Collegio Carlo Alberto

Analysis of the Gibbs Sampler for Gaussian Hierarchical Models via multigrid decomposition

Giacomo Zanella

Gareth Roberts

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Analysis of the Gibbs Sampler for Gaussian Hierarchical Models via multigrid decomposition

Giacomo Zanella[*] and Gareth Roberts[†]

March 17, 2017

Abstract

We study the convergence properties of the Gibbs Sampler in the context of posterior distributions arising from Bayesian analysis of Gaussian hierarchical models. We consider centred and non-centred parameterizations as well as their hybrids including the full family of *partially non-centred* parameterizations. We develop a novel methodology based on multi-grid decompositions to derive analytic expressions for the convergence rates of the algorithm for an arbitrary number of layers in the hierarchy, while previous work was typically limited to the two-level case. Our work gives a complete understanding for the three-level symmetric case and this gives rise to approximations for the non-symmetric case. We also give analogous, if less explicit, results for models of arbitrary level. This theory gives rise to simple and easy-to-implement guidelines for the practical implementation of Gibbs samplers on conditionally Gaussian hierarchical models.

1   Introduction

MCMC is established as the computational workhorse of most Bayesian statistical analyses for complex models. For conditionally conjugate models, the Gibbs sampler remains the most natural algorithm of choice, particularly for hierarchical models which have proved arguably the most popular general class of models for Bayesian analysis. However convergence can be a major issue and it has long been recognised that reparameterisation can drastically affect the convergence rate and hence efficiency of the sampler, see for example Hills and Smith [1992].

The popularity of hierarchical models stems partly from their flexibility, interpretability, and the fact that easily implementable MCMC algorithms

---

[*]Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Roentgen 1, 20136 Milan, Italy. giacomo.zanella@unibocconi.it

[†]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. gareth.o.roberts@warwick.ac.uk

such as the Gibbs sampler can be used to simulate their posterior distributions [Gelfand and Smith, 1990, Smith and Roberts, 1993]. Moreover, empirical experience with Gibbs samplers for hierarchical models (see for example Gelfand et al. [1995]) suggests that the Gibbs sampler's convergence properties are aided substantially by hierarchical structure, suggesting a synbiosis between hierarchical models and the Gibbs sampler. While there is some theoretical work to investigate this phenomenon (see for example Roberts and Sahu [1997], Meng and Van Dyk [1997], Papaspiliopoulos et al. [2003], Jones and Hobert [2004], Papaspiliopoulos et al. [2007], Yu and Meng [2011], Bass and Sahu [2016b]) comparatively little is known about the link between hierarchical models and MCMC convergence, and almost nothing for models of hierarchical depth greater than two. The present paper offers the first steps towards such an understanding.

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

$$\lim_{s \to \infty} \frac{\|P^s f - \mathbb{E}_\pi[f]\|_{L^2(\pi)}}{r^s} = 0 \qquad \forall f \in L^2(\pi), \tag{1}$$

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

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

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

The theoretical results given in this paper apply to the Gibbs sampler for Gaussian hierarchical models where we are able to extend and improve substantially on existing literature [Roberts and Sahu, 1997, Yu and Meng, 2011, Bass and Sahu, 2016b] both in terms of generality of hierarchical structure and the availability of explicit rates. We shall study the connections between rates of convergence and parameterisation, for instance studying different options for centred and non-centred (also called *sufficient* and *ancillary* in Yu and Meng [2011]) parameterisations and their links to explicit rates of convergence of their corresponding Gibbs samplers.

In general, the Gibbs sampler can be elegantly described in terms of orthogonal projections [Amit, 1991, 1996, Diaconis et al., 2010], although it is rarely possible to extract practical convergence information for complex Gibbs samplers from this theory. However one important family in which explicit and useful results concerning Gibbs sampler rates of converge are

available is the family of multivariate Gaussian distributions. This paper will take advantage of this explicit framework to study rates of convergence of the Gibbs sampler as they vary with the hierarchical structure of the model. Our study goes substantially beyond existing studies of this problem Roberts and Sahu [1997], Yu and Meng [2011], Bass and Sahu [2016b], Gao and Owen [2016] and leads to simple and practical recommendations for parameterisation choice in Gibbs sampling for hierarchical models.

Our results can be readily used for models which are conditionally Gaussian, for example for the very popular context (see for example Gelman et al. [2013]) where there exist unknown variances at various levels of the hierarchy. In that case our optimality results describe the optimal updating strategies for the hierarchical mean structure conditional on the variances. For such Gaussian hierarchical models of sufficient complexity, block updating of the entire Gaussian component together is typically infeasible ($O(n^3)$ in the dimension (n) of the Gaussian to be updated).

Furthermore, the Gaussian linear models we study epitomise Bayesian multilevel modeling more generally, in the sense that many more elaborate models share the qualitative features of these more simple models. Moreover, many non-Gaussian hierarchical models can be well-approximated by Gaussian ones (for example 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.

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. homogenous 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. 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 that corresponding to the coarsest level and thus derive explicit expressions for the rates of convergence in symmetric contexts. For 3-level non-symmetric models, we give bounds on rates based on comparisons with related symmetric models. Moereover we describe partial non-centering strategies and describe *bespoke* strategies which optimise the degree of non-centering at each branch.

Section 4 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

3

some non-symmetric cases. We again demonstrate a fundamental multi-grid decomposition in Theorem 15 where the coarsest level chain converges the slowest, and we give explicit formulae for optimal partial non-centering strategies.

2   Three level hierarchical linear models

We shall begin with a detailed study of the three-level case, giving a fairly complete analysis for the case of symmetric models, which we shall follow up with some new results for the non-symmetric case. Therefore we shall begin by considering the following three-level Gaussian hierarchical model.

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

$$y_{ijk} = \mu + a_i + b_{ij} + \epsilon_{ijk}, \tag{2}$$

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

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

Sampler GS$(1, 1)$. *Initialize $\mu(0)$, $\boldsymbol{a}(0)$ and $\boldsymbol{b}(0)$) and then iterate*
   1. *Sample $\mu(s + 1)$ from $p(\mu|\boldsymbol{a}(s), \boldsymbol{b}(s), \boldsymbol{y})$;*
   2. *Sample $a_i(s + 1)$ from $p(a_i|\mu(s + 1), \boldsymbol{b}(s), \boldsymbol{y})$ for all i;*
   3. *Sample $b_{ij}(s + 1)$ from $p(b_{ij}|\mu(s + 1), \boldsymbol{a}(s + 1), \boldsymbol{y})$ for all i and j,*
where $p(\mu|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y})$, $p(a_i|\mu, \boldsymbol{b}, \boldsymbol{y})$ and $p(b_{ij}|\mu, \boldsymbol{a}, \boldsymbol{y})$ are the full conditionals of Model S3 (see Appendix A for explicit expressions).

The parametrization $(\mu, \mathbf{a}, \mathbf{b})$ induced by (2) is often referred to as *non-centred parametrization* (NCP) and it is contrasted with the *centred parametrization* (CP) obtained by replacing a_i and b_{ij} with $\gamma_i = \mu + a_i$ and $\eta_{ij} = \gamma_i + b_{ij}$ respectively. Under the centred parametrization $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ the model formulation becomes

$$y_{ijk} \sim N(\eta_{ij}, \sigma_e^2), \qquad \eta_{ij} \sim N(\gamma_i, \sigma_b^2), \qquad \gamma_i \sim N(\mu, \sigma_a^2), \qquad p(\mu) \propto 1. \tag{3}$$

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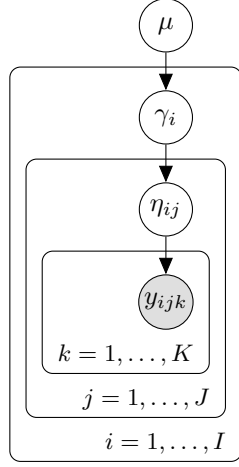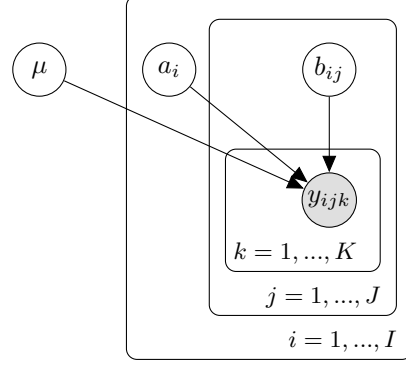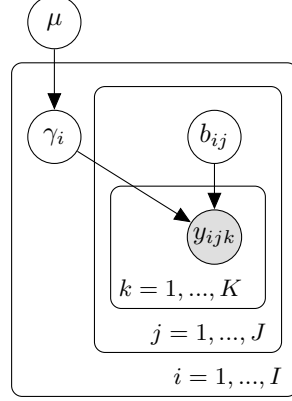
(a) Fully centred parametrization   (b) Fully non-centred parametrization
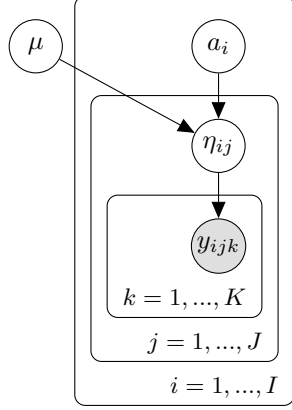
(c) Mixed parametrization: $(\mu, \mathbf{a}, \boldsymbol{\eta})$ (d) Mixed parametrization: $(\mu, \boldsymbol{\gamma}, \mathbf{b})$

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

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

   1. *Sample* $\mu(s+1)$ *from* $p(\mu|\boldsymbol{\gamma}(s), \boldsymbol{\eta}(s), \boldsymbol{y})$;

   2. *Sample* $\gamma_i(s+1)$ *from* $p(\gamma_i|\mu(s+1), \boldsymbol{\eta}(s), \boldsymbol{y})$ *for all* i;

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

where $p(\mu|\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{y})$, $p(\gamma_i|\mu, \boldsymbol{\eta}, \boldsymbol{y})$ *and* $p(\eta_{ij}|\mu, \boldsymbol{\gamma}, \boldsymbol{y})$ *are the full conditionals induced by* (3) *(see Appendix A for explicit expressions).*

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

2.1 Illustrative example

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

Figure 2 shows the mixing behaviour of the global mean μ and displays the potentially dramatic difference among mixing properties of the Gibbs Sampler under different parametrizations. Based on Figure 2 one would
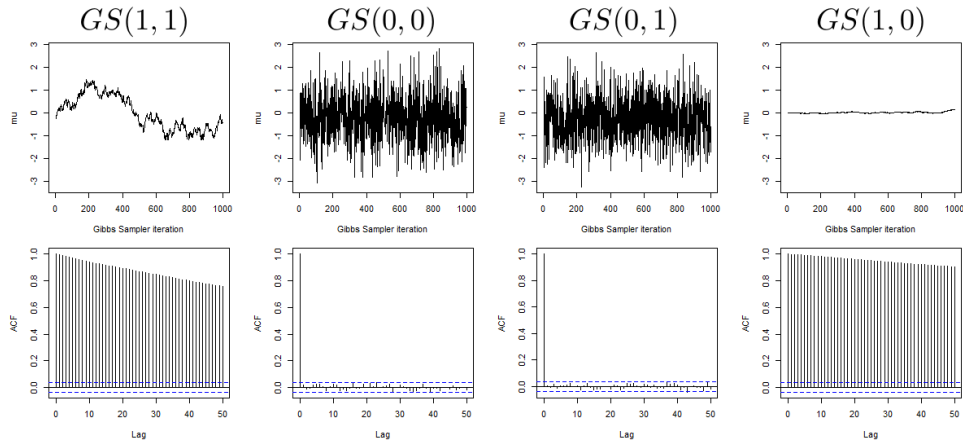

Figure 2: Mixing behaviour at level 0 (i.e. μ) under four different parametrizations. The ranges of the y-axes are constant across parametrizations.

certainly exclude using $GS(1,1)$ and $GS(1,0)$ to fit the model under consideration and may be tempted to deduce that both $GS(0,0)$ and $GS(0,1)$ lead to good mixing properties of the resulting chain. However, as an additional check, a cautious practitioner may also explore the mixing of the parameters at the first level, namely $\mathbf{a}$ and $\boldsymbol{\gamma}$. Figure 3 displays the behaviour of the global averages of such parameters, namely $a_. = \frac{\sum_i a_i}{I}$ and $\gamma_. = \frac{\sum_i \gamma_i}{I}$, in the first 3000 iterations. Again, we see a dramatic difference induced by
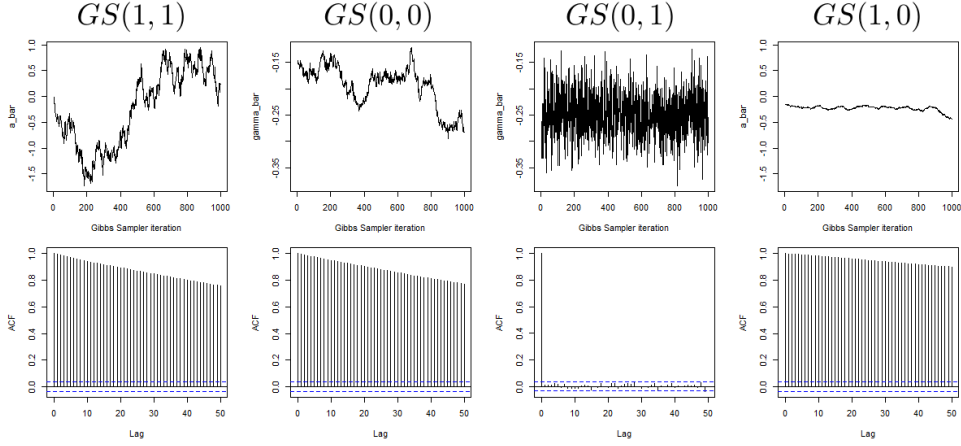
6

Figure 3: Mixing behaviour at level 1 (i.e. $a.$ and $\gamma.$) under four different parametrizations. The ranges of the y-axes are constant across parametrizations sharing the same parameters at level 1.

different parametrizations and, somehow surprisingly, we see that, despite having good mixing behaviour at level 0 (i.e. μ), $GS(0,0)$ displays very poor mixing behaviour at level 1 (i.e. $\boldsymbol{\gamma}$). It is then natural to explores also the mixing behaviour at level 2 and Figure 4 does 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 $GS(0,1)$
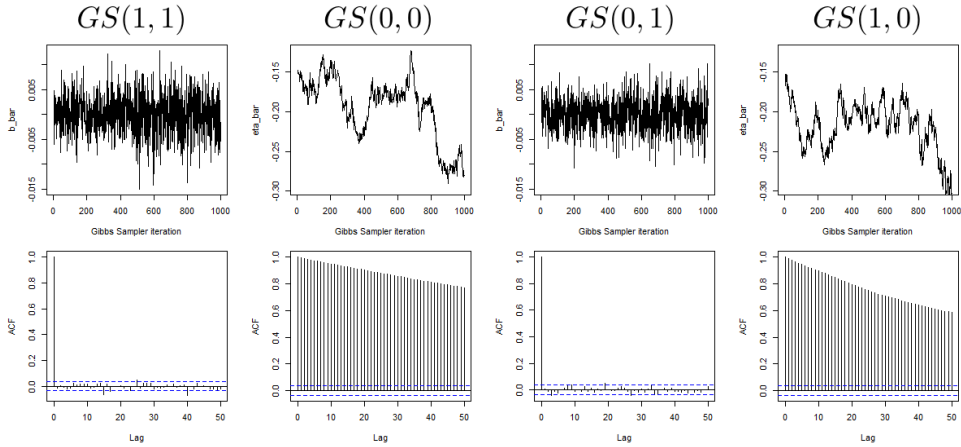
Figure 4: Mixing behaviour at level 2 (i.e. $b_{..}$ and $\eta_{..}$) under four different parametrizations. The ranges of the y-axes are constant across parametrizations sharing the same parameters at level 2.

are the only one achieving good mixing. Based on Figures 2, 3 and 4 it is natural to choose to fit the model using the sampler $GS(0,1)$ corresponding to the mixed parametrization $(\mu, \boldsymbol{\gamma}, \mathbf{b})$, as it is the only one providing a good mixing across all three levels.

