Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions

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;
Gibbs Sampler and Multigrid Decompositions

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, 2016a; 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 $(\beta(s))_{s=1,2,...}$ be a Markov chain with stationary distribution $\pi$ and transition operator defined by $P^s f(\beta(0)) = E[f(\beta(s))|\beta(0)]$. The rate of convergence $\rho(\beta(s))$ associated to $(\beta(s))_{s=1,2,...}$ is defined as the smallest number $\rho$ such that for all $r > \rho$

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

(1.1)

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

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

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

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

\[ y_{ijk} = \mu + a_i + b_{ij} + \epsilon_{ijk}, \quad (2.1) \]

where \( i, j \) and \( k \) run from 1 to \( I, J \) and \( K \) respectively and \( \epsilon_{ijk} \) are iid normal random variables with mean 0 and variance \( \sigma_e^2 \). 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 \( \mu \).

For the theoretical analysis, we will consider the variance terms \( \sigma_a^2 \), \( \sigma_b^2 \) and \( \sigma_e^2 \) 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 \( \mu \), \( \mathbf{a} \) 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 (2020) for explicit expressions).
As an illustrative example, we simulated data from Model S3 with

2.1 Illustrative example

G. Zanella and G. Roberts

the four different parametrizations, denoting them by noise in the measurements. We fit model S3 assuming the standard deviations (\(\sigma_a, \sigma_b, \sigma_c\)) to be unknown and placing weakly informative priors, namely \(\frac{1}{\sigma_a^2}, \frac{1}{\sigma_b^2}\) and \(\frac{1}{\sigma_c^2}\) 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 \((\mu, a, b)\) and \((\sigma_a, \sigma_b, \sigma_c)\). 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 \(\mu\) 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-
Gibbs Sampler and Multigrid Decompositions

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

- **(a) Fully centred parametrization**
- **(b) Fully non-centred parametrization**
- **(c) Mixed parametrization: \((\mu, a, \eta)\)**
- **(d) Mixed parametrization: \((\mu, \gamma, b)\)**

In rows 3-4 of Figure 2 display the behaviour of the global averages of such parameters, namely \(a = \frac{\sum_i a_i}{J}\) and \(\gamma = \frac{\sum_i \gamma_i}{J}\), 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. \(\mu\)), \(GS(0, 0)\) displays very poor mixing behaviour at level 1 (i.e. \(\gamma\)). 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 \((\mu, \gamma, b)\), as it is the only one mixing well at each of the three levels.
Figure 2: Mixing behaviour at level 0 ($\mu$; 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 multi-level 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 $\mu$ 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: $(\mu, a, b)$, $(\mu, \gamma, \eta)$ and the mixed parametrizations $(\mu, \gamma, b)$ and $(\mu, a, \eta)$. In order to provide a unified treatment, regardless of the chosen parametrization, we denote the parameters used by $\beta = (\beta(0), \beta(1), \beta(2))$ and the resulting Gibbs Sampler by $GS(\beta)$. For example, in the NCP case $\beta(0) = \mu$, $\beta(1) = a$, $\beta(2) = b$ and $GS(\beta)$ coincides with $GS(1,1)$. First consider the map $\delta$ sending $\beta = (\beta(0), \beta(1), \beta(2))$ to

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

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

$$
\delta(0)\beta(0) = \beta(0), \quad \delta(0)\beta(1) = \beta(1), \quad \delta(0)\beta(2) = \beta(2),
$$

$$
\delta(1)\beta(1) = \left(\beta_1^{(1)} - \beta_1^{(1)}, \ldots, \beta_I^{(1)} - \beta_I^{(1)}\right), \quad \delta(1)\beta(2) = \left(\beta_1^{(2)} - \beta_1^{(2)}, \ldots, \beta_I^{(2)} - \beta_I^{(2)}\right),
$$

$$
\delta(2)\beta(2) = \left(\beta_{11}^{(2)} - \beta_{11}^{(2)}, \beta_{12}^{(2)} - \beta_{12}^{(2)}, \ldots, \beta_{IJ}^{(2)} - \beta_{IJ}^{(2)}\right),
$$

where

$$
\beta(1) = \frac{\sum_i \beta_i^{(1)}}{I}, \quad \beta(2) = \frac{\sum_{i,j} \beta_{ij}^{(2)}}{IJ}, \quad \beta(2) = \frac{\sum_j \beta_{ij}^{(2)}}{J}.
$$
Figure 3: Illustration of Theorem 1. Left: the transition from $\beta(s)$ to $\beta(s+1)$ in Sampler $GS(\beta)$ follows the structure of a Gibbs Sampler with 3 components. Right: the transition from $\delta \beta(s)$ to $\delta \beta(s+1)$ in Sampler $GS(\beta)$ follows the structure of three independent Markov chains.

