a linear-algebraic grammar of hierarchical models, built upon an ingenious random walk construction on the underlying graphical model. As such, our comments will focus primarily on this auxiliary random walk and the signal decomposition which it accompanies.

One of the key innovations of this work is the auxiliary random walk X, which is crucial to defining the averaging operators $\phi^{(p)}$. The Markov chain X moves from the root node of the tree t_0, and up the tree one layer at a time, choosing among the children of the current note at each step. It bears acknowledging at the outset that the construction of this auxiliary random walk, as well as the conditions (S) and ($\tilde{\mathrm{S}}$), are somewhat opaque, and so we humbly seek to press the authors for some additional intuition. Why should these be the right transition probabilities? What does c represent? Is there an abstract property which characterises (directed?) Gaussian graphical models which possess this property? Is there some way of understanding the equivalence classes which are induced by c?

The construction of the averaging maps ϕ and the differencing maps δ is intriguing in that it provides a decomposition of the full signal β which is prescribed *intrinsically* by the model and its corresponding graphical structure. Here, we briefly present some observations and interpretations of this decomposition.

To begin with, we note that while $\phi^{(p)}$ and $\delta^{(p)}$ are defined rather abstractly through the process X, they are ultimately *linear* mappings of the multivariate Gaussian vector β_T, and thus are themselves also multivariate Gaussian vectors. The map $\phi_r^{(p)}\beta^{(d)}$ encodes a coarse summary of the coefficients of the signal at level d, as viewed from the vantage point of node r at level p. Similarly, the map $\delta_r^{(p)}\beta^{(d)}$ encodes the increment in this summary which is obtained by moving from $\mathrm{pa}(r)$ to r, furnishing the decomposition with a multi-resolution character.

Additionally, for each d, one can summarise certain properties of the ϕ, δ maps by noting that $M_p^{(d)} = \phi_{X_p}^{(p)}\beta^{(d)}$ admits a martingale-like structure under the evolution of X_p according to the auxiliary random walk on the tree, where p indexes the steps of the walk: for any $p < p' \leq d$, $\mathbb{E}[M_p^{(d)}|\beta_T, X_{p'}] = M_{p'}^{(d)}$. One wonders whether this perspective may be useful in extending the scope of these results to more general settings.

Another curious aspect of the decomposition is precisely what information is contained in each component. While one might interpret both of the indices (d, p) which

[*] University of Bristol, sam.power@bristol.ac.uk
[†] University of Bristol, andi.wang@bristol.ac.uk

parametrise δ as corresponding to levels within the tree, there is perhaps some utility in drawing a slight distinction: d corresponds to a level within the tree, but p corresponds to a pair of adjacent levels within the tree, or even the collection of edges which join those two levels.

One can then view the collection $\Delta^{(p)} := \{\delta^{(p)}\beta^{(d)} : d \in \{p, \cdots, k-1\}\}$ as containing the increments in information which are obtained by stepping across the various edges which pass from level $(p-1)$ to level p, and observing their subsequent impact at levels $d \geqslant p$. By contrast, $\Delta_{(d)} := \{\delta^{(p)}\beta^{(d)} : p \in \{0, \cdots, d\}\}$ holds the increments in information about only level d, as generated by stepping across edges from the root towards the pth level. The usual Gibbs sampler makes updates based on $\{\Delta_{(d)}\}$, but a key insight from the present work is that the same Gibbs sampler also induces a rich structure on $\{\Delta^{(p)}\}$, hinting at a fascinating duality structure.