7

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 Figure 2. In fact, while in two level model good mixing of the global hyperparameters such as μ typically indicates good global mixing, this is not true in other multi-level models. Indeed, it is legitimate to wonder whether diagnoses based only on the global means, like in Figures 2, 3 and 4, 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 3). 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, robust and theoretically grounded guidance in choosing a computationally efficient parametrization, given the huge impact it can have on computational performances. The centred parametrization is often a good choice but it can fail dramatically in case of highly noisy of weakly informative data and therefore it does not provide a "black-box" solution. In particular, models with more than two levels clearly shows a more complex structure than the two level ones. For example the simple illustrative example of this section shows that in three-level models both the fully centred and fully non-centred parametrizations can fail to produce computationally efficient samplers. The theoretical analysis developed in the next section will provide the required guidance. For example Theorem 4 below implies that, if $(I, J, K, \sigma_a, \sigma_b, \sigma_e) = (100, 100, 5, 10, 10^{-0.5}, 10)$, the L^2 rates of convergence (NB: numbers close to 1 mean slow convergence, see (1) and discussion thereof) of the various Gibbs Samplers under consideration are

$$\left(\rho_{(1,1)}, \rho_{(0,0)}, \rho_{(0,1)}, \rho_{(1,0)}\right) = (0.995, 0.998, 0.005, 0.999).$$

These numbers provide a quantitative and theoretically grounded description of the behaviour heuristically observed in this section and can be easily used to optimize performances (see e.g. Section 3.2).

3   Multigrid decomposition

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

by $GS(\boldsymbol{\beta})$. For example, in the NCP case $\boldsymbol{\beta}^{(0)} = \mu$, $\boldsymbol{\beta}^{(1)} = \mathbf{a}$, $\boldsymbol{\beta}^{(2)} = \mathbf{b}$ and $GS(\boldsymbol{\beta})$ coincides with $GS(1,1)$. First consider the map δ sending $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ to

$$\delta(\boldsymbol{\beta}) = \begin{pmatrix} \delta^{(0)}\boldsymbol{\beta} \\ \delta^{(1)}\boldsymbol{\beta} \\ \delta^{(2)}\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \delta^{(0)}\boldsymbol{\beta}^{(0)}, \ \delta^{(0)}\boldsymbol{\beta}^{(1)}, \ \delta^{(0)}\boldsymbol{\beta}^{(2)} \\ \delta^{(1)}\boldsymbol{\beta}^{(1)}, \ \delta^{(1)}\boldsymbol{\beta}^{(2)} \\ \delta^{(2)}\boldsymbol{\beta}^{(2)} \end{pmatrix}, \qquad (4)$$

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

$$\delta^{(0)}\beta^{(0)} = \beta^{(0)}, \qquad \delta^{(0)}\boldsymbol{\beta}^{(1)} = \beta_\cdot^{(1)}, \qquad \delta^{(0)}\boldsymbol{\beta}^{(2)} = \beta_{\cdot\cdot}^{(2)},$$

$$\delta^{(1)}\boldsymbol{\beta}^{(1)} = \left(\beta_1^{(1)} - \beta_\cdot^{(1)}, \ldots, \beta_I^{(1)} - \beta_\cdot^{(1)}\right), \ \delta^{(1)}\boldsymbol{\beta}^{(2)} = \left(\beta_{1\cdot}^{(2)} - \beta_{\cdot\cdot}^{(2)}, \ldots, \beta_{I\cdot}^{(2)} - \beta_{\cdot\cdot}^{(2)}\right),$$

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

where

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

It is easy to see that the map δ is a bijection between $\mathbb{R}^d$ and $\mathbb{R}^3 \times (\mathbb{R}^I)^* \times (\mathbb{R}^I)^* \times_{i=1}^I (\mathbb{R}^J)^*$, where $(\mathbb{R}^p)^* = \{(v_1, \ldots, v_p) \in \mathbb{R}^p : \sum_{i=1}^p v_i = 0\}$. The dimensionality of $\delta\boldsymbol{\beta}$ equals the one of $\boldsymbol{\beta}$, which is $1 + I + IJ$, because $\delta\boldsymbol{\beta}$ has $3 + 2I + IJ$ parameters and $2 + I$ constraints. It is worth noting that the three subspaces spanned by $(\delta^{(0)}\boldsymbol{\beta}^{(0)}, \delta^{(0)}\boldsymbol{\beta}^{(1)}, \delta^{(0)}\boldsymbol{\beta}^{(2)})$, $(\delta^{(1)}\boldsymbol{\beta}^{(1)}, \delta^{(1)}\boldsymbol{\beta}^{(2)})$ and $\delta^{(2)}\boldsymbol{\beta}^{(2)}$ respectively do not depend on the choice of parametrization $\boldsymbol{\beta}$. Thus the multi-grid decomposition is intrinsic to the model, and not dependent on the particular parameterization being considered.

Theorem 1. $(GS(\boldsymbol{\beta})$ *factorizes under* $\delta)$ *Let* $(\boldsymbol{\beta}(s))_s$ *be a Markov chain on* $\mathbb{R}^d$ *evolving according to* $GS(\boldsymbol{\beta})$. *Then* $(\delta^{(0)}\boldsymbol{\beta}(s))_s$, $(\delta^{(1)}\boldsymbol{\beta}(s))_s$ *and* $(\delta^{(2)}\boldsymbol{\beta}(s))_s$ *are three independent Markov chains evolving according to blocked Gibbs Samplers with blocks given by* $(\delta^{(0)}\boldsymbol{\beta}^{(0)}, \delta^{(0)}\boldsymbol{\beta}^{(1)}, \delta^{(0)}\boldsymbol{\beta}^{(2)})$, $(\delta^{(1)}\boldsymbol{\beta}^{(1)}, \delta^{(1)}\boldsymbol{\beta}^{(2)})$ *and* $\delta^{(2)}\boldsymbol{\beta}^{(2)}$ *respectively.*

Remark. *The posterior independence of* $\delta^{(0)}\boldsymbol{\beta}$, $\delta^{(1)}\boldsymbol{\beta}$ *and* $\delta^{(2)}\boldsymbol{\beta}$ *is a well-known fact following from properties of Gaussian distributions. The remarkable fact following from Theorem 1 is that also the Markov chains induced by the Gibbs Sampler are independent (note that the independence of the random vector under the target measure is a necessary but not sufficient condition for the independence of a corresponding MCMC scheme).*

Despite its simplicity, Theorem 1 provides a powerful tool to analyze the Markov chain of interest $(\boldsymbol{\beta}(s))_s$. In fact the factorization into independent
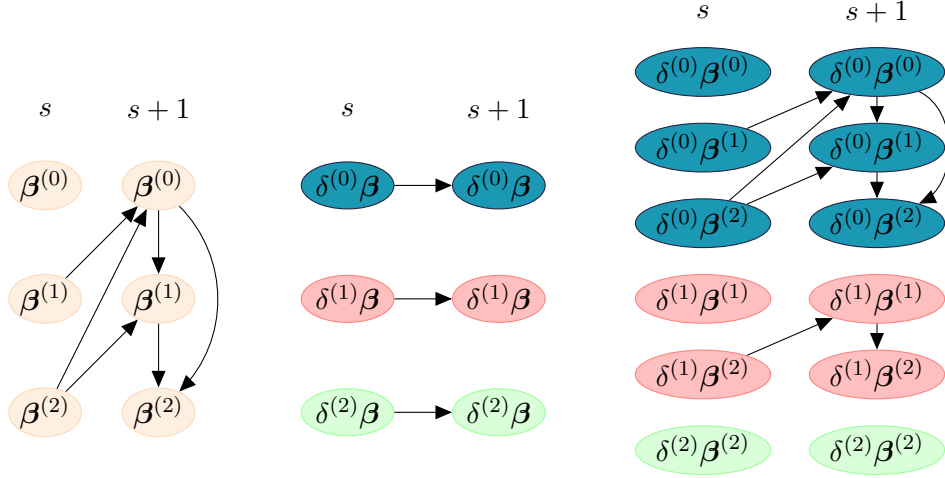
Figure 5: Illustration of Theorem 1. Left: transition from $\boldsymbol{\beta}(s)$ to $\boldsymbol{\beta}(s+1)$ in Sampler $GS(\boldsymbol{\beta})$. This is the common structure of a Gibbs Sampler with 3 blocks. Center: transition from $\delta\boldsymbol{\beta}(s)$ to $\delta\boldsymbol{\beta}(s+1)$ in Sampler $GS(\boldsymbol{\beta})$. There are three Markov chains, namely $(\delta^{(0)}\boldsymbol{\beta}(s))_s$, $(\delta^{(1)}\boldsymbol{\beta}(s))_s$ and $(\delta^{(2)}\boldsymbol{\beta}(s))_s$, evolving independently. Right: more detailed structure of the evolution of $(\delta^{(0)}\boldsymbol{\beta}(s))_s$, $(\delta^{(1)}\boldsymbol{\beta}(s))_s$ and $(\delta^{(2)}\boldsymbol{\beta}(s))_s$. These are Gibbs Samplers with respectively 3, 2 and 1 blocks.

Markov chains implies that the rate of convergence of $(\boldsymbol{\beta}(s))_t$ is simply given by the worst rate of convergence among $\delta^{(0)}\boldsymbol{\beta}(s)$, $\delta^{(1)}\boldsymbol{\beta}(s)$ and $\delta^{(2)}\boldsymbol{\beta}(s)$. Interestingly, the slowest chain is always the chain at the highest level, namely $\delta^{(0)}\boldsymbol{\beta}(s)$.

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

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

3.1 Explicit rates of convergence

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

Corollary 3. *(Rate of convergence of $GS(\boldsymbol{\beta})$) Given the notation of Theorem 1,*

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

While the original Markov chain $(\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ has $1 + I + IJ$ components and thus is typically high-dimensional, the skeleton chain $(\delta^{(0)}\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ is a

low-dimensional chain (namely 3-dimensional) amenable to direct analysis, for example using the tools developed in Roberts and Sahu [1997]. Therefore using Corollary 3 we can analytically evaluate the rates of convergence for the Gibbs Sampler under different parametrizations.

Theorem 4. *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*

$$\rho_{(0,0)} = 1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_b^2} \frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2} ,$$

$$\rho_{(1,1)} = \max\left\{ \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_e^2}, \frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2} \right\} ,$$

$$\rho_{(0,1)} = 1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_e^2} \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2} ,$$

$$\rho_{(1,0)} = \max\left\{ \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2 + \tilde{\sigma}_b^2}, \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2 + \tilde{\sigma}_e^2} \right\} ,$$

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

Theorem 4 provides explicit and informative formulas regarding the interaction between choice of parametrization and resulting efficiency of the Gibbs Sampler for Model . Figure 6 summarizes graphically the dependence
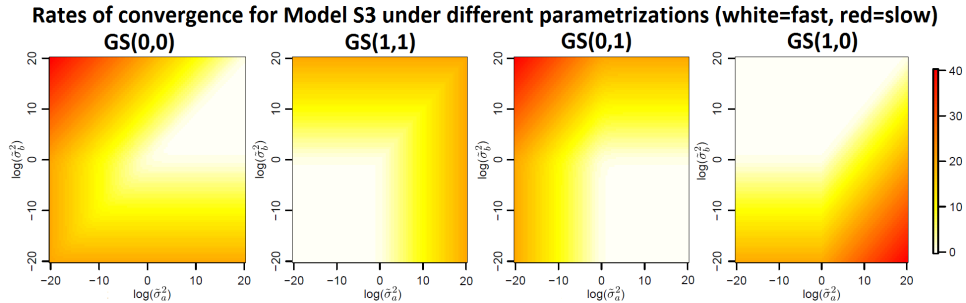
Figure 6: Plot of rates of convergence for three-levels Gaussian hierarchical models under different parametrizations. Color levels correspond to values of $\log(1 - \rho)$, where ρ is the rate of convergence, as a function of $\log(\tilde{\sigma}_a^2)$ and $\log(\tilde{\sigma}_b^2)$ for fixed $\log(\tilde{\sigma}_e^2) = 0$.

of the converge rates of different parametrizations from the values of the variances of various levels.

3.2 Conditionally optimal parametrization

A natural and practically relevant question is what is the optimal parametrization (among the four possible choices $(\mu, \mathbf{a}, \mathbf{b})$, $(\mu, \boldsymbol{\gamma}, \mathbf{b})$, $(\mu, \mathbf{a}, \boldsymbol{\eta})$ and $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$)

11

as a function of the normalized variance components $(\tilde{\sigma}_a^2, \tilde{\sigma}_b^2, \tilde{\sigma}_e^2)$. Using the formulas of Theorem 4 we can obtain a simple and easily interpretable answers.

Corollary 5 (Optimal parametrization for Model S3). *The rate of convergence of the Gibbs Sampler targeting Model S3 is minimized by the following choice of parametrization:*

- *if $\tilde{\sigma}_b^2 \geq \tilde{\sigma}_e^2$ use a centred parametrization $\boldsymbol{\eta}$ at the lowest level, while if $\tilde{\sigma}_b^2 < \tilde{\sigma}_e^2$ use a non-centred parametrization $\boldsymbol{b}$,*

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

The table in Figure 7 provides a graphical representation of the decision process. This simple rule guarantees that the resulting Gibbs Sampler has a

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

Figure 7: Optimal parametrization for 3-levels hierarchical models as a function of the normalized variance components.

rate of converges smaller than $\frac{2}{3}$ (thus guaranteeing a high sampling efficiency for fixed variances). The optimized convergence rate equals $\frac{2}{3}$ if and only if $\tilde{\sigma}_a^2 = \tilde{\sigma}_b^2 + \tilde{\sigma}_e^2$ and $\tilde{\sigma}_b^2 = \tilde{\sigma}_e^2$ (in which case all parametrizations are equivalent).

Figure 7 implies that the choice of parametrization of a given level (i.e. whether it is computationally convenient to use a center or non-center 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. This results extend previous intuition for the two-level case (e.g. Papaspiliopoulos et al. [2003]) to deeper hierarchical levels (in this case three levels).