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

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

While the posterior independence of $\delta(0) \beta$, $\delta(1) \beta$ and $\delta(2) \beta$ 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 $\delta(0) \beta$, $\delta(1) \beta$ and $\delta(2) \beta$, respectively, do not depend on the choice of parametrization $\beta$. For example, the four instances of $\delta(0) \beta$ – i.e. $(\mu, a, b \cdot), (\mu, a, \eta \cdot), (\mu, \gamma, b \cdot)$ and $(\mu, \gamma, \eta \cdot)$ – 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, $\beta(s)$. In fact the factorization into independent Markov chains implies that the rate of convergence of $\beta(s)$ is simply given by the slowest rate of convergence among $\delta(0) \beta(s)$, $\delta(1) \beta(s)$ and $\delta(2) \beta(s)$. Interestingly, the slowest chain is always the chain at the highest level $\delta(0) \beta(s)$, regardless of the choice of parametrization and the values of $(I, J, K, \sigma_a, \sigma_b, \sigma_c)$.

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

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

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

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

### 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 $\beta(s)$. Since this is a Gibbs Sampler targeting a multivariate Gaussian distribution, 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 $\Sigma$. However, given the high-dimensionality of $\beta(s)$, which has $1 + I + IJ$ parameters, it is hard to apply directly such results and in fact the convergence properties of $\beta(s)$ 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 $\delta^{(0)} \beta(s)$, 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

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

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

where $\hat{\sigma}_a^2 = \frac{\sigma_a^2}{1}, \hat{\sigma}_b^2 = \frac{\sigma_b^2}{IJ}$ and $\hat{\sigma}_e^2 = \frac{\sigma_e^2}{IJK}$.

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 \(\rho\) is the rate of convergence, as a function of \(\log(\tilde{\sigma}^2_a)\) and \(\log(\tilde{\sigma}^2_e)\) for fixed \(\log(\tilde{\sigma}^2_b) = 0\).

Recall that values of \(\rho\) 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 \((\mu, a, b), (\mu, \gamma, b), (\mu, a, \eta)\) and \((\mu, \gamma, \eta)\)) 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 \(\eta\) at the lowest level if and only if \(\tilde{\sigma}^2_b \geq \tilde{\sigma}^2_e\),
- use a centred parametrization \(\gamma\) 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 \(\rho\) 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

| \(\tilde{\sigma}^2_b \geq \tilde{\sigma}^2_e\) | \(\tilde{\sigma}^2_a \geq \tilde{\sigma}^2_b + \tilde{\sigma}^2_e\) | \(\tilde{\sigma}^2_a < \tilde{\sigma}^2_b + \tilde{\sigma}^2_e\) |
|---|---|---|
| \((\mu, \gamma, \eta)\) | \((\mu, \alpha, \eta)\) |
| \((\mu, \gamma, \eta)\) | \((\mu, \alpha, \beta)\) |

Table 1: Optimal parametrization as a function of the normalized variance components.
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 $\beta$ of the Gaussian component, such as the fully centred parametrization $\beta = (\mu, \gamma, \eta)$, and alternate updating $\beta|\sigma_a, \sigma_b, \sigma_e$ with $GS(\beta)$ and $(\sigma_a, \sigma_b, \sigma_e)|\beta$ with direct sampling (which is straightforward using the conditional independence of $\sigma_a, \sigma_b$ and $\sigma_e$ given $\beta$). Given Corollary 2, instead, one can at each iteration choose the optimal parametrization $\beta$ 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 $\beta|\sigma_a, \sigma_b, \sigma_e$ 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 $(\mu, \gamma, \eta)|\sigma_a, \sigma_b, \sigma_e$ with $GS((0, 0))$ and $(\sigma_a, \sigma_b, \sigma_e)|(\mu, \gamma, \eta)$ exactly; the sampler choosing the optimal parametrization $\beta$ according to Table 1 and then updating $\beta|\sigma_a, \sigma_b, \sigma_e$ with $GS(\beta)$ and $(\sigma_a, \sigma_b, \sigma_e)|\beta$ exactly; the sampler updating both $\beta|\sigma_a, \sigma_b, \sigma_e$ and $(\sigma_a, \sigma_b, \sigma_e)|\beta$ exactly (which can be implemented because the distribution of $\beta|\sigma_a, \sigma_b, \sigma_e$ 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 $\beta | (\sigma_a, \sigma_b, \sigma_e)$ in the Optimal sampler only requires univariate updates and has a potentially lower computational cost compared to a full multivariate block update of $\beta | (\sigma_a, \sigma_b, \sigma_e)$, 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 $GS(\beta)$ 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 $\beta | (\sigma_a, \sigma_b, \sigma_e)$ and $(\sigma_a, \sigma_b, \sigma_e) | \beta$) the difference between and exact update of $\beta$ 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 $\beta$ and $(\sigma_a, \sigma_b, \sigma_e)$, and in this case it is successful in doing so for the slowly mixing parameter $\sigma_b$ (see Figure 5, rows 4-6). The results suggest that the optimal choice of parametrization for $\beta$ 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 C$k$ (k-factors crossed-effects model). \( y_{i_1...i_k} = \mu + a_1^{(1)} i_1 + \cdots + a_k^{(k)} i_k + \epsilon_{i_1...i_k}, \) \( i_k = 1, \ldots, n_s, \) \( s = 1, \ldots, k, \) \( (4.1) \) with \( a_s^{(s)} \overset{iid}{\sim} N(0, 1/\tau_s) \) for \( s \in \{1, \ldots, k\} \), \( \epsilon_{i_1...i_k} \overset{iid}{\sim} N(0, 1/\tau_e) \) 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: \( \mathbf{y} = (y_{i_1...i_k})_{i_1,...,i_k}, \mathbf{a}^{(s)} = (a_s^{(s)})_{i_s}, \mathbf{a} = (\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}), \mathbf{a}^{(s)} = (\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(s-1)}, \mathbf{a}^{(s+1)}, \ldots, \mathbf{a}^{(k)}). \) The standard Gibbs Sampler to sample from the posterior distribution \( \mathcal{L}(\mu, \mathbf{a} | \mathbf{y}) \) of Model C$k$ is defined as follows.