We now pass brief comment on two of the key auxiliary lemmata, namely Lemma 4 and Lemma 3.1. Lemma 4 computes the conditional expectation

$$E[\delta^{(p)}\beta^{(d)}|\beta\backslash\beta^{(d)}] = \sum_{\ell} c_{d\ell}\delta^{(p)}\beta^{(\ell)},$$

which will allow the authors to conclude, in the Proof of Theorem 9, the equality in law

$$\mathcal{L}(\delta^{(p)}\beta^{(d)}|\beta\backslash\beta^{(d)}) = \mathcal{L}(\delta^{(p)}\beta^{(d)}|(\delta^{(p)}\beta^{(\ell)})_{\ell\geq p,\ell\neq d}). \tag{1}$$

This follows from the fact that for conditional Gaussian distributions, the conditioned values can only affect the conditional mean and have no influence on the conditional variance. This is crucial to establish that each $\delta^{(p)}\beta$ in fact evolves as a *Markov* chain for the specific scan order used in the paper. In fact, the proof of Lemma 4 actually shows that the same relation (1) holds with $\delta^{(p)}$ replaced with $\phi^{(p)}$. However, the $\phi^{(p)}$'s will not be conditionally independent, unlike the $\delta^{(p)}$'s.

While Lemma 4 can be read as a statement about the dependence structure of $\Delta^{(p)}$, Lemma 3.1. concerns instead the dependence structure of $\Delta_{(d)}$ under the same update. We again wonder whether some notion of duality is at play in the relevance of these two results.

Finally, in light of the richness of the random walk construction and the induced signal decomposition, we cannot help but wonder as to whether these tools might be fruitfully adapted to the case of undirected Gaussian graphical models.

References

Zanella, G. and Roberts, G. (2020). "Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions." *Bayesian Analysis*, 1–35.

Contributed Discussion

Peng Zhao[*], Hou-Cheng Yang[†], and Guanyu Hu[‡]

We would like to congratulate the authors on their excellent effort to this important topic. The author develops a multigrid decomposition approach that allows to focus on low-dimensional Markov chains and it derives analytic expressions for the convergence rates of the Gibbs sampler for various multi-level linear models structures. This novel approach is inherent to the model, rather than specific to the selected parameterization. We think this paper contributes to the quantitative understanding of the interaction between the Bayesian hierarchical structure and the behavior of the Markov chain Monte Carlo (MCMC) algorithm, which is the core and important of the actual success of Bayesian statistics.

1 Comments on Coordinate-Wise Scheme

In addition to Gibbs samplers, non-centered and centered parameterizations also exhibit distinctive behaviors in variational inference, which is another popular stream of Bayesian inference. Specifically, the coordinate ascent variational inference (CAVI) algorithm uses similar coordinate-wise updating in the commonly used mean-field (MF) framework: $q(\theta_i) \propto \exp[\mathbf{E}_{\theta_{-i}}\{\log P(y, \theta)\}]$. Unfortunately, the objective function associated with MF variational inference is usually non-convex, making it difficult to evaluate CAVI's convergence rate. It would be interesting to determine whether CAVI convergence can be analyzed using the tool developed and used in the paper. Specifically, if the variational distribution $q(\theta)$ under MF is Gaussian, then the trace of CAVI's updating would also be a Gaussian autoregressive process, similar to the settings of Theorem 1.1. The tractability of such an autoregressive transition kernel needs to be determined.

2 Comments on Non-Gaussian Hierarchical Model

In addition to Gaussian response, several other types of data such count-valued data, binary data and multiple response-types are also important for statistician. For the non-Gaussian data, the conjugate priors (Chen et al., 2008; Bradley et al., 2018, 2020, 2021) are well established under generalized linear model. The conjugate priors also lead Gibbs sampling algorithm without Metropolis-Hasting step. For the current discussion, it based on independent Normal case and has a symmetry assumption. However, we were concerning how this fantastic approaching and methodologies to apply for spatial/spatio-temporal statistics or even says, hierarchical dependent model. Those model might be non-stationary in time and space and asymmetric.