Corollary 5 allows for simple and effective strategies to ensure high sampling efficiency in practical implementations of Gibbs Sampling for Model S3 in the case of unknown variances. Common implementations choose a parametrization of the gaussian component, say $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$, and alternate sampling $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ with $GS(\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ and $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$ (for which direct sampling is possible and computationally cheap). Indeed it is sufficient to include an "if" statement in the Gibbs Sampling implementation, after sampling $(\sigma_a, \sigma_b, \sigma_e)|\boldsymbol{\beta}$, to choose the optimal parametrization $\boldsymbol{\beta}$ given $(\sigma_a, \sigma_b, \sigma_e)$ according to Figure 7. This will ensure that the sampling step $\boldsymbol{\beta}|(\sigma_a, \sigma_b, \sigma_e)$ will have a high efficiency. This kind of strategies will be particularly robust and effective when there is little prior information on

the values of $(\sigma_a, \sigma_b, \sigma_e)$, or during the transient phase of the MCMC when $(\sigma_a, \sigma_b, \sigma_e)$ may assume unexpected values, or when the variance components are not well identified in the posterior. In all these situations it will be difficult to choose a priori a good parametrization and therefore optimizing it on-the-fly will be a much more robust and effective strategy. Note that the process of choosing the optimal parametrization (Figure 7) has a computational cost which is negligible compared to a Gibbs Sampling sweep.

3.3 Bounds for centred parametrization in the non-symmetric case

The approach of Section 3 provides an elegant decomposition into independent Markov chains that allows to evaluate analytically the convergence rates of the Gibbs Sampler for Model S3. However, Model S3 has two over-restrictive assumptions. First it assumes the variances terms σ_b^2 and σ_e^2 to be constant across the hierarchy (i.e. not to depend on i and j) and secondly, more crucially, it assumes the study under consideration to be perfectly balanced (i.e. J and K not depending on i and j). In this section we consider the following more general case.

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

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

where variance components are assumed to be known.

Although the multigrid factorization of Theorem 1 does not apply directly to Model NS3, it can still be used to obtain upper and lower bounds on the rates of convergence.

Theorem 6. *Given an instance of Model NS3 we define*

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

Then if

$$
\min_{i=1,\ldots,I} r_{a,b}^{(i)} \geq \max_{i=1,\ldots,I} r_{a,b}^{(i)} r_{e,b}^{(i)} \tag{5}
$$

the rate of convergence of the Gibbs Sampler with centred parametrization $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ satisfies

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

13

The results of Theorem 6 suggest that as the number of datapoints increase the efficiency of the Gibbs sampler with centred parametrization increases. In fact, as K_{ij} increases assumption (5) is eventually satisfied and the bound on the convergence rate goes to 0 as J_i and K_{ij} increase. Theorem 5 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].

3.4 Handling heterogeneity with bespoke parametrizations

In non-symmetric cases such as Model NS3 all parametrizations previously considered, namely $(\mu, \mathbf{a}, \mathbf{b})$, $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$, $(\mu, \boldsymbol{\gamma}, \mathbf{b})$ and $(\mu, \mathbf{a}, \boldsymbol{\eta})$, may be computationally inefficient. In this more general scenario, the computationally optimal strategy will involve centering some branches of the hierarchy and non-centering others. We will refer to these strategies as *bespoke parametrizations*. In order to design a bespoke parametrization, a question of particular interest is whether the optimal choice on each branch of the hierarchy can be taken independently of other branches or not. In the former case, the optimization process would have a local nature and the resulting algorithm could be implemented easily and efficiently, while in the latter case (optimal decision on each branch influenced by other branches) the global nature of the optimization problem would impose implementation challenges.

While the rate of convergence of the general three level case (Model NS3) is intractable, the two-level case offers enough tractability and yet richness to provide insight in the choice of optimal heterogeneous parametrizations. We therefore 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*

$$
\begin{aligned}
p(\mu) &\propto 1 \\
\gamma_i &\sim N(\mu, 1/\tau_a) & i &= 1, \ldots, I, \\
y_{ij} &\sim N(\gamma_i, 1/\tau_{e,i}) & j &= 1, \ldots, J_i,
\end{aligned}
$$

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

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

$$
\rho_{CP} = \frac{\tau_a}{\tau_a + \tilde{\tau}_e} \quad \text{and} \quad \rho_{NCP} = \frac{\tilde{\tau}_e}{\tau_a + \tilde{\tau}_e}, \tag{6}
$$

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 λ_i equals 0 if component i is centred and 1 if it is non-centred.

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

$$\rho_{\lambda_1 \ldots \lambda_I} = \frac{\sum_{i\,:\,\lambda_i=1} \tilde{\tau}_i \frac{\tilde{\tau}_i}{\tilde{\tau}_i + \tau_a} + \sum_{i\,:\,\lambda_i=0} \tau_a \frac{\tau_a}{\tilde{\tau}_i + \tau_a}}{\sum_{i\,:\,\lambda_i=1} \tilde{\tau}_i + \sum_{i\,:\,\lambda_i=0} \tau_a} \,, \tag{7}$$

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

Equation (7) shows that in the non-symmetric case, the GS rate of convergence is given by a weighted average of the precision ratios $\frac{\tau_a}{\tilde{\tau}_i + \tau_a}$ and $\frac{\tilde{\tau}_i}{\tilde{\tau}_i + \tau_a}$ depending on whether each component is centred or not. This has clear analogies with the symmetric case in (6). The weights in the average of (7) are themselves a function of $(\lambda_1, \ldots, \lambda_I)$, thus introducing dependence across components in terms of centering and the overall convergence rate. However, starting from (7) we can see that for any $i \in \{1, \ldots, I\}$

$$\rho_{\lambda_1 \ldots \lambda_{i-1} 0 \lambda_{i+1} \ldots \lambda_I} - \rho_{\lambda_1 \ldots \lambda_{i-1} 1 \lambda_{i+1} \ldots \lambda_I} = \frac{\frac{\tau_a \tilde{\tau}_i}{\tau_a + \tilde{\tau}_i}(\tau_a - \tilde{\tau}_i)}{(\rho_{-i} + \tau_a)(\rho_{-i} + \tilde{\tau}_i)} \,, \tag{8}$$

where $\rho_{-i} = \sum_{\ell \neq i\,:\,\lambda_\ell=0} \tilde{\tau}_\ell \frac{\tilde{\tau}_\ell}{\tilde{\tau}_\ell + \tau_a} + \sum_{\ell \neq i\,:\,\lambda_\ell=1} \tilde{\tau}_a \frac{\tilde{\tau}_a}{\tilde{\tau}_\ell + \tau_a} \geq 0$. Equation 8 implies that $\rho_{\lambda_1 \ldots \lambda_{i-1} 0 \lambda_{i+1} \ldots \lambda_I} > \rho_{\lambda_1 \ldots \lambda_{i-1} 1 \lambda_{i+1} \ldots \lambda_I}$ if and only if $\tau_a > \tilde{\tau}_i$, which in turn implies the following corollary.

Corollary 8. *Let $\bar{\lambda}_i = \mathbb{1}(\tau_a > \tilde{\tau}_i)$ for all i from 1 to I. Then*

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

Corollary 8 implies that in Model NS2, the choice of optimal parametrization in each branch of the three can be carried out independently following the simple rule: for each i in $\{1, \ldots, I\}$ use centred parametrization γ_i if and only if $\tau_a \leq J_i \tau_{e,i}$, otherwise use a non-centred parametrization $a_i = \gamma_i - \mu$.

As mentioned before, the non-symmetric three level case (Model NS3) is not easily tractable and therefore we do not have neat results such as Corollary 8 regarding the design of bespoke parametrizations. However, motivated by Corollary 5, Corollary 8 and Theorem 10 below we can design an heuristically motivated criteria to design efficient bespoke parametrizations. For each $i = 1, \ldots, I$, let λ_i be the non-centering indicator of the i-th parameter (i.e. $\beta_i = \gamma_i - \lambda_i \mu$) and for each $j = 1, \ldots, J_i$ let λ_{ij} be the non-centering indicator of the ij-th parameter (i.e. $\beta_{ij} = \eta_{ij} - \lambda_{ij} \gamma_i$). The heuristic choice for $(\lambda_i)_i$ and $(\lambda_{ij})_{ij}$ is the following:

$$\lambda_{ij} = \mathbb{1}\left(\sigma_{b,i}^2 < \frac{\sigma_{e,ij}^2}{K_{ij}}\right) \quad \text{and} \quad \lambda_i = \mathbb{1}\left(\frac{1}{\sigma_a^2} > \sum_{j=1}^{J_i} \frac{1}{\sigma_{b,i}^2 + K_{ij}^{-1}\sigma_{e,ij}^2}\right). \tag{9}$$

It is easy to see that in the symmetric case equation (9) coincides with the criteria derived in Section 3.2 from the explicit rates of convergence analysis. In the non-symmetric case, these heuristic criteria can be motivated by observing that the non-centering indicators λ_i and λ_{ij} defined by (9) equal 1 or 0 depending on whether the corresponding optimal partial non-centering coefficients $\lambda_i^{(1,0)}$ and $\lambda_{ij}^{(2,1)}$ (see Theorem 9 below) are above or below a half. Note also that the definition of λ_{ij} in (9) is consistent with the non-symmetric two level case of Corollary 8.

3.5   Optimal partial non-centering

The centred and non-centred parametrizations can be seen as extrema of a continuum of parametrizations, namely partially non-centred parametrizations (PNCP) . PNCP can be obtained from the centred (or non-centred) parametrization by linear transformations which preserve the hierarchical structure. For example, in the symmetric context of Model S3, the family of PNCP can be obtained from the centred parametrization $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ as $\beta^{(0)} = \mu$, $\beta_i^{(1)} = \gamma_i - \lambda^{(1,0)}\mu$, and $\beta_{ij}^{(2)} = \eta_{ij} - \lambda^{(2,1)}\gamma_i - \lambda^{(2,0)}\mu$, for some real valued non-centering parameters $(\lambda^{(1,0)}, \lambda^{(2,0)}, \lambda^{(2,1)})$. In the more general non-symmetric context of Model NS3 one may allow the non-centering parameters to depend on the location in the hierarchical tree, thus obtaining

$$
\begin{aligned}
\beta^{(0)} &= \mu, \\
\beta_i^{(1)} &= \gamma_i - \lambda_i^{(1,0)}\mu, \\
\beta_{ij}^{(2)} &= \eta_{ij} - \lambda_{ij}^{(2,1)}\gamma_i - \lambda_{ij}^{(2,0)}\mu,
\end{aligned}
\tag{10}
$$

for some real valued $(\lambda_i^{(1,0)})_i$ and $(\lambda_{ij}^{(2,0)}, \lambda_{ij}^{(2,1)})_{ij}$.

Theorem 9. *Let $(\mu, \boldsymbol{\gamma}, \boldsymbol{\eta})$ be given by Model NS3 conditioned on some value of $\boldsymbol{y}$. If $(\lambda_i^{(1,0)})_i$ and $(\lambda_{ij}^{(2,0)}, \lambda_{ij}^{(2,1)})_{ij}$ are defined as*

$$
\lambda_i^{(1,0)} = \frac{\frac{1}{\sigma_a^2}}{\frac{1}{\sigma_a^2} + \sum_{j=1}^{J_i} \frac{1}{\sigma_{b,i}^2 + K_{ij}^{-1}\sigma_{e,ij}^2}}
$$

$$
\lambda_{ij}^{(2,0)} = 0 \quad and \quad \lambda_{ij}^{(2,1)} = \frac{K_{ij}^{-1}\sigma_{e,ij}^2}{\sigma_{b,i}^2 + K_{ij}^{-1}\sigma_{e,ij}^2} \,,
$$

then the resulting $(\beta^{(0)}, \beta_i^{(1)}, \beta_{ij}^{(2)})$ is a collection of independent random variables.

Theorem 9 implies that appropriate partial non-centering leads to i.i.d. samples even in 3 level models. While any Gaussian random vector can be factorized into independent random variables via some linear transformation

16

(e.g. a Cholesky decomposition), the interesting aspect of Theorem 9 is that for hierarchical models such factorization can be obtained within the class of hierarchical reparametrizations.

4   Hierarchical linear models with arbitrary number of levels

In this section we consider Gaussian hierarchical models with an arbitrary number of levels, namely k levels. We refer to the highest level of the hierarchy (i.e. the one furthest away from the data) as level 0, down to level $k-1$ being 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 $k = 3$.

4.1   Model formulation

In order to allow for more generality and keep the notation concise, in this section we will use a graphical models notation. In particular T will denote a finite tree with k levels and root $t_0 \in T$. For each node $t \in T$ we will denote by $pa(t)$ the unique parent of t and by $ch(t)$ the collection of children of t. Moreover we write $s \preceq t$ and $s \succeq t$ if s is respectively an ancestor or a descendent of t (with s and t possibly being equal) while $s \prec t$ and $s \succ t$ denote the same notions with the additional condition of $s \neq t$. For each node $t \in T$ we denote by $\ell(t)$ the level of node t (i.e. its distance from t_0). For each $d \in \{0, \dots, k-1\}$ we denote by $T_d = \{t \in T : \ell(t) = d\}$ the collection of nodes at level d. For example we have $T_0 = \{t_0\}$ and $T = \cup_{d=0}^{k-1} T_d$. Noisy observations will occur only at leaf nodes. The collection of leaf nodes is denoted as $T_L = \{t \in T : ch(t) = \emptyset\}$. For simplicity we assume that all leaf nodes are at level $k-1$, i.e. $T_L = T_{k-1}$, although this assumption could be easily relaxed allowing some branches to be longer than others.

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

$$y_t^{(i)} \sim N(\gamma_t, 1/\tau_t^{(e)}) \qquad\qquad t \in T_L, \qquad\qquad (11)$$

$$\gamma_t \sim N(\gamma_{pa(t)}, 1/\tau_t) \qquad\qquad t \in T\backslash t_0, \qquad\qquad (12)$$

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

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

Sampler GS($\boldsymbol{\gamma}_T$). *Initialize $\boldsymbol{\gamma}_T(0)$ and then iterate the following kernel:*
For $d = 0, \ldots, k-1$

- *Sample $\gamma_t(s+1)$ from $p(\gamma_t | \boldsymbol{\gamma}_{T_{d-1}}(s+1), \boldsymbol{\gamma}_{T_{d+1}}(s), \boldsymbol{y})$ for all $t \in T_d$,*

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

4.2 Non centering and hierararchical reparametrizations

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

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

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

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

and assume a flat prior on α_{t_0}.

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

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

$$\beta_t = \sum_{r \preceq t} \lambda_{tr} \gamma_r \qquad t \in T, \tag{13}$$

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 (13) by $\boldsymbol{\beta}_T = \Lambda \boldsymbol{\gamma}_T$.

Analogously to Section 3.5, we refer to the family of hierarchical reparametrizations of $\boldsymbol{\gamma}_T = (\gamma_t)_{t \in T}$ as *partially non-centred parametrizations* (PNCP) of Model NSk. Note that (13) does not span the space of all linear transformations of $\boldsymbol{\gamma}_T$. In fact $\Lambda = (\lambda_{tr})_{r \preceq t, t \in T}$ can be thought as a $|T| \times |T|$ matrix $\Lambda = (\lambda_{tr})_{r, t \in T}$ inducing a linear transformation of $\boldsymbol{\gamma}_T$ with the additional sparsity constraint of being zero on all elements λ_{tr} such that $r \not\preceq t$. The following Lemma shows that the definition of the class of PNCP does not depend on the starting parametrization used to formulate Model NSk. For example, one could equivalently define the class of PNCP of Model NSk as the collection of hierarchical reparametrization of the non-centred parametrization $\boldsymbol{\alpha}_T$ of Example 1.