Sampler GS-crossed. At each iteration
1. sample $\mu$ from $\mathcal{L}(\mu | \mathbf{a}, \mathbf{y}),$ 
2. sample $\mathbf{a}^{(s)}$ from $\mathcal{L}(\mathbf{a}^{(s)} | \mu, \mathbf{a}^{(s-1)}, \mathbf{y})$ with $s$ going from 1 to $k.$
Model \( C_k \) 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 \( \bar{a}^{(s)} = \frac{1}{n_s} \sum_{i=1}^{n_s} a_i^{(s)} \) and \( \delta a^{(s)} = (a^{(s)} - \bar{a}^{(s)}) \),

\[
\bar{a}^{(s)} = \frac{1}{n_s} \sum_{i=1}^{n_s} a_i^{(s)} \quad \text{and} \quad \delta a^{(s)} = (a^{(s)} - \bar{a}^{(s)}) ,
\]

for each \( s \in \{1, \ldots, k\} \) and proved the following result.

**Theorem 4** (Papaspiliopoulos et al. (2019)). Let \( ((\mu, a^{(1)}(t), \ldots, a^{(k)}(t)))_{t=1}^{\infty} \) be the Markov chain generated by Sampler GS-crossed. Then the time-wise transformations \( ((\mu, \bar{a}^{(1)}(t), \ldots, \bar{a}^{(k)}(t)))_{t=1}^{\infty} \) and \( (\delta a^{(1)}(t), \ldots, \delta a^{(k)}(t))_{t=1}^{\infty} \) are \( (k + 1) \) independent Markov chains. Moreover, the rate of convergence of \( ((\mu, a^{(1)}(t)))_{t=1}^{\infty} \) is

\[
\rho = \max_{s \in \{1, \ldots, k\}} \frac{N \tau_e}{N \tau_e + n_s \tau_s}.
\]

Theorem 4 implies that the convergence properties of Sampler GS-crossed deteriorate as \( N \) increases because \( \max_{s \in \{1, \ldots, k\}} (N \tau_e)^{-1}(N \tau_e + n_s \tau_s) \) goes to 1 as \( N \rightarrow \infty \). 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 \( C_k \) 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 \( C_k \). 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

\[
(\mu, \beta^{(1)}, \beta^{(2)}) = (\mu, a^{(1)} + (1 - \lambda_1)\mu, a^{(2)} + (1 - \lambda_2)\mu), \quad \text{for } (\lambda_1, \lambda_2) \in \{0, 1\}^2.
\]

Each parametrization corresponds to a different Gibbs Sampler, which at each iteration updates \( \mu \) from \( \mathcal{L}(\mu|\beta^{(1)}, \beta^{(2)}, y) \), \( \beta^{(1)} \) from \( \mathcal{L}(\beta^{(1)}|\mu, \beta^{(2)}, y) \), and \( \beta^{(2)} \) from \( \mathcal{L}(\beta^{(2)}|\mu, \beta^{(1)}, y) \). 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 \( \rho \) is the rate of convergence (or its lower bound for \( \rho_{11} \)), as a function of \( \log(r_1/(1 - r_1)) \) and \( \log(r_2/(1 - r_2)) \).

**Theorem 5.** Let \( r_1 = \frac{N_e}{N_e + n_1 \tau_1} \) and \( r_2 = \frac{N_e}{N_e + n_2 \tau_2} \). Then we have

\[
\begin{align*}
\rho_{11} &= \max\{r_1, r_2\}, \\
\rho_{10} &= 1 - r_2(1 - r_1), \\
\rho_{01} &= 1 - r_1(1 - r_2), \\
\rho_{00} &= 1 + r_1 r_2 - \min\{r_1, r_2\}.
\end{align*}
\]

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 \((\lambda_1, \lambda_2)\) 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 \((\mu, a^{(1)}, \ldots, a^{(k)})\) in Model Ck are not identifiable, in the sense that the mapping \((\mu, a^{(1)}, \ldots, a^{(k)}) \to \mathcal{L}(y|\mu, a^{(1)}, \ldots, a^{(k)})\) is not injective. While this is
Gibbs Sampler and Multigrid Decompositions

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 \( C_k \) 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 \( (a_1^{(s)}, \ldots, a_{n_s}^{(s)}) \) weighted by some non-negative terms \( (w_1^{(s)}, \ldots, w_{n_s}^{(s)}) \) 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 \( (w_1^{(s)}, \ldots, w_{n_s}^{(s)}) \).