[*] Department of Statistics, Texas A&M University, College Station, TX, pzhao@stat.tamu.edu
[†] Department of Statistics, Florida State University, Tallahassee, FL, hy15e@my.fsu.edu
[‡] Department of Statistics, University of Missouri, Columbia, MO, gh7mr@missouri.edu

3 Comments on Sparse, Dependent, and Mixture Models

The data model discussed in the article is the independent Gaussian response model, where the parameters have additive effects. Modern statistical applications may include other kinds of hierarchical parameters as a result of sparsity, dependence, and heterogeneity. For example, in the sparsity regime, shrinkage priors are commonly used in high-dimensional applications, where the hierarchical parameters play a multiplicity role: for the famous horseshoe prior (Carvalho et al., 2010), $\theta \sim N(0, \lambda^2\tau^2)$, where $\lambda, \tau \sim \mathrm{Ca}^+(0, 1)$, where $\mathrm{Ca}^+(0, 1)$ is the half-Cauchy distribution. The non-centered parameterization for such prior would be multiplicative $\theta = \beta\lambda\tau$ with $\beta \sim N(0, 1)$ and $\lambda, \tau \sim \mathrm{Ca}^+(0, 1)$. A second challenge would be to analyze the convergence rate of the MCMC sequence of horseshoe prior under centered and non-centered parameterization.

The Gaussian process (Gelfand and Schliep, 2016) has emerged as the most valuable tool in the toolkit for spatially dependent data. It is also important to consider convergence rate of such dependent model with Gibbs samplers. The final important model is the mixture model under nonparametric Bayesian framework. The nonparametric Bayesian model offers a good choice to simultaneously estimate the number of clusters and cluster configurations. Based on proper choice of base distribution in such mixture model, a Gibbs sampler algorithm (Neal, 2000) can be used for such model. Compared with tradition model, the multi-grid decomposition should be carefully implemented under mixture models.

References

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)." *Bayesian Analysis*, 13(1): 253–310. MR3773410. doi: https://doi.org/10.1214/17-BA1069. 1383

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020). "Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family." *Journal of the American Statistical Association*, 115(532): 2037–2052. MR4189775. doi: https://doi.org/10.1080/01621459.2019.1677471. 1383

Bradley, J. R. et al. (2021). "Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model." *Bayesian Analysis*. 1383

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. 1384

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). "Bayesian variable selection and computation for generalized linear models with conjugate priors." *Bayesian analysis (Online)*, 3(3): 585. MR2434404. doi: https://doi.org/10.1214/08-BA323. 1383

Gelfand, A. E. and Schliep, E. M. (2016). "Spatial statistics and Gaussian processes: A beautiful marriage." *Spatial Statistics*, 18: 86–104. MR3573271. doi: `https://doi.org/10.1016/j.spasta.2016.03.006`. 1384

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of computational and graphical statistics*, 9(2): 249–265. MR1823804. doi: `https://doi.org/10.2307/1390653`. 1384

Rejoinder

Giacomo Zanella[*] and Gareth Roberts[†]

1 Introduction

We thank all discussants for their interesting and stimulating contributions. Given the tremendous breath covered by these contributions, we will not be able to answer all points raised therein in a comprehensive way, but instead we select the ones we found more stimulating and we felt we had more to comment on.

2 Extensions, limitations and applications

2.1 Model extensions

Various authors discuss and propose extensions to alternative or more general classes of models. Hazra and Huser discuss applications to Gaussian Process regression models for spatial statistics and extreme value theory. See also Bass and Sahu (2017, 2019) for related work. Yang and Liu explore in great details extensions to vector hierarchical models with non-diagonal covariance matrices and more general regression models with covariates. In doing so they also highlight various challenges in deriving explicit results and bounds for such more general models. Zhao, Yang, and Hu ask about extensions to general exponential families. We agree that this would be an interesting line of research and in our paper we only discuss extensions to Poisson regression models in Remark 2. However, as many discussant highlight, it is highly non-trivial to extend the very explicit results we derive in this paper to non-Gaussian models. This would clearly be a major breakthrough and advance in the Markov chain Monte Carlo (MCMC) literature if feasible. Nonetheless, our current results are already directly relevant to optimize samplers for models featuring conditionally Gaussian distributions (see e.g. discussion on conditionally optimal parametrizations) and are also relevant in large data limits, as for example discussed in (Roberts and Sahu, 2001).