Lemma 10. *If $\boldsymbol{\beta}_T$ is a hierarchical reparametrization of $\boldsymbol{\gamma}_T$, then also $\boldsymbol{\gamma}_T$ is a hierarchical reparametrization of $\boldsymbol{\beta}_T$.*

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

Sampler $GS(\boldsymbol{\beta}_T)$. *Initialize $\boldsymbol{\beta}_T(0)$ and then iterate the following kernel:*
For $d = 0, \ldots, k-1$

- *Sample $\beta_t(s+1)$ from $p(\beta_t | (\boldsymbol{\beta}_{T_p}(s+1))_{0 \leq p < d}, (\boldsymbol{\beta}_{T_p}(s))_{d < p \leq k-1}, \boldsymbol{y})$ for all $t \in T_d$,*

where $p(\beta_t | (\boldsymbol{\beta}_{T_p})_{0 \leq p < d}, (\boldsymbol{\beta}_{T_p})_{d < p \leq k-1}, \boldsymbol{y}) = p(\beta_t | \boldsymbol{\beta}_{T \setminus t}, \boldsymbol{y})$ is the full conditional distribution of β_t given by Model NSk.

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

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

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

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

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

Lemma 11. *Property (H) is closed under hierarchical re-parametrizations, meaning that if $\boldsymbol{\beta}_T$ satisfies (H) then any hierarchical re-parametrization of $\boldsymbol{\beta}_T$ satisfies (H) too.*

Corollary 12. *Any PNCP $\boldsymbol{\beta}_T$ of Model NSk satisfies (H).*

4.3 Symmetry assumption

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

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

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

$$P(X_{\ell+1} = t \,|\, X_\ell = r) = \frac{\rho_{tr}^2}{\sum_{t' \in ch(r)} \rho_{t'r}^2} \mathbb{1}(t \in ch(r)) \,. \tag{14}$$

Equation (14) implies that at each step the Markov chain $\mathbf{X}$ jumps from the current state r to one of its children $t \in ch(r)$ choosing t proportionally to the squared partial correlation between β_r and β_t. Since $\ell(X_d) = d$ almost surely for all $d \in \{0, \dots, k-1\}$ we can use the following simplified notation: for any t and r in T we use $P(t)$, $P(t|r)$ and $P(t \cap r)$ to denote respectively $P(X_{\ell(t)} = t)$, $P(X_{\ell(t)} = t \,|\, X_{\ell(r)} = r)$ and $P(X_{\ell(t)} = t \cap X_{\ell(r)} = r)$.

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

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

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

$$\tilde{\beta}_t = \beta_t \sqrt{\frac{Q_{tt}}{P(t)}} \qquad\qquad t \in T \,. \tag{15}$$

Given (15), condition (S) can be written, in terms of the precision matrix of $\tilde{\boldsymbol{\beta}}_T = (\tilde{\beta}_t)_{t \in T}$ as

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

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

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

Lemma 13. *Property ($\tilde{\text{S}}$) is closed under symmetric hierarchical reparametrizations, meaning that if $\tilde{\boldsymbol{\beta}}_T$ satisfies ($\tilde{\text{S}}$) then any symmetric hierarchical reparametrization of $\tilde{\boldsymbol{\beta}}_T$ satisfies ($\tilde{\text{S}}$) 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*

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

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

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

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 $\mathbf{y} = (y_{i_1,\ldots,i_{k-1},j})_{i_1,\ldots,i_{k-1},j}$, satisfies (S). In this case the random walk $\mathbf{X}$ defined by (14) coincides with the natural random walk going through T.

Example 2 (Weakly symmetric case). *Another special case of Model NSk satisfying (S) is given by the case of a general tree T and precision terms defined as $\tau_t = \frac{\tau_{\ell(t)}}{\prod_{s \prec t} |ch(s)|}$ for all $t \in T$ and $\tau_t^{(e)} = \frac{\tau_e}{n_t \prod_{s \prec t} |ch(s)|}$, 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 $\boldsymbol{\gamma}_T$.*

Example 3 (Non-symmetric cases). *In both cases previously considered (Model Sk and Example 2) the auxiliary Markov chain $\mathbf{X}$ defined in (14) follows a natural random walk, in the sense that at each time the chain chooses the next state uniformly at random among children nodes. However, condition (S) is also satisfied by non-symmetric cases where $\mathbf{X}$ is not a natural random walk. In particular any instance of Model NSk such that*

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

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 (14)).*

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 13 shows that any symmetric hierarchical reparametrization of a vector satisfying (S̃) still satisfies (S̃). 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 (S̃) (after rescaling).

Moreover, note that the exact analysis we will now provide for the Gibbs sampler on models satisfying (S̃) can be used to provide bound on general cases that do not satisfy (S̃) (see for example Theorem 6).

4.4   Multigrid decomposition

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

$$
\begin{aligned}
\phi_r^{(p)} \boldsymbol{\beta}^{(d)} =& \mathbb{E}[\beta_{X_d} | \boldsymbol{\beta}_T, X_p = r] && r \in T_p \\
=& \sum_{t \in T_d} \beta_t P(t|r) && (16)
\end{aligned}
$$

where $\mathbf{X} = (X_\ell)_{\ell=0}^{k-1}$ is the Markov chain defined by (14). Loosely speaking $\phi^{(p)} \boldsymbol{\beta}^{(d)} = \mathbb{E}[\beta_{X_d} | \boldsymbol{\beta}_T, X_p]$ can be interpreted as the averages of $\boldsymbol{\beta}^{(d)}$ at the coarseness corresponding to the p-th level of the hierarchy. In particular $\phi^{(d)} \boldsymbol{\beta}^{(d)} = \boldsymbol{\beta}^{(d)}$ and $\phi_{t_0}^{(0)} \boldsymbol{\beta}^{(d)} = \mathbb{E}[\beta_{X_d} | \boldsymbol{\beta}_T]$.

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

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

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

$$
\delta_r^{(p)} \boldsymbol{\beta}^{(d)} = \phi_r^{(p)} \boldsymbol{\beta}^{(d)} - \phi_{pa(r)}^{(p-1)} \boldsymbol{\beta}^{(d)} \qquad\qquad r \in T_p \qquad (17)
$$

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

Lemma 14 (*p*-residuals interact only with *p*-residuals)**.** *Let $\boldsymbol{\beta}_T$ be a Gaussian random vector satisfying* (H) *and* (S̃)*. Then for any p and d with $0 \le p \le d \le k-1$, for all $t \in T_p$ we have the identity*

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

Given Lemma 14 we can prove the following multigrid decomposition for hierarchical linear models.

Theorem 15 (Multigrid decomposition for k levels)**.** *Let $(\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ be a Markov chain evolving according to GS($\boldsymbol{\beta}_T$) with $\boldsymbol{\beta}_T$ satisfying* (H) *and* (S̃)*. Then $(\delta^{(0)}\boldsymbol{\beta}(s))_s$, ..., $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$ are k independent Markov chains. Moreover, each chain $\delta^{(p)}\boldsymbol{\beta}(s) = (\delta^{(p)}\boldsymbol{\beta}^{(d)}(s))_{d=p}^{k-1}$ evolves according to the following blocked Gibbs sampler scheme with $(k-p)$ blocks: for d going from p to $k-1$ sample*

$$\delta^{(p)}\boldsymbol{\beta}^{(d)}(s+1) \sim \mathcal{L}\left(\delta^{(p)}\boldsymbol{\beta}^{(d)}|(\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s+1))_{p\le\ell<d}, (\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s))_{d<\ell\le k-1}\right), \tag{18}$$

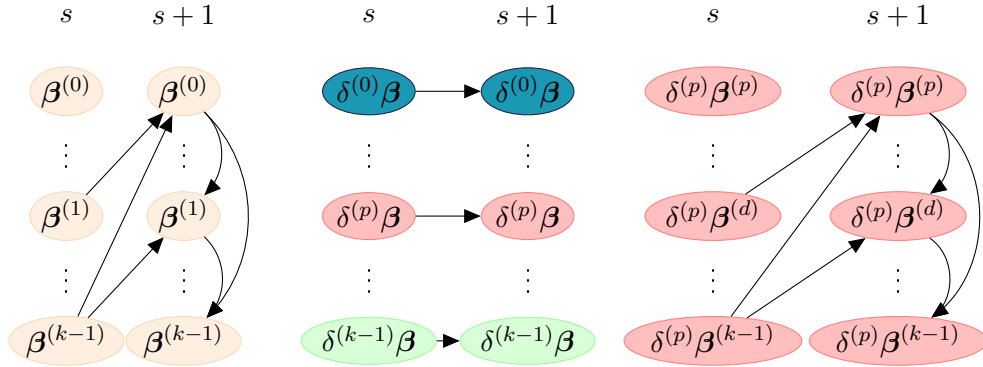*where $\mathcal{L}(X|Y)$ denotes the conditional distribution of X given Y.*

Figure 8: Illustration of Theorem 15. Left: transition from $\boldsymbol{\beta}(s)$ to $\boldsymbol{\beta}(s+1)$ in Sampler GS($\boldsymbol{\beta}_T$). This is the common structure of a Gibbs Sampler with k blocks. Center: transition from $\delta\boldsymbol{\beta}(s)$ to $\delta\boldsymbol{\beta}(s+1)$ in Sampler GS($\boldsymbol{\beta}_T$). There are k Markov chains, namely $(\delta^{(p)}\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ for $p \in \{0,\ldots,k-1\}$, evolving independently. Right: for each $p \in \{0,\ldots,k-1\}$ the transition from $\delta^{(p)}\boldsymbol{\beta}(s) = (\delta^{(p)}\boldsymbol{\beta}^{(d)}(s))_{d=p}^{k-1}$ to $\delta^{(p)}\boldsymbol{\beta}(s+1) = (\delta^{(p)}\boldsymbol{\beta}^{(d)}(s+1))_{d=p}^{k-1}$ follows a Gibbs Sampler scheme with $(k-p)$ blocks.

Theorem 15 implies that running a Gibbs sampler $(\boldsymbol{\beta}(s))_s$ targeting distributions satisfying (H) is equivalent to running k independent blocked Gibbs Samplers, one for each level of coarseness from $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ to $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$.

Corollary 16. *Let $\boldsymbol{\beta}_T$ satisfies* (H) *and* (S̃). *Then the rate of convergence of GS($\boldsymbol{\beta}_T$) is given by* $\max\{\rho_0, \ldots, \rho_{k-1}\}$ *where for each* $p \in \{0, \ldots, k-1\}$, ρ_p *is the rate of convergence of* $(\delta^{(p)}\boldsymbol{\beta}(s))_s$.

4.5 Hierarchical ordering of rates

The convergence properties of these k Markov chains are closely related, as the following theorem suggests.

Theorem 17. *The rate of convergence of* $(\delta^{(p)}\boldsymbol{\beta}(s))_s$ *is given by the largest modulus eigenvalue of* $(\mathbb{I}_{k-p} - L)^{-1}U$. *Here* $\mathbb{I}_{k-p}$ *is the* $(k-p)$ *dimensional identity matrix, while* L *and* U *are, respectively, the strictly lower and strictly upper triangular part of* $(c_{d\ell})_{d,\ell=p}^{k-1}$, *with* $c_{d\ell}$ *given by* (S̃).

Theorem 17 implies that, from the rates of convergence point of view, the k Markov chains updating $\delta^{(p)}\boldsymbol{\beta}$ 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 $C = (c_{d\ell})_{d,\ell=p}^{k-1}$ is given by (S̃). This suggests that the convergence properties of the sub-chains will typically improve from that of $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ to that $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$ and that the rate of convergence of $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ will typically determine the rate of the whole sampler GS($\boldsymbol{\beta}_T$). In particular, in the centred parametrization case we can show that the rate of convergence is monotonically non-increasing from $(\delta^{(0)}\boldsymbol{\beta}(s))_s$ to $(\delta^{(k-1)}\boldsymbol{\beta}(s))_s$.

Theorem 18. (Ordering of rates for centred parametrization) *Let* $\boldsymbol{\gamma}$ *be a Gaussian vector satisfying* (T) *and* (S̃) *and let* $(\boldsymbol{\gamma}(s))_{s\in\mathbb{N}}$ *be the corresponding Markov chain evolving according to GS($\boldsymbol{\gamma}_T$). Then the convergence rates of the k independent Markov chains* $(\delta^{(0)}\boldsymbol{\gamma}(s))_s$, ..., $(\delta^{(k-1)}\boldsymbol{\gamma}(s))_s$ *satisfy*

$$\rho(\delta^{(0)}\boldsymbol{\gamma}(s)) \geq \rho(\delta^{(1)}\boldsymbol{\gamma}(s)) \geq \cdots \geq \rho(\delta^{(k-1)}\boldsymbol{\gamma}(s)) = 0. \tag{19}$$

In Theorem 18 we needed the additional assumption (T) to prove (19). The reason is that, while in most cases the convergence rates of DSGS targeting a n-th dimensional Gaussian distribution improves if one of the coordinates is conditioned to a fixed value and the DSGS has to sample only from 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. Theorem 18 implies the following corollary.

Corollary 19. *The rate of convergence of a DSGS targeting a Gaussian distribution satisfying* (T) *and* (S̃) *is given by the largest squared eigenvalue of the k-dimensional matrix* $C - \mathbb{I}_k$, *where* $C = (c_{d\ell})_{d,\ell=0}^{k-1}$ *is defined by* (S̃) *and* $\mathbb{I}_k$ *is the k-dimensional matrix.*

In particular, considering the special case of Model Sk it is easy to deduce the following Corollary.

Corollary 20. *The rate of convergence of $GS(\boldsymbol{\gamma}_T)$ targeting Model Sk is given by the largest squared eigenvalue of the k-dimensional matrix*

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

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

4.6 Example: rates of convergence for 4-level models

The results developed in Sections 4.4 and 4.5 allow to analyze hierarchical models with an arbitrary number of levels. For example we could consider 4-level extensions of Model S3.