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

\[
\rho = \max_{s \in \{1, \ldots, k\}} \left( \frac{N \tau_c (1 - q_s)}{N \tau_c + n_s \tau_s} \right),
\]

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 \( s = 1, \ldots, k \). 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 \( s \in \{1, \ldots, k\} \) is given by

\[
\rho = \max_{s \in \{1, \ldots, k\}} \left( \frac{N \tau_c}{N \tau_c + n_s \tau_s} \right) \frac{n_s - 1}{n_s},
\]

while the rate of convergence of Sampler GS-crossed conditioned on \( \bar{a}^{(s)} = 0 \) for each \( s \in \{1, \ldots, k\} \) equals 0, i.e. the sampler produces i.i.d. draws from the posterior distribution \( \mathcal{L}(\mu, a|y, \bar{a}^{(1)} = \cdots = \bar{a}^{(k)} = 0) \). 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 \( s \in \{1, \ldots, k\} \) leads to a convergence rate that can still go to 1 as the data size 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 \( s = 1, \ldots, k \) and \( j = 1, \ldots, n_s \).
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 \(C_k\) in the context of count data.

**Model \(C_kP\)** (Poisson crossed-effects model).

\[
y_{i_1, \ldots, i_k} \sim \text{Poisson}(\mu a^{(1)}_{i_1} \cdots a^{(k)}_{i_k}) \quad i_s = 1, \ldots, n_s \text{ for } s \in \{1, \ldots, k\},
\]

with \(a^{(s)}_{i_s} \sim \text{Gamma}(\alpha_s, \beta_s)\) for \(s = 1, \ldots, k\) and \(\mu \sim \text{Gamma}(\alpha_\mu, \beta_\mu)\).

Consider sampling from the posterior distribution \(\mathcal{L}(\mu, a | y)\) of Model \(C_kP\) using the standard Gibbs Sampler that, similarly to Sampler GS-crossed, at each iteration updates \(\mu\) from \(\mathcal{L}(\mu | a, y)\) and then \(a^{(s)}\) from \(\mathcal{L}(a^{(s)} | \mu, a^{(\sim s)}, y)\) for \(s = 1, \ldots, k\).

We explore the extent to which the conclusions drawn from Theorem 6 apply also to Model \(C_kP\) by means of simulations. We consider the case \(k = 2\) with three different combinations of values of \(n_1\) and \(n_2\). The data \((y_{i_1, i_2})\) are generated from the model with \((a^{(1)}_{i_1})\) and \((a^{(2)}_{i_2})\) sampled from their prior distributions and \(\mu\) 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 sample 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 \(C_kP\), 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 \(C_kP\), such as Hamiltonian Monte Carlo.
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 \( \mu \), \( a_1^{(1)} \) and \( a_2^{(2)} \)) 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 \texttt{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 \( ((\mu, a)(t))_{t=1}^{\infty} \) induced by the Gibbs Sampler can be decomposed into \( (k + 1) \) independent Markov chains \( (\mu, \tilde{a}^{(1)}(t))_{t=1}^{\infty}, \ldots, (\tilde{a}^{(k)}(t))_{t=1}^{\infty} \), where \( \tilde{a}^{(s)} = \sum_{i_s} a^{(s)}_{i_s} \) and \( \tilde{a}^{(s)}_{i_s} = a^{(s)}_{i_s}/\tilde{a}^{(s)} \). In this case the rate of convergence of the original chain coincides with the one of \( ((\mu, \tilde{a}^{(1)}, \ldots, \tilde{a}^{(k)})(t))_{t=1}^{\infty} \), which evolves according to a \( (k + 1) \)-dimensional Gibbs Sampler with full conditionals given by:

\[
\mu|y, \tilde{a} \sim \text{Gamma}(\alpha_\mu + y, \beta_\mu + \prod_{s=1}^{k} \tilde{a}^{(s)}),
\]

\[
\tilde{a}^{(s)}|y, \mu, \tilde{a}^{(-s)} \sim \text{Gamma}(I\alpha_s + y, \beta_s + \mu \prod_{s \neq \ell} \tilde{a}^{(\ell)}) \quad \text{for } s \in \{1, \ldots, k\},
\]

where \( y = \sum_{i_1, \ldots, i_k} y_{i_1, \ldots, i_k} \). 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 work 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 \(\sigma_b^2\) and \(\sigma_e^2\), 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

\[
p(\mu) \propto 1, \quad \gamma_i \sim N(\mu, \tau_a^{-1}), \quad y_{ij} \sim N(\gamma_i, \tau_{e,i}^{-1}), \quad i = 1, \ldots, I; j = 1, \ldots, J_i,
\]

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 $\tau_e$. They showed that the convergence rates induced by the centred and non-centred parametrizations respectively are

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

(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 $(\mu, \beta_1, \ldots, \beta_I)$ defined by $I$ indicators $(\lambda_1, \ldots, \lambda_I) \in \{0, 1\}^I$ as $\beta_i = \gamma_i - \lambda_i \mu$, meaning that $\lambda_i$ 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 $(\lambda_1, \ldots, \lambda_I) \in \{0, 1\}^I$ is