2.2 Algorithmic extension

Both Flegal and Zhou and Zhou discuss about extensions of our results from deterministic-scan Gibbs Samplers to random-scan ones. We expect the results to extend as already partially outlined by Zhou and Zhou. Zhao, Yang, and Hu ask about extensions to Coordinate Ascent Variational Inference (CAVI) algorithms for mean-field variational approximations. This is a good point and in fact the convergence rate of Gibbs Sampling and CAVI are often related, and they coincide in the Gaussian case

[*] Department of Decision Sciences, BIDSA and IGIER; Bocconi University, Milan, Italy, giacomo.zanella@unibocconi.it

[†] Department of Statistics, University of Warwick, Coventry, UK, gareth.o.roberts@warwick.ac.uk

(Tan and Nott, 2014). Similarly, older work has already shown close connection to other algorithms (e.g. conditional maximisation expectation maximization Sahu and Roberts, 1999). We are indeed working on extending the multigrid analysis approach to study performances of CAVI for multilevel models, both in terms of convergence rates and accuracy of mean-field approximations. See also the recent work in Tan (2021) on reparametrization techniques in variational methods. Various authors also comment about and explore the role of partial noncentering, which is only briefly considered in the paper.

Finally, Robert discusses the potential relevance of the proposed multigrid analysis to other MCMC algorithms, such as Hamiltonian Monte Carlo (HMC). We agree that this would be a very interesting line of research. We also take this occasion to mention that Bob Carpenter suggested to us in private communication some careful modifications to the Stan code we used (which is publicly available in the supplementary material) to optimize the implementation of the HMC and No-U-Turn sampler (NUTS) algorithms that we run in Section 5.1. His modifications are able to reduce the runtime of NUTS by a factor of two. This does not change the qualitative conclusions of Section 5.1, neither in terms of relative comparison between NUTS-v1, v2 and v3 not in terms of comparison to Gibbs-v1, v2 and v3, but it is important to mention. The optimized Stan code is available at https://github.com/gzanella/multigrid. Note also that the HMC and NUTS runtimes reported in Table 3 do not include the burn-in or warm-up time, but only the sampling one. This may not have been clear from the experiment description in Section 5.1.

3 Conditionally optimal parametrizations

Both Robert as well as Irie and Sugasawa ask clarifications about the practical implementation of the "conditionally optimal parametrization" discussed in Section 3.2. As Irie and Sugasawa explain, Corollary 2 characterizes the optimal parametrization for $\boldsymbol{\beta}$ given the variance parameters $(\sigma_a, \sigma_b, \sigma_e)$. In Section 3.2 we point out that, given those results, "one can at each iteration choose the optimal parametrization $\boldsymbol{\beta}$ given $(\sigma_a, \sigma_b, \sigma_e)$ according to Table 1". This scheme, which we refer to as *Optimal* in Figure 5, is exactly the approach described by Irie and Sugasawa in the fourth paragraph of their discussion: to "switch from one parametrization to another at each iteration of a Gibbs sampler" by checking "the inequalities of Corollary 2 with the latest samples [of $(\sigma_a, \sigma_b, \sigma_e)$] plugged-in". As mentioned in the paper, this approach does not add significant computational costs compared to usual Gibbs Sampling with a fixed parametrization, as it basically only requires to check the inequalities in Corollary 2 and then add an *if* statement pointing to the part of the code implementing the desired sampler (i.e. the one with the conditionally optimal parametrization). See e.g. https://github.com/gzanella/multigrid for an R implementation. Irie and Sugasawa wonder whether such adaptive strategy which changes parameterisation *on the fly*, depending on current variance components induce bias and whether the "resulting stationary distribution might be different from the original posterior distribution." This is not the case as shown by the following proposition.