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

$$
y_{ijk\ell} = \mu + a_i + b_{ij} + c_{ijk} + \epsilon_{ijk\ell}, \tag{20}
$$

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

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

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

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

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

If instead $\lambda_1 = 0$ (i.e. using $(\gamma_i)_i$ with $\gamma_i = \mu + a_i$ at level 1) we have four different parametrization with rates of convergence given by

$$\rho_{000} = \frac{1}{2}\left(1 + r_{2,3}(r_{4,3} - r_{1,2}) + \sqrt{(1 + r_{2,3}(r_{4,3} - r_{1,2}))^2 - 4r_{2,1}r_{2,3}r_{4,3}}\right),$$

$$\rho_{001} = \frac{1}{2}\left(1 + r_{2,4}(r_{3,4} - r_{1,2}) + \sqrt{(1 + r_{2,4}(r_{3,4} - r_{1,2}))^2 - 4r_{2,1}r_{2,4}r_{3,4}}\right),$$

$$\rho_{011} = \frac{1}{2}\left(1 - r_{1,4}r_{4,2}r_{4,3} + r_{2,4}r_{3,4} + \sqrt{(1 - r_{1,4}r_{4,2}r_{4,3} + r_{2,4}r_{3,4})^2 - 4r_{2,4}r_{3,4}}\right),$$

$$\rho_{010} = \frac{1}{2}\left(1 - r_{1,3}r_{3,2}r_{3,4} + r_{2,3}r_{4,3} + \sqrt{(1 - r_{1,3}r_{3,2}r_{3,4} + r_{2,3}r_{4,3})^2 - 4r_{2,3}r_{4,3}}\right).$$

These rates can be easily derived from Corollary 16 and Theorem 17. Even if some of the rates obtained have more complicated expressions than in the two and three-levels case, especially when the highest level is centred, the theory developed in this section can be easily applied to models with an arbitrary number of levels. It is worth noting that in this 4-level case the skeleton chain $\delta^{(0)}\boldsymbol{\beta}$ is always the slowest chain for all mixed parametrizations (can be checked by inspection), even if we provided a general proof of this fact only for the centred parametrization. The expressions given here can be easily used to derive conditionally optimal parametrizations for Model S4 given the rescaled variance components $(\tilde{\sigma}_i^2)_{i=1}^4$. For example, choosing whether to center or not each level by comparing the level-specific rescaled variances with the sum of the rescaled variances of the lower levels like in Section 3.2 leads to rates of convergence upper bounded by $\frac{3}{4}$. Note that, while this strategy is optimal for 3-level models (see Section 3.2 and Corollary 5), the same strategy is not the optimal one for 4 levels. In fact the supremum over $(\tilde{\sigma}_i^2)_{i=1}^4$ of $\min_{(\lambda_1,\lambda_2,\lambda_3)\in\{0,1\}^3}\rho_{(\lambda_1,\lambda_2,\lambda_3)}$ is strictly smaller than $\frac{3}{4}$, but the improvement is small and does not seem sufficient to motivate a more complicated strategy.

4.7 Optimal partial non-centering

In Section 4.2, we defined the class of Partially Non-Centered Parametrizations (PNCP) $\boldsymbol{\beta}_T$ of Model NSk. As for the 3-level case (see Section 3.5), even in this context one can find a optimal PNCP that factorizes the unknown parameters $\boldsymbol{\beta}_T$ into independent random variables.

Theorem 21. *Let $\boldsymbol{\gamma}_T$ be given by Model NSk (conditioned on some value of $\boldsymbol{y}$). Denote by $Q = (Q_{tr})_{t,r\in T}$ its precision matrix and define $\Lambda = (\lambda_{tr})_{r\preceq t,t\in T}$*

as

$$\lambda_{tt} = 1 \qquad\qquad\qquad t \in T,$$

$$\lambda_{t\,pa(t)} = \frac{Q_{t\,pa(t)}}{D_{tt}} \qquad\qquad\qquad t \in T\backslash\{t_0\},$$

$$\lambda_{tr} = 0 \qquad\qquad\qquad t \in T,\ r \notin \{t, pa(t)\},$$

where

$$D_{tt} = Q_{tt} - \sum_{t' \in ch(t)} \lambda_{t't}^2 D_{t't'} \qquad\qquad t \in T\backslash\{t_0\}\,. \qquad (21)$$

Then the resulting hierarchical reparametrization $\boldsymbol{\beta}_T = \Lambda\boldsymbol{\gamma}_T$ is a collection of independent random variables.

Note that the re-parametrization coefficients $\Lambda = (\lambda_{tr})_{r \preceq t, t \in T}$ defined in can be easily calculated iteratively starting from the leaves going up to the root of T. In fact the optimal PNCP is related to the multiscale or change-of-resolution Kalman filter (see e.g. Chou et al., 1994) and more generally to belief propagation algorithms for trees.

5 Conclusions and future work

In this work we studied the convergence properties of Gibbs Sampler algorithms in the context of Gaussian hierarchical models. To do so we developed a novel analytic approach based on multigrid decompositions that allows to factorize the Markov chain of interest into independent and easier to analyze sub-chains. This decomposition enables us to evaluate explicitly the L^2-rate of convergence in symmetric (e.g. Model Sk) and weakly-symmetric cases (e.g. Examples 2 and 3). These results can then be used to provide upper bounds in the more general non-symmetric case (e.g. Theorem 6).

We believe that the results presented in this work can provide a novel and significant contribution to understanding the behavior of MCMC algorithms (and in particular the Gibbs Sampler) in the extremely popular context of hierarchical linear models. In fact, together with explicit formulas for L^2-rates of convergences, the multigrid decomposition we presented here 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 18), algorithmic scalability with width of the hierarchy but not with height (e.g. Theorem 4 and Corollary 20) and good performances of centred parametrization in data-rich contexts (Theorem 6). Also, it is worth noting that these results are not limited to hierarchical models with two levels but rather can be applied to models with an arbitrary number of layers.

The results presented here have also various practically useful implications in terms of algorithmic implementations of MCMC for hierarchical linear models. For example, our analysis provides easy and theoretically grounded indications to choose the computationally optimal parametrization on-the-fly (see Sections 3.1, 3.2), how to handle heterogeneity (see Section 3.4) and indicates which parameters to monitor in the convergence diagnostic process (see Theorem 2 and discussion at the end of Section 2.1). Also, the bounds in Theorem 6 can be used to have a theoretical guarantee of good convergence properties with large datasets.

The present work could be extended in many directions. First one could extend the results for non-symmetric cases by improving and generalizing the bounds of Theorem 6, by providing upper bounds on the rates of convergence for the optimal bespoke parametrization (Section 3.4) and ideally by weakening the symmetry assumption in (S). Also it would be worth exploring more deeply the use of partially-non centred parametrization and their computational feasibility. Finally an ambitious aim, involving non-trivial generalizations of the approach of Amit [1996], would be to extend to other tractable distributions within the exponential family beyond the Gaussian.

In terms of classes of models considered, a natural and important extension would be to consider the multivariate case (where each parameter γ_t is a multivariate random vector) and the regression case. This would make the results presented here applicable to a large class of widely used models. 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 considering and 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, 2016a). Also, it would be interesting to compare nested structures like the one considered here to crossed-effect structures (see e.g. Gao and Owen, 2016).

In terms of algorithms considered it would be relatively easy to extend the machinery developed in this work to other Gibbs Sampling schemes, such as the random scan Gibbs Sampling case, where the component to be updated is chosen uniformly at random at each step. In this case the resulting sub-chains would not evolve completely independently because they would be coupled by the random choice of which component to update. Nevertheless each sub-chain would still be a Markov chain marginally and the overall rate of the vector of coupled Markov chains would still be amenable to direct analysis. Also, in the random scan case the reversibility of the induced Markov chains would allow to prove orderings results under weaker assumptions than Theorem 18.

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 it-

28

eration and the target distribution. 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).

Acknowledgments

References

Yali Amit. On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.*, 38(1):82–99, 1991.

Yali Amit. Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.*, 24(1):122–140, 1996.

Mark R Bass and Sujit K Sahu. A comparison of centring parameterisations of gaussian process-based models for bayesian computation using mcmc. *Statistics and Computing*, pages 1–22, 2016a.

Mark R Bass and Sujit K Sahu. A comparison of centring parameterisations of gaussian process-based models for bayesian computation using mcmc. *Statistics and Computing*, pages 1–22, 2016b.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Kenneth Chien-ko Chou, Alan S Willsky, and Albert Benveniste. Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, 39(3):464–478, 1994.

Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Stochastic alternating projections. *Illinois J. Math.*, 54(3):963–979, 2010.

Katelyn Gao and Art B Owen. Efficient moment calculations for variance components in large unbalanced crossed random effects models. *arXiv preprint arXiv:1602.00346*, 2016.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Alan E Gelfand, Sujit K Sahu, and Bradley P Carlin. Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488, 1995.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis, 3rd edition, 2013.

Jonathan Goodman and Alan D Sokal. Multigrid monte carlo method. conceptual foundations. *Physical Review D*, 40(6):2035, 1989.

Susan E Hills and Adrian FM Smith. Parameterization issues in bayesian inference. *Bayesian statistics*, 4:227–246, 1992.

James E Johndrow, Jonathan C Mattingly, Sayan Mukherjee, and David Dunson. Approximations of Markov Chains and Bayesian Inference. *arXiv preprint arXiv:1508.03387*, 2015.

Galin L Jones and James P Hobert. Sufficient burn-in for gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2): 784–817, 2004.

Jun S Liu and Chiara Sabatti. Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, pages 353–369, 2000.

Xiao-Li Meng and David Van Dyk. 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, 1997.

O Papaspiliopoulos, Gareth O Roberts, and M Skold. Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West, editors, *Bayesian Statistics 7*, pages 307–326. Oxford University Press, New York, 2003.

Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

Bala Rajaratnam and Doug Sparks. 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*, 2015.

Gareth O Roberts and Jeffrey S Rosenthal. Complexity bounds for markov chain monte carlo algorithms via diffusion limits. *Journal of Applied Probability*, 53(02):410–420, 2016.

Gareth O Roberts and Sujit K Sahu. 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, 1997.

Sujit K Sahu and Gareth O Roberts. On convergence of the em algorithmand the gibbs sampler. *Statistics and Computing*, 9(1):55–64, 1999.

Adrian FM Smith and Gareth O Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.

Joe Whittaker. Graphical models in applied multivariate analysis, 1990.

Yaming Yu and Xiao-Li Meng. 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, 2011.

A   Full conditional distributions of $\mathbf{GS}(1,1)$ and $\mathbf{GS}(0,0)$

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

$$\mu(s+1) \sim N\left(y_{\cdots} - a_{\cdot}(s) - b_{\cdot\cdot}(s), \frac{\sigma_e^2}{n}\right)$$

$$a_i(s+1) \sim N\left(\frac{\sigma_a^2(y_{i\cdot\cdot} - \mu(s+1) - b_{i\cdot}(s))}{\sigma_a^2 + \frac{\sigma_e^2}{JK}}, \frac{\sigma_a^2 \frac{\sigma_e^2}{JK}}{\sigma_a^2 + \frac{\sigma_e^2}{JK}}\right) \qquad \forall i$$

$$b_{ij}(s+1) \sim N\left(\frac{\sigma_b^2(y_{ij\cdot} - \mu(s+1) - a_i(s+1))}{\sigma_b^2 + \frac{\sigma_e^2}{K}}, \frac{\sigma_b^2 \frac{\sigma_e^2}{K}}{\sigma_b^2 + \frac{\sigma_e^2}{K}}\right) \qquad \forall i,j$$

where we use the dot subscript to indicate averaging over one dimension, meaning that $a_{\cdot} = \sum_i \frac{a_i}{I}$, $b_{\cdot\cdot} = \sum_{i,j} \frac{b_{ij}}{IJ}$, $y_{\cdots} = \sum_{i,j,k} \frac{y_{ijk}}{IJK}$, $b_{i\cdot} = \sum_j \frac{b_{ij}}{J}$, $y_{i\cdot\cdot} = \sum_{j,k} \frac{y_{ijk}}{JK}$ and $y_{ij\cdot} = \sum_k \frac{y_{ijk}}{K}$.

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

$$\mu(s+1) \sim N\left(\gamma_{\cdot}(s), \frac{\sigma_a^2}{I}\right)$$

$$\gamma_i(s+1) \sim N\left(\frac{\frac{\sigma_b^2}{J}\mu(s+1) + \sigma_a^2 \eta_{i\cdot}(s)}{\frac{\sigma_b^2}{J} + \sigma_a^2}, \frac{\sigma_a^2 \frac{\sigma_b^2}{J}}{\sigma_a^2 + \frac{\sigma_b^2}{J}}\right) \qquad \forall i$$

$$\eta_{ij}(s+1) \sim N\left(\frac{\frac{\sigma_e^2}{K}\gamma_i(s+1) + \sigma_b^2 y_{ij\cdot}}{\sigma_b^2 + \frac{\sigma_e^2}{K}}, \frac{\sigma_b^2 \frac{\sigma_e^2}{K}}{\sigma_b^2 + \frac{\sigma_e^2}{K}}\right) \qquad \forall i,j$$

where as before the dot subscript indicates averaging over indices.

B Theorem proofs

In this section we list proofs following the order of appearance in the paper. This is different from the mathematical chronology. To follow the latter, the reader should start from the results (lemmas, theorems and corollaries) for k-level models, namely the results from Lemma 10 to Theorem 21 and then move to the ones for 3-level models, namely the results from Theorem 1 to Theorem 9.

Proof of Theorems 1, 2 and 4

Theorems 1, 2 and 4 are substantially special cases of the Theorems of Section 4. In particular Theorem 1 is a special case of Theorem 15 for $k = 2$ and Theorems 2 and 4 can be directly verified as follows. Using Corollary 16 we can evaluate the rates of convergence of the three subchains $\rho(\delta^{(0)}\boldsymbol{\beta}(s))$, $\rho(\delta^{(1)}\boldsymbol{\beta}(s))$ and $\rho(\delta^{(2)}\boldsymbol{\beta}(s))$ for the four parametrzations under consideration in Section 3 and check by inspection that

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

and that the rates of convergence $\rho(\delta^{(0)}\boldsymbol{\beta}(s))$ are the ones given by Theorem 4. In particular the rates of convergence of $\delta^{(0)}\boldsymbol{\beta}(s)$, $\delta^{(1)}\boldsymbol{\beta}(s)$ and $\delta^{(2)}\boldsymbol{\beta}(s)$ under $GS(\boldsymbol{\beta})$ are given by the following Table.

	$\rho(\delta^{(0)}\boldsymbol{\beta}(s))$	$\rho(\delta^{(1)}\boldsymbol{\beta}(s))$	$\rho(\delta^{(2)}\boldsymbol{\beta}(s))$
$(\mu,\mathbf{a},\mathbf{b})$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_e^2} \vee \frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_e^2}\frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	0
$(\mu,\boldsymbol{\gamma},\boldsymbol{\eta})$	$1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_b^2}\frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_b^2}\left(1 - \frac{\tilde{\sigma}_b^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}\right)$	0
$(\mu,\boldsymbol{\gamma},\mathbf{b})$	$1 - \frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_e^2}\frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_e^2}\left(1 - \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}\right)$	0
$(\mu,\mathbf{a},\boldsymbol{\eta})$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_b^2} \vee \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	$\frac{\tilde{\sigma}_a^2}{\tilde{\sigma}_a^2+\tilde{\sigma}_b^2}\frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_b^2+\tilde{\sigma}_e^2}$	0

Figure 9: Rates of convergence of $\delta^{(0)}\boldsymbol{\beta}(s)$, $\delta^{(1)}\boldsymbol{\beta}(s)$ and $\delta^{(2)}\boldsymbol{\beta}(s)$ for $GS(\boldsymbol{\beta})$ under various parametrizations.

Proof of Corollary 3

By Theorem 1 the rate of convergence of the whole chain $(\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ coincides with the maximum of the rates of the subchains, meaning that $\rho(\boldsymbol{\beta}(s)) = \max\{\rho(\delta^{(0)}\boldsymbol{\beta}(s)), \rho(\delta^{(1)}\boldsymbol{\beta}(s)), \rho(\delta^{(2)}\boldsymbol{\beta}(s))\}$. By Theorem 2 the latter maximum equals $\rho(\delta^{(0)}\boldsymbol{\beta}(s))$.