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

(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 $\frac{\tau_a}{\tau_a + \tilde{\tau}_e}$ and $\frac{\tilde{\tau}_e}{\tau_a + \tilde{\tau}_e}$ 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 $(\lambda_1, \ldots, \lambda_I)$, 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 $\{1, \ldots, I\}$ use centred parametrization $\gamma_i$ 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 $\tilde{\lambda}_i = 1(\tau_a > \tilde{\tau}_i)$ for all $i$ from 1 to $I$. Then

$$\rho_{\tilde{\lambda}_1 \ldots \tilde{\lambda}_I} \leq \rho_{\lambda_1 \ldots \lambda_I} \quad \text{for any} \ (\lambda_1 \ldots \lambda_I) \in \{0, 1\}^I.$$

By (6.2), the strategy described in Corollary 3 ensures that $\rho_{\tilde{\lambda}_1 \ldots \tilde{\lambda}_I} \leq 1/2$. 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

$$p(\mu) \propto 1$$

$$\gamma_i \sim N(\mu, \sigma_{\gamma_i}^2) \quad i = 1, \ldots, I,$$

$$\eta_{ij} \sim N(\gamma_i, \sigma_{\eta_{ij}}^2) \quad j = 1, \ldots, J_i,$$

$$y_{ijk} \sim N(\eta_{ij}, \sigma_{y_{ijk}}^2) \quad k = 1, \ldots, K_{i,j},$$

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

$$r^{(i)}_{a,b} = \frac{\sigma^2_a}{\sigma^2_a + J_i^{-1} \sigma^2_b}, \quad \text{and} \quad r^{(i)}_{e,b} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{K^{-1}_{ij} \sigma^2_{e,ij}}{\sigma^2_{b,j} + K^{-1}_{ij} \sigma^2_{e,ij}}.$$  

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

$$\rho \leq 1 - \frac{1}{I} \sum_{i=1}^{I} r^{(i)}_{a,b} + \max_{i=1, \ldots, I} r^{(i)}_{a,b} r^{(i)}_{e,b}.$$  

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 $s \leq t$ and $s \geq t$ 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 ∈ \{0, \ldots, k - 1\} we denote by \(T_d = \{t ∈ T : ℓ(t) = d\}\) the collection of nodes at level \(d\). For example we have \(T_0 = \{t_0\}\) and \(T = \bigcup_{d=0}^{k-1} T_d\). Noisy observations will occur only at leaf nodes. The collection of leaf nodes is denoted as \(T_L = \{t ∈ T : ch(t) = \emptyset\}\). 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 NS\(k\) (\(k\)-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

\[
y_t^{(i)} \sim N(\gamma_t, 1/\tau_t^{(e)}) \quad t ∈ T_L, \tag{7.1}
\]

\[
\gamma_t \sim N(\gamma_{pa(t)}, 1/\tau_t) \quad t ∈ T\setminus t_0, \tag{7.2}
\]

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

We are interested in sampling from the posterior distribution of \(\gamma_T = (\gamma_t)_{t ∈ T}\) given some observations \(y = (y_t)_{t ∈ T_L}\). The centred parametrization \(\gamma_T\) of Model NS\(k\) leads to the following Gibbs Sampler.

**Sampler GS(\(\gamma_T\)).** Initialize \(\gamma_T(0)\) and then iterate the following kernel:

For \(d = 0, \ldots, k - 1\), sample \(\gamma_t(s+1)\) from \(p(\gamma_t|\gamma_{T_{d-1}}(s+1), \gamma_{T_{d+1}}(s), y)\) for all \(t ∈ T_d\), where \(p(\gamma_t|\gamma_{T_{d-1}}, \gamma_{T_{d+1}}, y) = p(\gamma_t|\gamma_{T\setminus t_0}, y)\) is the full conditional distribution of \(\gamma_t\) given by Model NS\(k\). When \(d\) equals 0 or \(k - 1\) the levels \(\gamma_{T_{d-1}}\) and \(\gamma_{T_{d+1}}\) have to be replaced by empty sets in the conditioning.

### 7.2 Non-centering and hierarchical reparametrizations

Model NS\(k\) 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 NS\(k\), define

\[
y_t^{(i)} \sim N\left(\sum_{r ≤ t} \alpha_r, 1/\tau_t^{(e)}\right) \quad t ∈ T_L, \tag{7.1}
\]

\[
\alpha_t \sim N(0, 1/\tau_t) \quad t ∈ T\setminus t_0, \tag{7.2}
\]

and assume a flat prior on \(\alpha_{t_0}\).

The non-centred parametrization \(α_T\) can be obtained as a linear transformation of the centred version \(\gamma_T\) of Model NS\(k\). More generally, we will consider the class of parametrizations that can be obtained by reparametrizing \(\gamma_T\) as follows.
**Definition 1** (Hierarchical reparametrizations). Let \( \gamma_T = (\gamma_t)_{t \in T} \) be a random vector with elements indexed by a tree \( T \). We say that \( \beta_T = (\beta_t)_{t \in T} \) is a hierarchical (linear) reparametrization of \( \gamma_T \) if

\[
\beta_t = \sum_{r \preceq t} \lambda_{tr} \gamma_r \quad t \in T,
\]

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 \( \beta_T = \Lambda \gamma_T \).