Proposition 1. *Suppose π is a joint distribution for $(\sigma, \boldsymbol{\beta}) \in \mathcal{X} \times \mathcal{Y}$, and $\{P_\lambda, \lambda \in \Lambda\}$ is a family of Markov kernels which all leave π invariant and do not change σ, i.e. which* *satisfy* $P_\lambda((\sigma,\boldsymbol{\beta}),\{\sigma\}\times\mathcal{Y})=1$ *for any* $\sigma,\boldsymbol{\beta}$ *and* λ*. Define a kernel* Q *that selects* λ *depending on the current value of* σ*, i.e.* $Q((\sigma,\boldsymbol{\beta}),A):=P_{\lambda(\sigma)}((\sigma,\boldsymbol{\beta}),A)$ *for* $A\subseteq\mathcal{X}\times\mathcal{Y}$*, where* $\lambda:\mathcal{X}\to\Lambda$ *is an arbitrary function. Then* Q *is also invariant for* π*.*

Proof. Since P_λ is π-invariant and does not change σ it follows that $P_\lambda((\sigma,\cdot),\{\sigma\}\times\cdot)$ is a $\pi(\boldsymbol{\beta}|\sigma)$-invariant kernel on $\mathcal{Y}$. To see that, for every $A\subset\mathcal{X}\times\mathcal{Y}$ and $\sigma\in\mathcal{X}$ define $A_\sigma:=\{\boldsymbol{\beta}\,:\,(\boldsymbol{\beta},\sigma)\in A\}\subseteq\mathcal{Y}$ and

$$\int_{\mathcal{Y}}P_\lambda((\sigma,\boldsymbol{\beta}),A)\pi(d\boldsymbol{\beta}|\sigma)=\int_{\mathcal{Y}}P_\lambda((\sigma,\boldsymbol{\beta}),\{\sigma\}\times A_\sigma)\pi(d\boldsymbol{\beta}|\sigma)=:g(\sigma,A_\sigma),$$

where the first equality holds since $P_\lambda((\sigma,\boldsymbol{\beta}),\{\sigma\}\times\mathcal{Y})=1$ and thus $g(\sigma,A_\sigma)$ depends only on A through A_σ. By π-invariance $\int g(\sigma,A_\sigma)\pi(d\sigma)=\pi(A)$ for every $A\subseteq\mathcal{X}\times\mathcal{Y}$, meaning that $g(\sigma,A_\sigma)$ equals (a version of) the conditional distribution $\pi(\boldsymbol{\beta}\in A_\sigma|\sigma)$ as desired. Thus $\int_{\mathcal{Y}}P_\lambda((\sigma,\boldsymbol{\beta}),A)\pi(d\boldsymbol{\beta}|\sigma)=\pi(\boldsymbol{\beta}\in A_\sigma|\sigma)$ $\pi(\sigma)$-almost surely for any λ and

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}}\pi(\sigma,\boldsymbol{\beta})Q((\sigma,\boldsymbol{\beta}),A)d\sigma d\boldsymbol{\beta}&=\int_{\mathcal{X}}\int_{\mathcal{Y}}\pi(\sigma,\boldsymbol{\beta})P_{\lambda(\sigma)}((\sigma,\boldsymbol{\beta}),A)d\boldsymbol{\beta}d\sigma\\&=\int_{\mathcal{X}}\pi(\boldsymbol{\beta}\in A_\sigma|\sigma)\pi(d\sigma)=\pi(A)\end{aligned}$$

meaning that Q is π-invariant. □

The above proposition applies to the on-the-fly parametrization discussed above. In such case P_λ denote the Gibbs Samplers updating $\boldsymbol{\beta}$ with different parametrizations and $\lambda(\sigma)$ the optimal choice of parametrization described in Table 1. Recall that Gibbs Samplers with different parametrizations, e.g. $GS(0,0)$, $GS(1,0)$, $GS(0,1)$ and $GS(1,1)$ defined in Section 1, can be interpreted as different transition kernels for the same variables $\boldsymbol{\beta}$ (under any arbitrarily chosen parametrization used as reference). We thank Irie and Sugasawa for giving us the chance to clarify this important and subtle point.