Proof of Corollary 5

Follows from Theorem 4 by checking that for both $\boldsymbol{\alpha} = \boldsymbol{\gamma}$ and $\boldsymbol{\alpha} = \mathbf{a}$, the inequality $\rho_{(\mu,\boldsymbol{\alpha},\boldsymbol{\eta})} \leq \rho_{(\mu,\boldsymbol{\alpha},\mathbf{b})}$ holds if and only if $\tilde{\sigma}_b^2 \geq \tilde{\sigma}_e^2$; and for both $\boldsymbol{\alpha} = \boldsymbol{\gamma}$

and $\boldsymbol{\alpha} = \mathbf{a}$ the inequality $\rho_{(\mu,\gamma,\boldsymbol{\alpha})} \leq \rho_{(\mu,\mathbf{a},\boldsymbol{\alpha})}$ holds if and only if $\tilde{\sigma}_a^2 \geq \tilde{\sigma}_b^2 + \tilde{\sigma}_e^2$.

Proof of Theorem 6

Given an instance of Model NS3 with variance terms $(\sigma_a^2, (\sigma_{b,i}^2)_i, (\sigma_{e,ij}^2)_{ij})$ satisfying (5), the proof will proceed by comparing the original Gibbs Sampler with an auxiliary Gibbs Sampler targeting a different instance of Model NS3 with variance terms $(\sigma_a^2, (\sigma_{b,i}^2)_i, (\bar{\sigma}_{e,ij}^2)_{ij})$ satisfying (S*) and thus allowing direct analysis using Corollary 19. In the context of Model NS3, (S*) reduces to requiring $\sum_{j=1}^{J_i} \rho_{\gamma_i \eta_{ij}}^2$ to be constant over i, where $\rho_{\gamma_i \eta_{ij}}$ is the partial correlation $Corr(\gamma_i, \eta_{ij} | \mu, (\gamma_\ell)_{\ell \neq i}, (\eta_{\ell s})_{(\ell s) \neq (ij)})$ as in Section 4.3. By computing the partial correlations of Model NS3 it can be checked that $\sum_{j=1}^{J_i} \rho_{\gamma_i \eta_{ij}}^2 = r_{a,b}^{(i)} r_{e,b}^{(i)}$, where $r_{a,b}^{(i)}$ and $r_{e,b}^{(i)}$ are defined in Theorem 6. For each $i = 1, \ldots, I$ we define auxiliary variance terms $(\bar{\sigma}_{e,ij}^2)_{j=1}^{J_i}$ such that $\bar{\sigma}_{e,ij}^2 \geq \sigma_{e,ij}^2$ for all $j = 1, \ldots, J_i$ and

$$r_{a,b}^{(i)} \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{K_{ij}^{-1} \bar{\sigma}_{e,ij}^2}{\sigma_{b,i}^2 + K_{ij}^{-1} \bar{\sigma}_{e,ij}^2} = \max_{\ell=1,\ldots,I} r_{a,b}^{(\ell)} r_{e,b}^{(\ell)}. \tag{22}$$

Such $(\bar{\sigma}_{e,ij}^2)_{j=1}^{J_i}$ exist because $r_{a,b}^{(i)} \geq \max_{\ell=1,\ldots,I} r_{a,b}^{(\ell)} r_{e,b}^{(\ell)}$ by (5) and the left hand side of (22) can take any value in $(0, r_{a,b}^{(i)}]$ for $(\bar{\sigma}_{e,ij}^2)_{j=1}^{J_i}$ belonging to $[0, \infty)$. (22) implies that the instance of Model NS3 with variance terms $(\sigma_a^2, (\sigma_{b,i}^2)_i, (\bar{\sigma}_{e,ij}^2)_{ij})$ satisfies (S*) with $c_0 = \sum_{i=1}^{I} \rho_{\mu \gamma_i}^2 = 1 - \frac{1}{I} \sum_{i=1}^{I} r_{a,b}^{(i)}$ and $c_1 = \sum_{j=1}^{J_i} \rho_{\gamma_i \eta_{ij}}^2 = \max_{\ell=1,\ldots,I} r_{a,b}^{(\ell)} r_{e,b}^{(\ell)}$. As discussed in Example 3, for models with centred parametrization like Model NS3, (S*) implies (S) and, after rescaling, ($\tilde{\text{S}}$). In this case the matrix $C = (c_{dp})_{d,p=0}^2$ is given by

$$C = \begin{pmatrix} 1 & \sqrt{c_0} & 0 \\ \sqrt{c_0} & 1 & \sqrt{c_1} \\ 0 & \sqrt{c_1} & 1 \end{pmatrix}.$$

Therefore, by Corollary 19, the rate of convergence of the Gibbs Sampler targeting the posterior distribution of Model NS3 with variance terms $(\sigma_a^2, (\sigma_{b,i}^2)_i, (\bar{\sigma}_{e,ij}^2)_{ij})$ is given by $c_0 + c_1$, which is the largest squared eigenvalue of $C - \mathbb{I}_3$, where $\mathbb{I}_3$ is the 3-dimensional identity matrix.

Finally we show that the Gibbs Sampler rate of convergence induced by the auxiliary variance terms $(\sigma_a^2, \sigma_{b,i}^2, \bar{\sigma}_{e,ij}^2)$ is greater or equal than the original one given by $(\sigma_a^2, \sigma_{b,i}^2, \sigma_{e,ij}^2)$. Denote by Q the precision matrix of the original posterior distribution and by $\bar{Q}$ the auxiliary one. By deriving Q and $\bar{Q}$ from the definition of Model NS3, it is easy to see that the only terms of Q and $\bar{Q}$ affected by replacing $\sigma_{e,ij}^2$ with $\bar{\sigma}_{e,ij}^2$ are $(Q_{\eta_{ij} \eta_{ij}})_{ij}$ and $(\bar{Q}_{\eta_{ij} \eta_{ij}})_{ij}$. Moreover $\bar{\sigma}_{e,ij}^2 \geq \sigma_{e,ij}^2$ implies $Q_{\eta_{ij} \eta_{ij}} = \frac{1}{\sigma_{b,i}^2} + \frac{K_{ij}}{\sigma_{e,ij}^2} \geq \frac{1}{\sigma_{b,i}^2} + \frac{K_{ij}}{\bar{\sigma}_{e,ij}^2} = \bar{Q}_{\eta_{ij} \eta_{ij}}$.

The result then follows from Theorem 7 of Roberts and Sahu [1997], which states that the convergence rate of a deterministic scan Gibbs Sampler with single-site update is a non-increasing function of the diagonal elements of the target precision matrix (when the off-diagonal terms are kept constant).

Proof of Theorem 7

The Markov chain under consideration is a Gibbs Sampler sweeping through $(\mu, \beta_1, \ldots, \beta_I)|\mathbf{y}$ for some observed data $\mathbf{y} = ((y_{ij})_{j=1}^{J_i})_{i=1}^I$ assuming Model NS2 and $\beta_i = \gamma_i - \lambda_i \mu$ with $\lambda_i \in \{0, 1\}$. To compute the Gibbs Sampler rate of convergence we first need to compute the $(I+1) \times (I+1)$ matrix A indexed by $(\alpha_1, \alpha_2) \in \{\mu, \beta_1, \ldots, \beta_I\} \times \{\mu, \beta_1, \ldots, \beta_I\}$ and defined as $A_{\alpha_1 \alpha_2} = -\frac{Q_{\alpha_1 \alpha_2}}{Q_{\alpha_1 \alpha_1}}$ for $\alpha_1 \neq \alpha_2$ and 0 for $\alpha_1 = \alpha_2$, where Q is the precision matrix of $(\mu, \beta_1, \ldots, \beta_I)|\mathbf{y}$. See [Roberts and Sahu, 1997, Sec.2.2] for more details on the derivation of such A-matrices. By computing the precision matrix of $(\mu, \beta_1, \ldots, \beta_I)|\mathbf{y}$ it is easy to see that A is given by

$$
\begin{pmatrix}
0 & A_{\mu\beta_1} & \cdots & A_{\mu\beta_I} \\
A_{\beta_1\mu} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
A_{\beta_I\mu} & 0 & \cdots & 0
\end{pmatrix}
$$

where for all i

$$
A_{\mu\beta_i} = \frac{-\lambda_i \tilde{\tau}_i + (1 - \lambda_i)\tau_a}{\sum_{\ell=1}^I \lambda_\ell \tilde{\tau}_\ell + (1 - \lambda_\ell)\tau_a}, \quad A_{\beta_i\mu} = \frac{-\lambda_i \tilde{\tau}_i + (1 - \lambda_i)\tau_a}{\tilde{\tau}_i + \tau_a}.
$$

From Roberts and Sahu [1997, Thm.1] the rate of convergence of the Gibbs sampler of interest equals the largest modulus eigenvalue of $B = (\mathbb{I}_{I+1} - L)^{-1}U$, where $\mathbb{I}_{I+1}$ is the $(I+1)$-dimensional identity matrix, L is the lower triangular part of A and $U = A - L$. Simple calculations show that in this case

$$
B = \begin{pmatrix} 0 & A_{\mu\beta_1} & \cdots & A_{\mu\beta_I} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ A_{\beta_1\mu} \\ \vdots \\ A_{\beta_I\mu} \end{pmatrix} = \begin{pmatrix}
0 & A_{\mu\beta_1} & \cdots & A_{\mu\beta_I} \\
0 & A_{\mu\beta_1}A_{\beta_1\mu} & 0 & A_{\mu\beta_I}A_{\beta_1\mu} \\
\vdots & \vdots & \ddots & \vdots \\
0 & A_{\mu\beta_1}A_{\beta_I\mu} 0 & \cdots & A_{\mu\beta_I}A_{\beta_I\mu}.
\end{pmatrix}
$$

Finally note that B has I eigenvalues equal to 0 and one equal to

$$
\sum_{i=1}^I A_{\mu\beta_i}A_{\beta_i\mu} = \frac{\sum_{i:\lambda_i=1} \tilde{\tau}_i \frac{\tilde{\tau}_i}{\tilde{\tau}_i+\tau_a} + \sum_{i:\lambda_i=0} \tau_a \frac{\tau_a}{\tilde{\tau}_i+\tau_a}}{\sum_{i:\lambda_i=1} \tilde{\tau}_i + \sum_{i:\lambda_i=0} \tau_a}.
$$

Proofs of Lemmas 10 and 11

In order to prove Lemmas 10 and 11 we need some preliminary results on matrices $M = (M_{tr})_{t,r \in T}$ indexed by elements of a tree T.

Lemma 22 (Triangular matrices on trees). *Suppose that a matrix* $L = (L_{tr})_{t,r \in T}$ *satisfies the following lower-triangularity condition*

$$L_{tr} = 0 \text{ unless } t \succeq r \text{ and } L_{tt} \neq 0 \quad \forall t \in T. \tag{L_m}$$

Then L is invertible and its inverse satisfies (L_m). *Similarly, if* $U = (U_{tr})_{t,r \in T}$ *satisfies the following upper-triangularity condition*

$$U_{tr} = 0 \text{ unless } t \preceq r \text{ and } U_{tt} \neq 0 \quad \forall t \in T, \tag{U_m}$$

then U is invertible and its inverse still satisfies (U_m).

Proof. Suppose that L satisfies (L_m). Also without loss of generality suppose $L_{tt} = 1$ for all $t \in T$ by rescaling. Then write $L = (\mathbb{I}_T + N)$ where $\mathbb{I}_T$ is the $|T| \times |T|$ identity matrix and N satisfies the following strict lower-triangularity condition

$$N_{tr} = 0 \text{ unless } t \succ r. \tag{L_m^*}$$

Consider $N_{tr}^2 = \sum_{s \in T} N_{ts} N_{sr} = \sum_{s \in T : r \prec s \prec t} N_{ts} N_{sr}$. From the last expression it follows that $N_{tr}^2 \neq 0$ implies $r \prec t$ and $|\ell(r) - \ell(t)| \geq 2$, where $\ell(t)$ denote the level of t in the tree T. Iterating the same argument we have that $N_{tr}^p \neq 0$ implies $r \prec t$ and $|\ell(r) - \ell(t)| \geq p$. It follows that N^p satisfies (L_m^*) for all $p \geq 1$ and that $N^p = \mathbf{0}_T$ for all $p \geq k$ where $\mathbf{0}_T$ is the $|T| \times |T|$ zero matrix. Here k indicates the number of levels of T, as in Section 4. From $N^p = \mathbf{0}_T$ for $p \geq k$ it follows that

$$(\mathbb{I}_T + N)(\mathbb{I}_T + \sum_{p=1}^{k-1} (-1)^p N^p) = \mathbb{I}_T + (-1)^{k-1} N^k = \mathbb{I}_T$$

and therefore $L^{-1} = (\mathbb{I}_T + N)^{-1} = \mathbb{I}_T + \sum_{p=1}^{k-1} (-1)^p N^p$. Since N^p satisfies (L_m^*) for all $p \geq 1$ it follows that $L^{-1} = \mathbb{I}_T + \sum_{p=1}^{k-1} (-1)^p N^p$ satisfies (L_m).

The analogous statement for (U_m) can be deduced by observing that U satisfies (U_m) if and only if its transpose satisfies (L_m). $\qquad\square$

Lemma 22 can be used to deduce Lemma 10 in a straightforward way.

Proof of Lemma 10. Suppose that $\boldsymbol{\beta}_T = \Lambda \boldsymbol{\gamma}_T$ is a hierarchical reparametrization of $\boldsymbol{\gamma}_T$. This is equivalent to say that $\Lambda = (\Lambda_{tr})_{t,r,\in T}$ is a matrix satisfying (L_m) and that for all $t \in T$, $\boldsymbol{\beta}_t = \sum_{s \in T} \Lambda_{ts} \gamma_s$. Lemma 22 implies that Λ is invertible and that $Z = \Lambda^{-1}$ satisfies (L_m). Therefore $\boldsymbol{\gamma}_T = Z \boldsymbol{\beta}_T$ is a hierarchical reparametrization of $\boldsymbol{\beta}_T$. $\qquad\square$

To prove Lemma 11 we need an additional preliminary result.

Lemma 23 (Closure of H_m). *Suppose that $M = (M_{tr})_{t,r \in T}$ satisfies the following condition*

$$M_{tr} = 0 \ \text{unless } t \preceq r \text{ or } r \preceq t. \tag{H_m}$$

Then if we multiply M from the right with a martix L satisfying (L_m), the product ML still satisfy (H_m). Similarly if we multiply M from the left with a martix U satisfying (U_m), the product UM still satisfy (H_m).

Proof. Consider $(UM)_{tr} = \sum_{s \in T} U_{ts} M_{sr}$. From (H_m) and (U_m), the elements $U_{ts} M_{sr}$ in the latter sum are non-zero only when s belongs to the intersection of $\{s \in T : s \succeq t\}$ and $\{s \in T : s \succeq r \text{ or } s \preceq r\}$. It is easy to see that the intersection of these two sets is non-empty only if $t \succeq r$ or $t \preceq r$. Therefore UM satisfies (H_m). The argument to show that ML satisfies (H_m) is analogous. $\qquad\square$

We now combine Lemmas 22 and 23 to prove Lemma 11.