Using terminology from Papaspiliopoulos et al. (2003), we refer to the family of hierarchical reparametrizations of \( \gamma_T = (\gamma_t)_{t \in T} \) as partially non-centred parametrizations (PNCP) of Model NSk. Note that (7.3) does not span the space of all linear transformations of \( \gamma_T \). In fact \( \Lambda = (\lambda_{tr})_{r \preceq t, t \in T} \) can be thought as a \( |T| \times |T| \) matrix \( \Lambda = (\lambda_{tr})_{r \preceq t, t \in T} \) inducing a linear transformation of \( \gamma_T \) with the additional sparsity constraint of being zero on all elements \( \lambda_{tr} \) 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 \( \alpha_T \) of Example 1.

**Lemma 1.** If \( \beta_T \) is a hierarchical reparametrization of \( \gamma_T \), then also \( \gamma_T \) is a hierarchical reparametrization of \( \beta_T \).

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 \( \beta_T \) the corresponding Gibbs Sampler scheme \( GS(\beta_T) \) is defined analogously to \( GS(\gamma_T) \).

**Sampler GS(\beta_T).** Initialize \( \beta_T(0) \) and then iterate the following kernel:

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

Sampler \( GS(\beta_T) \) 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:

\[
\beta_r \perp \beta_t | \beta_{T \setminus \{r, t\}} \quad \text{unless} \ r \preceq t \text{ or } t \preceq r.
\]

Note that this is a weaker condition than the one holding for the centred parametrization \( \gamma_T \). In the latter case, the conditional independence graph corresponds exactly to the tree \( T \), meaning that if \( r \npreceq t \)

\[
\gamma_r \perp \gamma_t | \gamma_{T \setminus \{r, t\}} \quad \text{unless} \ r = pa(t) \text{ or } t = pa(r).
\]

Despite being weaker than (T), condition (H) still guarantees that all parameters at the same level are conditionally independent (thus simplifying the update of \( \beta_{T_r}|\beta_{T \setminus T_d} \)) and that the full conditional distribution of each \( \beta_t \) 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 $\beta_T$ satisfies (H) then any hierarchical re-parametrization of $\beta_T$ satisfies (H) too.

Corollary 4. Any PNCP $\beta_T$ of Model NSk satisfies (H).

7.3 Symmetry assumption

To provide a full analysis of the convergence properties of Sampler GS$(\beta_T)$ we need a symmetry assumption that we now define. Let $\rho_{tr}$ denote the partial correlation $\text{Corr}(\beta_t, \beta_r \mid \beta_{T\setminus\{t,r\}})$, namely

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

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

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

(7.4)

Equation (7.4) implies that at each step $X$ jumps from the current state $r$ to one of its children $t \in ch(r)$ choosing $t$ proportionally to the squared partial correlation between $\beta_r$ and $\beta_t$. Since $\ell(X_d) = d$ almost surely for all $d \in \{0, \ldots, k-1\}$ 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 $P(X_{\ell(t)} = t)$, $P(X_{\ell(t)} = t \mid X_{\ell(r)} = r)$ 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

$$\rho_{tr} = c_{t(r)\ell(t)} \sqrt{P(t|r)} \quad r \leq t,$$

(S)

and $\rho_{tr} = 0$ if $r \not\leq t$ and $t \not\leq r$. Note that $\rho_{tr}$ is invariant to coordinate-wise rescaling of $\beta_T$ and therefore both property (S) and the transition kernel of $X$ are invariant to rescalings. Therefore we can consider, without loss of generality, the following rescaled version of $\beta_T$ defined by

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

(7.5)

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

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

(\dot{S})

The rescaled version $\tilde{\beta}_T$ will be helpful to design an appropriate multigrid decomposition of $\beta_T$. Also, property (\dot{S}) is closed under symmetric hierarchical parametrizations.
**Definition 2** (Symmetric hierarchical reparametrizations). We say that $\beta_T = \Lambda \alpha_T$ is a symmetric hierarchical reparametrization of $\alpha_T$ if the coefficients of $\Lambda = (\lambda_{tr})_{t \leq \ell, t \in T}$ depend only on the levels of $r$ and $t$ in the hierarchy $T$.

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

$$y_{i_1,\ldots,i_{k-1},j} \sim N(\gamma_{(i_1,\ldots,i_{k-1})}^{(k-1)},1/\tau_e) \quad (i_1,\ldots,i_{k-1},j) \in [I_1] \times \cdots \times [I_{k-1}] \times [J],$$

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

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

Here $(\tau_1,\ldots,\tau_{k-1},\tau_e)$ are known precisions and the root parameter $\gamma^{(0)}$ is given a flat prior $p(\gamma^{(0)}) \propto 1$. For each $d \in \{1,\ldots,k-1\}$ 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 $y = (y_{i_1,\ldots,i_{k-1},j})_{i_1,\ldots,i_{k-1},j}$, satisfies $(S)$. In this case the random walk $X$ 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 $\tau_t = \prod_{r \in \text{ch}(t)} \tau_r$, for all $t \in T$ and $\tau^{(e)}_t = \frac{\tau_t}{\tau_1 \prod_{r \in \text{ch}(t)} \tau_r}$, where $(\tau_1,\ldots,\tau_k,\tau_e) \in \mathbb{R}^{k+1}$ 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 $\gamma_T$.