4 Connections with auxiliary Markov chains

Sam Power and Andi Wang ask great questions about the interpretability of the results in Section 7 of the paper, concerning k level hierarchical models. Firstly we address their insightful observation that $\phi^{(p)}_{X_p}\beta^{(p)}$ admits a martingale-like structure. In fact this is closely related to the independence of the $\{\delta^{(i)},i=0,1\ldots k-1\}$ projections. The martingale property is precisely that of uncorrelated increments which, together with Gaussianity of β gives the required independence. Of course something more general is required though when moving away from the Gaussian context as being uncorrelated no longer implies independence. So it is not clear (to us at least) how we can move beyond the Gaussian case using these ideas.

Secondly, Sam and Andi ask (very reasonably) for more intuition into the construction and properties of the ϕ and δ and the c matrix. In more simplified settings (for instance just 2 level models) insight is provided in Papaspiliopoulos et al. (2003, 2007). We just give some insights which illuminate Corollary 7 in the paper in particular and hopefully clarify the role of the cs. Within model NSk (in fact in most of the models in our paper) the full conditional for any component s depends only on its parent and children. The dependence on the parent defines the local influence of the prior while the dependence on its children encodes the influence of data. Intuition is clearest in the case of Model Sk in the paper and we'll restrict attention to that. We recall that the rate of convergence of the Gibbs sampler for Model Sk can be characterised as the principal (Perron-Frobenius) eigenvalue of the non-negative matrix:

$$M_k = \begin{pmatrix} 0 & r_1 & & & \\ (1-r_2) & 0 & r_2 & & \\ & \dots & \dots & \dots & \\ & & (1-r_{k-2}) & 0 & r_{k-2} \\ & & & (1-r_{k-1}) & 0 \end{pmatrix},$$

where it turns out that we write explicitly the cs in terms of their weighted parameters (the r_is) which are normalised to have unit row sum (at least for $1 < i < k-1$). In the Sk model, each child has equal influence, and the relative influences of parent and children is constant across all other components t with $\ell(t) = \ell(s)$. In fact we see that the full conditional for s has mean given by

$$r_{\ell(s)}\bar{c}(s) + (1 - r_{\ell(s)})pa(s)$$

with $\bar{c}(s)$ denoting the mean of its children.

However we can gain much from a probabilistic interpretation of M_k. Indeed M_k is a sub-stochastic Markov matrix describing a skip-free Markov chain with absorption at -1 or k. The rate of convergence described by the result therefore corresponds to the quasi-stationary decay rate (see for example Collet et al. (2012)), a measure of the rate at which this Markov chain leaves $0, 1, \dots, k-1$. Many interesting conclusions can be drawn from this analogy. For instance the flat prior case corresponds to $\tau_0 = 0$ and in this case there is no possibility of exit to -1 and thus this corresponds to a Gibbs sampler which converges slower than under any proper prior. In fact the rate of convergence is monotonically decreasing in τ_0.

Moreover, if we restrict ourselves to this case ($\tau_0 = 0$), we can see how increasing I_ℓ has the effect of pushing the Markov chain towards higher values and thus exit will be more rapid. Therefore the rate of convergence will be monotonically decreasing as a function of I_ℓ for each ℓ. (Such an argument can be properly formulated within the framework of stochastic monotonicity Daley (1968). This Markov chain is not always stochastically monotone, but instead we can study a lazy version thereof which is stochastically monotone.) Note that the dependence on I_ℓ is far more subtle in the case of $\tau_0 \neq 0$ as absorption is now possible from 0 as well as $k-1$. In general, it is clear that the full implications of Corollary 7 still require further investigation.

Note that outside the Gaussian case, similar ideas to these can be used to characterise qualitative convergence properties of Gibbs samplers on non-Gaussian hierarchical models, see Papaspiliopoulos and Roberts (2008). A characteristic feature of the Gaussian case is that the relative influences of parents and children are constant, i.e. the elements r_i in Corollary 7 are independent of the actual values of $pa(s)$ and $\bar{c}(s)$. Outside the Gaussian case this is no longer the case, but we can still seek to study asymptotic version of the tri-diagonal matrix in Corollary 7. Papaspiliopoulos and Roberts (2008) actually discovers that the irreducibility properties of these asymptotic matrices are closely linked to qualitative stability properties of the underlying Gibbs sampler (such as geometric and uniform ergodicity).