Proof of Lemma 11. Suppose that $\boldsymbol{\beta}_T$ satisfies (H). This is equivalent to saying that its precision matrix $Q^{(\boldsymbol{\beta})} = (Q_{tr}^{(\boldsymbol{\beta})})_{t,r \in T}$ satisfies (H_m). Consider a hierarchical reparametrization of $\boldsymbol{\beta}_T$ denoted by $\Lambda \boldsymbol{\beta}_T$. Then its precision matrix is given by $Q^{(\Lambda\boldsymbol{\beta})} = (\Lambda^T)^{-1} Q^{(\boldsymbol{\beta})} \Lambda^{-1}$ where Λ^T denote the transpose of Λ. By definition of hierarchical reparametrizations, Λ satisfies (L_m). Therefore Lemma 22 implies that Λ^{-1} satisfies (L_m) and, consequently, Lemma 23 implies that $Q^{(\boldsymbol{\beta})} \Lambda^{-1}$ satisfies (H_m). Since Λ satisfies (L_m), then Λ^T satisfies (U_m) and thus Lemma 22 implies that $(\Lambda^T)^{-1}$ satisfies (U_m). We can then apply Lemma 23 to $(\Lambda^T)^{-1}$ and $Q^{(\boldsymbol{\beta})} \Lambda^{-1}$ to deduce that $(\Lambda^T)^{-1} Q^{(\boldsymbol{\beta})} \Lambda^{-1}$ satisfies (H_m) and thus $\Lambda \boldsymbol{\beta}_T$ satisfies (H). $\qquad\square$

Corollary 12 follows easily from equation (T) and Lemma 11.

Proof of Lemma 13

The strategy to prove Lemma 13 is similar to the one used to prove Lemmas 10 and 11 above, with the difference that we have to check that the symmetry condition is preserved under the operations considered. To do so we first prove two auxiliary lemmas.

Lemma 24 (Symmetric triangular matrices on trees). *Suppose that a matrix $L = (L_{tr})_{t,r \in T}$ satisfies the following symmetric lower-triangularity condition*

$$L_{tr} = l_{\ell(t)\ell(r)} \mathbb{1}(t \succeq r) \ \text{and } l_{dd} \neq 0 \text{ for all } d \in \{0, \ldots, k-1\}, \tag{SL_m}$$

where $(l_{pd})_{p,d=0}^{k-1}$ is a $k \times k$ real valued matrix. Then L is invertible and its inverse satisfies (SL_m). Similarly, if $U = (U_{tr})_{t,r \in T}$ satisfies the following symmetric upper-triangularity condition

$$U_{tr} = u_{\ell(t)\ell(r)} \mathbb{1}(t \preceq r) \text{ and } u_{dd} \neq 0 \text{ for all } d \in \{0, \ldots, k-1\}, \qquad (\text{SU}_m)$$

with $(u_{pd})_{p,d=0}^{k-1}$ being a $k \times k$ real valued matrix, then U is invertible and its inverse still satisfies (SU_m).

Proof. Suppose that L satisfies (SL_m). Also without loss of generality suppose $L_{tt} = 1$ for all $t \in T$ by rescaling. Since L also satisfies (L_m), arguing as in the proof of Lemma 22 we can write $L^{-1} = \mathbb{I}_T + \sum_{p=1}^{k-1}(-1)^p N^p$ where $N = L - \mathbb{I}_T$ and N satisfies the following symmetric strict lower-triangularity condition

$$N_{tr} = l_{\ell(t)\ell(r)} \mathbb{1}(t \succ r) \text{ and } l_{dd} \neq 0 \text{ for all } d \in \{0, \ldots, k-1\}. \qquad (\text{SL}_m^*)$$

From $N_{tr}^2 = \sum_{s \in T} N_{ts} N_{sr} = \sum_{s \in T \,:\, r \prec s \prec t} N_{ts} N_{sr} = \sum_{\ell' \,:\, \ell(r) \prec \ell' \prec \ell(t)} l_{\ell(t)\ell'} l_{\ell'\ell(r)}$ we deduce that N^2 still satisfies (SL_m^*) for some different $(l_{pd})_{p,d=0}^{k-1}$. Iterating the same argument we have that N^p satisfies (SL_m^*) for all $p \geq 1$. Since $L^{-1} = \mathbb{I}_T + \sum_{p=1}^{k-1}(-1)^p N^p$ it follows that L^{-1} satisfies (SL_m).

The analogous statement for (SU_m) can be deduced by observing that U satisfies (SU_m) if and only if its transpose satisfies (SL_m). $\square$

Lemma 25. *Let $M = (M_{tr})_{t,r \in T}$ and $L = (L_{tr})_{t,r \in T}$ satisfy respectively $(\tilde{\text{S}})$ and (SL_m). Then the product ML satisfies $(\tilde{\text{S}})$ for some different matrix $(c_{pd})_{p,d=1}^{k-1}$. Similarly, if $U = (U_{tr})_{t,r \in T}$ satisfies (SU_m) then the product UM satisfy $(\tilde{\text{S}})$.*

Proof. First consider the product ML. Lemma 23 implies that ML satisfies (H_m) and therefore $(ML)_{tr} = 0$ unless $t \preceq r$ or $t \succeq r$. Given $t \preceq r$ or $t \succeq r$ and using $(\tilde{\text{S}})$ and (SL_m) we have

$$\begin{aligned}
(ML)_{tr} &= \sum_{s \in T} M_{ts} L_{sr} = \sum_{s \in T \,:\, s \succeq r} M_{ts} L_{sr} \\
&= \sum_{s \in T \,:\, s \succeq (t \vee r)} M_{ts} L_{sr} + \sum_{s \in T \,:\, r \preceq s \prec t} M_{ts} L_{sr} \\
&= \sum_{s \in T \,:\, s \succeq (t \vee r)} m_{\ell(t)\ell(s)} l_{\ell(s)\ell(r)} P(s) + \sum_{s \in T \,:\, r \preceq s \prec t} m_{\ell(t)\ell(s)} l_{\ell(s)\ell(r)} P(t) \\
&= P(t \vee r) \sum_{\ell' = \ell(t \vee r)}^{k-1} m_{\ell(t)\ell'} l_{\ell'\ell(r)} + P(t) \sum_{\ell' = \ell(r)}^{\ell(t)-1} m_{\ell(t)\ell'} l_{\ell'\ell(r)} \\
&= P(t \vee r) \sum_{\ell' = \ell(r)}^{k-1} m_{\ell(t)\ell'} l_{\ell'\ell(r)},
\end{aligned}$$

where the sum from $\ell(r)$ to $\ell(t)$ equals 0 if $\ell(r) \geq \ell(t)$. The latter equation implies that ML satisfies (S̃).

To prove that the product UM satisfies (S̃) note that M and U satisfying (S̃) and (SU$_m$) respectively is equivalent to M^T and U^T satisfying (S̃) and (SL$_m$) respectively. Therefore, by the first part of this Lemma, $(UM)^T = M^T U^T$ satisfies (SL$_m$) and thus $((UM)^T)^T = UM$ satisfies (S̃). $\qquad\square$

We now combine Lemmas 24 and 25 to prove Lemma 13.

Proof of Lemma 13. Suppose that the precision matrix $Q^{(\tilde{\boldsymbol{\beta}})} = (Q_{tr}^{(\tilde{\boldsymbol{\beta}})})_{t,r \in T}$ of $\tilde{\boldsymbol{\beta}}_T$ satisfies (S̃). Consider a symmetric hierarchical reparametrization of $\tilde{\boldsymbol{\beta}}_T$ denoted by $\Lambda\tilde{\boldsymbol{\beta}}_T$. Then its precision matrix is given by $Q^{(\Lambda\tilde{\boldsymbol{\beta}})} = (\Lambda^T)^{-1}Q^{(\tilde{\boldsymbol{\beta}})}\Lambda^{-1}$ where Λ^T denote the transpose of Λ. By definition of symmetric hierarchical reparametrizations, Λ satisfies (SL$_m$). Therefore Lemma 24 implies that Λ^{-1} satisfies (SL$_m$) and, consequently, Lemma 25 implies that $Q^{(\tilde{\boldsymbol{\beta}})}\Lambda^{-1}$ satisfies (S̃). Since Λ satisfies (SL$_m$), then Λ^T satisfies (SU$_m$) and thus Lemma 22 implies that $(\Lambda^T)^{-1}$ satisfies (SU$_m$). We can then apply Lemma 25 to $(\Lambda^T)^{-1}$ and $Q^{(\tilde{\boldsymbol{\beta}})}\Lambda^{-1}$ to deduce that $Q^{(\Lambda\tilde{\boldsymbol{\beta}})} = (\Lambda^T)^{-1}Q^{(\tilde{\boldsymbol{\beta}})}\Lambda^{-1}$ satisfies (S̃). $\qquad\square$

Proof of Lemma 14

Suppose $\boldsymbol{\beta}_T$ has zero mean (otherwise replace $\boldsymbol{\beta}_T$ by $\boldsymbol{\beta}_T - \mathbb{E}[\boldsymbol{\beta}]$). As in Section 4.4, given any r and t in T we denote $P(X_{\ell(t)} = t | X_{\ell(r)} = r)$ by $P(t|r)$. Using (H), for any $d \in \{0, \ldots, k-1\}$ and $t \in T_d$, we can write the full conditional expectation of β_t as

$$\mathbb{E}[\beta_t | \boldsymbol{\beta}_{T \setminus t}] = \sum_{r \prec t} A_{tr}\beta_r + \sum_{r \succ t} A_{tr}\beta_r\,,$$

where $A_{tr} = -\frac{Q_{tr}}{Q_{tt}}$ for any $r \neq t$. Note that (S̃) implies

$$A_{tr} = \frac{c_{\ell(t)\ell(r)}P(r \cap t)}{P(t)} = c_{\ell(t)\ell(r)}P(r|t)\,. \qquad (23)$$

It follows

$$\mathbb{E}[\beta_t | \boldsymbol{\beta}_{T \setminus t}] = \sum_{r \prec t} c_{d\ell(r)}P(r|t)\beta_r + \sum_{r \succ t} c_{d\ell(r)}P(r|t)\beta_r$$
$$= \sum_{\ell \neq d} c_{d\ell}\mathbb{E}[\beta_{X_\ell} | \boldsymbol{\beta}_T, X_d = t]\,.$$

38

Since the last equation does not depend on $\boldsymbol{\beta}^{(d)}$ we have $\mathbb{E}[\beta_t|\boldsymbol{\beta}_{T\setminus t}] = \mathbb{E}[\beta_t|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}]$. For any $r \in T_p$, by definition of $\phi_r^{(p)}\boldsymbol{\beta}^{(d)}$, we have

$$
\begin{aligned}
\mathbb{E}[\phi_r^{(p)}\boldsymbol{\beta}^{(d)}|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}] &= \sum_{t\in T_d} \Pr(t|r)\mathbb{E}[\beta_t|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}] \\
&= \sum_{t\in T_d} \Pr(t|r)\sum_{\ell\neq d} c_{d\ell}\mathbb{E}[\beta_{X_\ell}|\boldsymbol{\beta}_T, X_d = t] \\
&= \sum_{\ell\neq d} c_{d\ell}\sum_{t\in T_d} \Pr(t|r)\mathbb{E}[\beta_{X_\ell}|\boldsymbol{\beta}_T, X_d = t] \\
&= \sum_{\ell\neq d} c_{d\ell}\mathbb{E}[\beta_{X_\ell}|\boldsymbol{\beta}_T, X_p = r].
\end{aligned}
$$

From the latter equation and the definition of $\delta_r^{(p)}\boldsymbol{\beta}^{(d)}$ it follows

$$
\begin{aligned}
\mathbb{E}[\delta_r^{(p)}\boldsymbol{\beta}^{(d)}|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}] &= \mathbb{E}[\phi_r^{(p)}\boldsymbol{\beta}^{(d)}|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}] - \mathbb{E}[\phi_{pa(r)}^{(p-1)}\boldsymbol{\beta}^{(d)}|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}] \\
&= \sum_{\ell\neq d} c_{d\ell}\left(\mathbb{E}[\beta_{X_\ell}|\boldsymbol{\beta}_T, X_p = r] - \mathbb{E}[\beta_{X_\ell}|\boldsymbol{\beta}_T, X_{p-1} = pa(r)]\right) \\
&= \sum_{\ell\in\{p,\ldots,k-1\}\setminus d} c_{d\ell}\delta_r^{(p)}\boldsymbol{\beta}^{(\ell)}.
\end{aligned}
$$

Proof or Theorem 15

To prove Theorem 15 we first need the following lemma.

Lemma 26. *Given $d \in \{0,\ldots,k-1\}$ and $p, p' \in \{0,\ldots,d\}$ with $p \neq p'$,*

$$
\delta^{(p)}\boldsymbol{\beta}^{(d)}\perp\delta^{(p')}\boldsymbol{\beta}^{(d)}\,|\,\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}\,.
$$

Proof. Let $d \in \{0,\ldots,k-1\}$, $p, p' \in \{0,\ldots,d\}$ with $p < p'$ and $\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}$ be fixed. To make the notation more compact we denote $\mathbb{E}[\cdot|\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}]$ by $\tilde{\mathbb{E}}[\cdot]$, $\phi_r^{(p)}\boldsymbol{\beta}^{(d)}$ by $\tilde{\phi}_r^{(p)}$ and $P(X_{p'} = r'|X_p = r)$ by $P(r'|r)$ for all $r \in T_p$ and $r' \in T_{p'}$. By replacing β_t with $\beta_t - \tilde{\mathbb{E}}[\beta_t]$ we can suppose without loss of generality that $\tilde{\mathbb{E}}[\beta_t] = 0$ for all $t \in T_d$ and therefore $\tilde{\mathbb{E}}[\tilde{\phi}_r^{(p)}] = \tilde{\mathbb{E}}[\tilde{\phi}_{r'}^{(p')}] = 0$ and $\tilde{\mathbb{E}}[\delta_r^{(p)}\boldsymbol{\beta}^{(d)}] = \tilde{\mathbb{E}}[\delta_{r'}^{(p')}\boldsymbol{\beta}^{(d)}] = 0$ for all $r \in T_p$ and $r' \in T_{p'}$. By definition $\tilde{\phi}_r^{(p)} = \sum_{s\in T_{p'}} P(s|r)\tilde{\phi}_s^{(p')}$ and therefore

$$
\tilde{\mathbb{E}}[\tilde{\phi}_r^{(p)}\tilde{\phi}_{r'}^{(p')}] = \sum_{s\in T_{p'}} P(s|r)\tilde{\mathbb{E}}[\tilde{\phi}_s^{(p')}\tilde{\phi}_{r'}^{(p')}] = P(r'|r)\tilde{\mathbb{E}}[(\tilde{\phi}_{r'}^{(p')})^2]\,, \qquad (24)
$$

where we used $\tilde{\mathbb{E}}[\tilde{\phi}_s^{(p')}\tilde{\phi}_{r'}^{(p')}] = 0$ for $r' \neq s$ and $r', s \in T_{p'}$, which follows from the conditional independence of $(\beta_t)_{t\in T_d, t\succeq r'}$ and $(\beta_t)_{t\in T_d, t\succeq s}$ given $\boldsymbol{\beta}\setminus\boldsymbol{\beta}^{(d)}$.