**Example 3** (Non-symmetric cases). In both Model Sk and Example 2 the auxiliary Markov chain $X$ 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 $X$ is not a natural random walk. In particular any instance of Model NSk such that

$$\sum_{r \in \text{ch}(t)} \rho^2_{tr} = c(t) \quad \text{for all } t \in T \setminus T_L,$$

for some $(k-1)$-dimensional vector $(c_0,\ldots,c_{k-2})$ 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 (S) still satisfies (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 (S) (after rescaling). Moreover, note that the exact analysis we will now provide for the Gibbs sampler on models satisfying (S) can be used to provide bound on general cases that do not satisfy (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 \( \beta_T \) satisfying (H) and (S). Our aim is to provide a transformation of \( \beta_T \) 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 \( \beta_{T_d} = (\tilde{\beta}_t)_{t \in T_d} \) by \( \beta^{(d)} \). 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 \( \phi^{(p)} : \mathbb{R}^T \rightarrow \mathbb{R}^T \) is defined as

\[
\phi^{(p)}_r \beta^{(d)} = \mathbb{E}[\beta_{X_d} | \beta_T, X_p = r] = \sum_{t \in T_d} \beta_t P(t|r) \quad r \in T_p, \tag{7.6}
\]

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

**Example 4** (Averaging operators in the symmetric case). Let \( \beta_T = \gamma_T \) be given by Model Sk. Then

\[
\phi^{(p)}_r \beta^{(d)} = \frac{1}{\prod_{t=p+1}^d I_t} \left( \sum_{t \in T_d : t \geq r} \beta_t \right) \quad r \in T_p.
\]

Given \( \phi \), we define the residual operators \( \delta^{(p)} : \mathbb{R}^T \rightarrow \mathbb{R}^T \) as \( \delta^{(p)} = (\delta^{(p)}_r)_{r \in T_p} \) with \( \delta^{(p)}_r : \mathbb{R}^T \rightarrow \mathbb{R} \) defined as

\[
\delta^{(p)}_r \beta^{(d)} = \phi^{(p)}_r \beta^{(d)} - \phi^{(p-1)}_{r \uparrow (r)} \beta^{(d)} \quad r \in T_p, \tag{7.7}
\]

for \( 1 \leq p \leq d \leq k - 1 \) and \( \delta^{(0)} \beta^{(d)} = \phi^{(0)} \beta^{(d)} \) 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 \( \beta_T \) be a Gaussian random vector satisfying (H) and (S). Then for any \( p \) and \( d \) with \( 0 \leq p \leq d \leq k - 1 \), for all
Theorem 10. The rate of convergence of chains described above as follows.

Let \( r \in T_p \) we have the identity

\[
E[\delta_r^{(p)}(\beta)^{d}] - E[\delta_r^{(p)}(\beta)^{d}] = \sum_{\ell \in \{p, ..., k-1\}\setminus d} c_{\delta_r^{(p)}}(\delta_r^{(p)}(\beta)^{\ell}) - E[\delta_r^{(p)}(\beta)^{\ell}] \cdot
\]

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

\[
\delta(p)\beta^{(d)}(s+1) \sim \mathcal{L}(\delta(p)\beta^{(d)}(s+1)_{\ell \leq d}, \delta(p)\beta^{(d)}(s)_{d < \ell \leq k-1}),
\]

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

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

Corollary 5. Let \( \beta_\tau \) satisfy (H) and (S). Then the rate of convergence of \( GS(\beta_\tau) \) is given by \( \max\{\rho_0, \ldots, \rho_{k-1}\} \) where for each \( p \in \{0, \ldots, k-1\} \), \( \rho_p \) is the rate of convergence of \( (\delta(p)\beta(s))_{s \in \mathbb{N}} \).

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 \( (\delta(p)\beta(s))_{s \in \mathbb{N}} \) is given by the largest modulus eigenvalue of \((I_{k-p} - L)^{-1}U\). Here \( I_{k-p} \) is the \( (k-p) \) dimensional identity matrix, while \( L \) and \( U \) are, respectively, the strictly lower and strictly upper triangular part of \((c_{\delta_r})_{d, \ell=p}^{k-1}\) with \( c_{\delta_r} \) given by (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 \( \delta(p)\beta \) for \( p = 0, \ldots, k-1 \) 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 \( C = (c_{\delta_r})_{d, \ell=p}^{k-1} \) is given by (S). This suggests that the convergence properties of the sub-chains will typically improve from that of \( (\delta(0)\beta(s))_{s \in \mathbb{N}} \) to that of \( (\delta(k-1)\beta(s))_{s \in \mathbb{N}} \) and that the rate of convergence of \( (\delta(0)\beta(s))_{s \in \mathbb{N}} \) will typically determine the rate of the whole sampler \( GS(\beta_\tau) \). 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 \( (\delta(0)\beta(s))_{s \in \mathbb{N}} \) to \( (\delta(k-1)\beta(s))_{s \in \mathbb{N}} \).
Theorem 11 (Ordering of rates for centred parametrization). Let $\gamma$ be a Gaussian vector satisfying (T) and ($\tilde{S}$) 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 $(\delta^{(0)}\gamma(s)),...,(\delta^{(k-1)}\gamma(s))$ satisfy