5 Concluding comments

There are a number of interesting and substantial comments to which we have not been able to address in this short rejoinder. Various contributions provide a broader perspective on the multigrid approach developed in our paper. In particular, Yang and Liu discuss the fundamental differences in the intuition behind our multigrid decomposition and more classical multigrid methods. Furthermore, Zhou and Zhou introduce an elegant matrix reformulation of our framework, which we very much hope will lead to substantial generalisations of our work.

It is exciting to see the substantial interest in this topic. One of us remembers fondly the days of the early 1990s sometimes referred to as the "MCMC revolution" in Bayesian Statistics. It is no surprise that this happened at exactly the same time as the prolific growth in the use of multilevel models (see for example Smith and Roberts (1993)). At the time it was empirically observed that Gibbs samplers often converged very rapidly, and scaled extremely well on models with hierarchical structure. Understanding this phenomenon has been an important theoretical challenge, but we hope that our paper makes some contribution towards this.

Our strategy has been to obtain results which strongly rely on the structural properties of nested and crossed models. In this way we have been able to obtain very explicit results, albeit only in the Gaussian case. Nevertheless, it seems reasonable that our methodology should provide useful practical guidance beyond the precise settings where our theory sits, though of course with no absolute guarantees.

Finally we wish to finish be reiterating our thanks to all the participants of this meeting for their insightful, substantial and encouraging comments. Also we'd like to record thanks to Michele Guindani, Mark Steel and Tommaso Rigon for organising this successful event.

References

Bass, M. R. and Sahu, S. K. (2017). "A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC." *Statistics and Computing*, 27(6): 1491–1512. MR3687322. doi: https://doi.org/10.1007/s11222-016-9700-z. 1386

Bass, M. R. and Sahu, S. K. (2019). "Dynamically Updated Spatially Varying Parameterizations of Hierarchical Bayesian Models for Spatial Data." *Journal of Computa-*

tional and Graphical Statistics, 28(1): 105–116. MR3939375. doi: https://doi.org/10.1080/10618600.2018.1482761. 1386

Collet, P., Martínez, S., and San Martín, J. (2012). *Quasi-stationary distributions: Markov chains, diffusions and dynamical systems*. Springer Science & Business Media. MR2986807. doi: https://doi.org/10.1007/978-3-642-33131-2. 1389

Daley, D. J. (1968). "Stochastically Monotone Markov chains." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, (10): 305–317. MR0242270. doi: https://doi.org/10.1007/BF00531852. 1389

Papaspiliopoulos, O. and Roberts, G. (2008). "Stability of the Gibbs sampler for Bayesian hierarchical models." *The Annals of Statistics*, 95–117. MR2387965. doi: https://doi.org/10.1214/009053607000000749. 1389, 1390

Papaspiliopoulos, O., Roberts, G., and Sköld (2003). "Non-centered parameterisations for hierarchical models and data augmentation." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, volume 307. Oxford University Press, USA. MR2003180. 1389

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 59–73. MR2408661. doi: https://doi.org/10.1214/088342307000000014. 1389

Roberts, G. O. and Sahu, S. K. (2001). "Approximate predetermined convergence properties of the Gibbs sampler." *Journal of Computational and Graphical Statistics*, 10(2): 216–229. MR1939698. doi: https://doi.org/10.1198/10618600152627915. 1386

Sahu, S. K. and Roberts, G. O. (1999). "On convergence of the EM algorithmand the Gibbs sampler." *Statistics and Computing*, 9(1): 55–64. 1387

Smith, A. F. and Roberts, G. O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1): 3–23. MR1210421. 1390

Tan, L. S. (2021). "Use of model reparametrization to improve variational Bayes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1): 30–57. MR4220983. doi: https://doi.org/10.1111/rssb.12399. 1387

Tan, S. L. and Nott, D. J. (2014). "Variational approximation for mixtures of linear mixed models." *Journal of Computational and Graphical Statistics*, 23(2): 564–585. MR3215825. doi: https://doi.org/10.1080/10618600.2012.761138. 1387