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

(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 $\gamma$ be a Gaussian vector satisfying (T) and ($\tilde{S}$). Then the rate of convergence of $GS(\gamma_T)$ is given by the largest squared eigenvalue of the $k$-dimensional matrix $C - I_k$, where $C = (c_{d\ell})_{d,\ell=0}^{k-1}$ is defined by ($\tilde{S}$) and $I_k$ is the $k$-dimensional matrix.

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

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

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

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

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 $S_3$.

Model $S_4$ (Symmetric 4-levels hierarchical model). Suppose

$$y_{ijkl} = \mu + a_i + b_{ij} + c_{ijk} + e_{ijkl},$$

(7.10)
where $i$, $j$, $k$ and $\ell$ run from 1 to $I$, $J$, $K$ and $L$ respectively and $c_{ijkl}$ are iid normal random variables with mean $\theta$ and variance $\sigma^2$. We employ a standard Bayesian model specification with $a_i \sim N(0, \sigma^2_{a_i})$, $b_{ij} \sim N(0, \sigma^2_{b_{ij}})$, $c_{ijkl} \sim N(0, \sigma^2_{c_{ijkl}})$ and a flat prior on $\mu$.

In order to fit Model S4 with a Gibbs Sampler like $GS(\beta_T)$, one could consider centering or non-centering each of the three levels $(a_i)_i$, $(b_{ij})_{ij}$ and $(c_{ijkl})_{ijkl}$. 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

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

where $\hat{\sigma}^2_1 = \frac{\sigma^2}{I}$, $\hat{\sigma}^2_2 = \frac{\sigma^2}{J}$, $\hat{\sigma}^2_3 = \frac{\sigma^2}{KL}$ and $\hat{\sigma}^2_4 = \frac{\sigma^2}{JLR}$. If $\lambda_1 = 1$ (i.e. using the non-centred parametrization $(a_i)_i$ at level 1) the rates are

$$\begin{align*}
\rho_{111} &= \max\{r_{1,4}, r_{2,4}, r_{3,4}\}, \\
\rho_{110} &= \max\{r_{1,3}, r_{2,3}, r_{4,3}\}, \\
\rho_{100} &= \max\{r_{1,2}, 1 - r_{2,3}r_{3,4}\}, \\
\rho_{101} &= \max\{r_{1,2}, 1 - r_{2,4}r_{4,3}\}.
\end{align*}$$

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 $\delta^{(0)}\beta$ is always the slowest chain for all centred and non-centred parametrisations (which can be checked by computing the rates of convergence of $\delta^{(1)}\beta$, $\delta^{(2)}\beta$ and $\delta^{(3)}\beta$ using Theorem 10 and comparing those to the ones of $\delta^{(0)}\beta$), 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 $(\hat{\sigma}^2_i)_{i=1}^4$. 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 $\gamma_t$ 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, 2016b).

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 ModelCkP 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 $\beta_T = (\beta_t)_{t \in T}$.

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. 19

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. 2

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/1033065202. 2, 10, 30, 31

Bass, M. R. and Sahu, S. K. (2016a). “A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC.” *Statistics and Computing*, 1–22. MR3687322. doi: https://doi.org/10.1007/s11222-016-9700-z. 2

Bass, M. R. and Sahu, S. K. (2016b). “A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC.” *Statistics and Computing*, 1–22. MR3687322. doi: https://doi.org/10.1007/s11222-016-9700-z. 30

Bhatia, R. (2013). *Matrix analysis*, volume 169. Springer Science & Business Media. MR1477662. doi: https://doi.org/10.1007/978-1-4612-0653-8. 27
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.

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

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).

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

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

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. MR365913. doi: https://doi.org/10.1214/17-EJS1236.

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.

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.

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

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.

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

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

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

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

Papaspiliopoulos, O. and Zanella, G. (2017). “A note on MCMC for nested multilevel regression models via belief propagation.” arXiv preprint arXiv:1704.06064.
Gibbs Sampler and Multigrid Decompositions

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/

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.*

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.

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.

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.

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.

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.

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

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

Whittaker, J. (1990). “Graphical models in applied multivariate analysis.” MR1112133.

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.

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.

Zanella, G. and Roberts, G. (2020). “Supplementary Material of “Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions”.” *Bayesian Analysis.* doi: https://doi.org/10.1214/20-BA1242SUPP.
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 (CRIStM), EP/R034710/1 (CoSInES), and EP/R018561/1 (Bayes for Health).