Note that $P(r'|r)$ in (24) could be 0. From $\tilde{\mathbb{E}}[\beta_t] = 0$ for any $t \in T_d$ and $(\tilde{\text{S}})$ we have $\tilde{\mathbb{E}}[\beta_t^2] = P(t)^{-1}$ and therefore

$$
\begin{aligned}
\tilde{\mathbb{E}}[(\tilde{\phi}_{r'}^{(p')})^2] =& \tilde{\mathbb{E}}[(\sum_{t \in T_d} P(t|r')\beta_t)^2] = \sum_{t \in T_d} P(t|r')^2 \tilde{\mathbb{E}}[\beta_t^2] \\
=& \sum_{t \in T_d} \frac{P(t|r')^2}{P(t)} = \frac{1}{P(r')} \sum_{t \in T_d} P(t|r') = \frac{1}{P(r')} \,. \qquad (25)
\end{aligned}
$$

Combining (24) and (25) we have $\tilde{\mathbb{E}}[\tilde{\phi}_r^{(p)} \tilde{\phi}_{r'}^{(p')}] = 0$ if $r' \not\succ r$ and

$$
\tilde{\mathbb{E}}[\tilde{\phi}_r^{(p)} \tilde{\phi}_{r'}^{(p')}] = \frac{P(r'|r)}{P(r')} = \frac{1}{P(r)} \qquad\qquad \text{if } r' \succ r \,. \qquad (26)
$$

From the last equality and the definition of $\delta_r^{(p)} \boldsymbol{\beta}^{(d)}$ in (17) we have

$$
\begin{aligned}
&\tilde{\mathbb{E}}[\delta_r^{(p)} \boldsymbol{\beta}^{(d)} \delta_{r'}^{(p')} \boldsymbol{\beta}^{(d)}] = \\
&\tilde{\mathbb{E}}[(\tilde{\phi}_r^{(p)} - \tilde{\phi}_{pa(r)}^{(p-1)})(\tilde{\phi}_{r'}^{(p')} - \tilde{\phi}_{pa(r')}^{(p'-1)})] = \\
&\tilde{\mathbb{E}}[\tilde{\phi}_r^{(p)} \tilde{\phi}_{r'}^{(p')} - \tilde{\phi}_{pa(r)}^{(p-1)} \tilde{\phi}_{r'}^{(p')} - \tilde{\phi}_r^{(p)} \tilde{\phi}_{pa(r')}^{(p'-1)} + \tilde{\phi}_{pa(r)}^{(p-1)} \tilde{\phi}_{pa(r')}^{(p'-1)}] = \\
&\frac{P(r'|r)}{P(r')} - \frac{P(r'|pa(r))}{P(r')} - \frac{P(pa(r')|r)}{P(pa(r'))} + \frac{P(pa(r')|pa(r))}{P(pa(r'))} = 0,
\end{aligned}
$$

where the last equality is trivial if $pa(r) \not\prec r'$ and can be deduced from (26) otherwise. The desired conditional independence follows from $\tilde{\mathbb{E}}[\delta_r^{(p)} \boldsymbol{\beta}^{(d)} \delta_{r'}^{(p')} \boldsymbol{\beta}^{(d)}] = 0 = \tilde{\mathbb{E}}[\delta_r^{(p)} \boldsymbol{\beta}^{(d)}] \tilde{\mathbb{E}}[\delta_{r'}^{(p')} \boldsymbol{\beta}^{(d)}]$ for all $r \in T_p$ and $r' \in T_{p'}$. $\qquad \square$

Proof or Theorem 15. Theorem 15 follows easily from Lemmas 14 and 26 as follows. For each $d \in \{0, \ldots, k-1\}$ the sampling step

$$
\boldsymbol{\beta}^{(d)}(s+1) \sim \mathcal{L}\left(\boldsymbol{\beta}^{(d)} | (\boldsymbol{\beta}^{(\ell)}(s+1))_{0 \le \ell < d}, (\boldsymbol{\beta}^{(\ell)}(s))_{d < \ell \le k-1}\right) \qquad (27)
$$

in Sampler GS($\boldsymbol{\beta}_T$) is equal in distribution to sampling jointly the (d+1) residuals

$$
(\delta^{(p)} \boldsymbol{\beta}^{(d)}(s+1))_{p=0}^d \sim \mathcal{L}\left((\delta^{(p)} \boldsymbol{\beta}^{(d)})_{p=0}^d | (\boldsymbol{\beta}^{(\ell)}(s+1))_{0 \le \ell < d}, (\boldsymbol{\beta}^{(\ell)}(s))_{d < \ell \le k-1}\right) \,.
$$

From the conditional independence statement in Lemma 26 the latter is equivalent to sampling independently each residual $\delta^{(p)} \boldsymbol{\beta}^{(d)}(s+1)$ from

$$
\mathcal{L}\left(\delta^{(p)} \boldsymbol{\beta}^{(d)} | (\boldsymbol{\beta}^{(\ell)}(s+1))_{0 \le \ell < d}, (\boldsymbol{\beta}^{(\ell)}(s))_{d < \ell \le k-1}\right) \,.
$$

Moreover, from Lemma 14

$$
\mathcal{L}\left(\delta^{(p)} \boldsymbol{\beta}^{(d)} | \boldsymbol{\beta} \backslash \boldsymbol{\beta}^{(d)}\right) = \mathcal{L}\left(\delta^{(p)} \boldsymbol{\beta}^{(d)} | (\delta^{(p)} \boldsymbol{\beta}^{(\ell)})_{\ell \in \{p,\ldots,k-1\} \backslash d}\right) \,.
$$

Therefore the original sampling step in (27) is equivalent to sampling independently

$$\delta^{(p)}\boldsymbol{\beta}^{(d)}(s+1) \ \sim \ \mathcal{L}\left(\delta^{(p)}\boldsymbol{\beta}^{(d)}|\big(\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s+1)\big)_{p\leq\ell<d}, \big(\delta^{(p)}\boldsymbol{\beta}^{(\ell)}(s)\big)_{d<\ell\leq k-1}\right),$$
(28)

for $p = 0,\ldots,d$. The thesis follows from the equivalence between (27) and (28). $\qquad\square$

Proof of Corollary 16

Proof of Corollary 16. The map from $\boldsymbol{\beta}_T$ to $(\delta^{(0)}\boldsymbol{\beta},\ldots,\delta^{(k-1)}\boldsymbol{\beta})$ is an injective linear transformation. The injectivity holds because for any $d \in \{0,\ldots,k-1\}$ and $t \in T_d$ we can reconstruct β_t from $(\delta^{(0)}\boldsymbol{\beta}^{(d)},\ldots,\delta^{(d)}\boldsymbol{\beta}^{(d)})$

$$\sum_{r\preceq t}\delta_r^{(\ell(r))}\boldsymbol{\beta}^{(d)} = \phi_{t_0}^{(0)}\boldsymbol{\beta}^{(d)} + \sum_{t_0\prec r\preceq t}\left(\phi_r^{(\ell(r))}\boldsymbol{\beta}^{(d)} - \phi_{pa(r)}^{(\ell(r)-1)}\boldsymbol{\beta}^{(d)}\right) = \phi_t^{(d)}\boldsymbol{\beta}^{(d)} = \beta_t.$$

It follows that $(\delta\boldsymbol{\beta}(s))_{s\in\mathbb{N}} = (\delta^{(0)}\boldsymbol{\beta}(s),\ldots,\delta^{(k-1)}\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$ is a Markov chain with the same rate of convergence of the original chain $(\boldsymbol{\beta}(s))_{s\in\mathbb{N}}$. Then the thesis follows from Theorem 15 and the fact that the rate of convergence of a collection of independent Markov chains equals the supremum of the rates of convergence of the single chains. $\qquad\square$

Proof of Theorem 17

Proof of Theorem 17. We are interested in the rate of convergence of the blocked sampler targeting $\delta^{(p)}\boldsymbol{\beta} = (\delta^{(p)}\boldsymbol{\beta}^{(p)},\ldots,\delta^{(p)}\boldsymbol{\beta}^{(k-1)})$ and evolving according to (18). Consider first the case $p \in \{1,\ldots,k-1\}$. Note that $\delta^{(p)}\boldsymbol{\beta}$ has a singular variance-covariance matrix because for each $t \in T_{p-1}$ and $d \in \{p,\ldots,k-1\}$ it follows from (16) and (17) that

$$\sum_{r\in ch(t)} P(r|t)\delta_r^{(p)}\boldsymbol{\beta}^{(d)} = 0 \tag{29}$$

and therefore some elements of $(\delta^{(p)}\boldsymbol{\beta})$ are linear combinations of the others. In order to use standard tools it is more convenient to work with non-singular Gaussian random vectors. To do so it is sufficient to consider a sub-vector of $\delta^{(p)}\boldsymbol{\beta}$ obtained by removing from T_p one children node for each parent node in T_{p-1}. More formally, let f be an arbitrary map from T_{p-1} to T_p such that $f(t) \in ch(t)$ for all $t \in T_{p-1}$ and then define the subset $T_p' \subseteq T_p$ as $T_p' = T_p \backslash f(T_{p-1})$. It is then easy to see that the resulting sub-vector $\delta_{T_p'}^{(p)}\boldsymbol{\beta} = (\delta_{T_p'}^{(p)}\boldsymbol{\beta}^{(p)},\ldots,\delta_{T_p'}^{(p)}\boldsymbol{\beta}^{(k-1)})$ with $\delta_{T_p'}^{(p)}\boldsymbol{\beta}^{(d)} = (\delta_r^{(p)}\boldsymbol{\beta}^{(d)})_{r\in T_p'}$ for all $d \in \{p,\ldots,k-1\}$ has an invertible variance-covariance matrix. Moreover, since each $\delta^{(p)}\boldsymbol{\beta}^{(d)}$ is a function of the corresponding $\delta_{T_p'}^{(p)}\boldsymbol{\beta}^{(d)}$ via (29) it follows

that the blocked sampler targeting $\delta^{(p)}\boldsymbol{\beta}$ and evolving according to (18) is equivalent in distribution to a blocked Gibbs Sampler targeting $\delta^{(p)}_{T'_p}\boldsymbol{\beta}$ and evolving according to

$$\delta^{(p)}_{T'_p}\boldsymbol{\beta}^{(d)}(s+1) \sim \mathcal{L}\left(\delta^{(p)}_{T'_p}\boldsymbol{\beta}^{(d)}|(\delta^{(p)}_{T'_p}\boldsymbol{\beta}^{(\ell)}(s+1))_{p\leq\ell<d},(\delta^{(p)}_{T'_p}\boldsymbol{\beta}^{(\ell)}(s))_{d<\ell\leq k-1}\right),$$
(30)

for $d \in \{p,\dots,k-1\}$. Let $A_{\delta^{(p)}_{T'_p}\boldsymbol{\beta}} = (A_{\delta^{(p)}_r\boldsymbol{\beta}^{(d)}\,\delta^{(p)}_{r'}\boldsymbol{\beta}^{(d')}})_{r,r'\in T'_p,\,d,d'\in\{p,\dots,k-1\}}$ be the A-matrix associated to the Gibbs sampler in (30), defined by

$$\mathbb{E}[\delta^{(p)}_r\boldsymbol{\beta}^{(d)}|\delta^{(p)}_{T'_p}\boldsymbol{\beta}\backslash\delta^{(p)}_r\boldsymbol{\beta}^{(d)}] - \mathbb{E}[\delta^{(p)}_r\boldsymbol{\beta}^{(d)}] =$$

$$\sum_{\delta^{(p)}_{r'}\boldsymbol{\beta}^{(d')}\in\delta^{(p)}_{T'_p}\boldsymbol{\beta}\backslash\delta^{(p)}_r\boldsymbol{\beta}^{(d)}} A_{\delta^{(p)}_r\boldsymbol{\beta}^{(d)}\,\delta^{(p)}_{r'}\boldsymbol{\beta}^{(d')}}\left(\delta^{(p)}_{r'}\boldsymbol{\beta}^{(d')} - \mathbb{E}[\delta^{(p')}_{r'}\boldsymbol{\beta}^{(d')}]\right).\quad (31)$$

See [Roberts and Sahu, 1997, Sec.2.2] for more details on the properties of such A-matrices. Then Lemma 14 implies that $A_{\delta^{(p)}_r\boldsymbol{\beta}^{(d)}\,\delta^{(p)}_{r'}\boldsymbol{\beta}^{(d')}} = c_{dd'}$ if $r = r'$ and $d' \in \{p,\dots,k-1\}\backslash d$ and 0 otherwise. The latter is equivalent to the equation

$$A_{\delta^{(p)}_{T'_p}\boldsymbol{\beta}} = \left(C^{(p)} - \mathbb{I}_{k-p}\right)\otimes\mathbb{I}_{|T'_p|}$$
(32)

where $C^{(p)}$ is the $(k-p)\times(k-p)$ square matrix $C^{(p)} = (c_{dd'})_{d,d'=p}^{k-1}$, $\mathbb{I}_n$ denotes the n dimensional identity matrix, $\otimes$ denotes the Kronecker product of matrices and $|T'_p| = |T_p| - |T_{p-1}|$ is the cardinality of T'_p. Then note that from Roberts and Sahu [1997, Thm.1] the rate of convergence of the Gibbs sampler in (30) equals the largest modulus eigenvalue of $B = (\mathbb{I}_{(k-p)|T'_p|} - L)^{-1}U$, where L is the lower triangular part of $A_{\delta^{(p)}_{T'_p}\boldsymbol{\beta}}$ and $U = A_{\delta^{(p)}_{T'_p}\boldsymbol{\beta}} - L$.
Using basic properties of the Kroenecker product we can see that

$$\begin{aligned} B =& (\mathbb{I}_{(k-p)}\otimes\mathbb{I}_{|T'_p|} - \tilde{L}\otimes\mathbb{I}_{|T'_p|})^{-1}\left(\tilde{U}\otimes\mathbb{I}_{|T'_p|}\right)\\ =& \left((\mathbb{I}_{(k-p)} - \tilde{L})\otimes\mathbb{I}_{|T'_p|}\right)^{-1}\left(\tilde{U}\otimes\mathbb{I}_{|T'_p|}\right)\\ =& \left((\mathbb{I}_{(k-p)} - \tilde{L})^{-1}\otimes\mathbb{I}_{|T'_p|}\right)\left(\tilde{U}\otimes\mathbb{I}_{|T'_p|}\right)\\ =& \left((\mathbb{I}_{(k-p)} - \tilde{L})^{-1}\tilde{U}\right)\otimes\mathbb{I}_{|T'_p|} \end{aligned}$$

where $\tilde{L}$ is the lower triangular part of $(C^{(p)}-\mathbb{I}_{k-p})$ and $\tilde{U} = (C^{(p)}-\mathbb{I}_{k-p})-\tilde{L}$. From $B = \tilde{B}\otimes\mathbb{I}_{|T'_p|}$, where $\tilde{B} = (\mathbb{I}_{(k-p)} - \tilde{L})^{-1}\tilde{U}$, it follows that the unique eigenvalues of B are the same as the unique eigenvalues of $\tilde{B}$ and thus the largest modulus eigenvalue of B equals the one of $\tilde{B}$. The case $p = 0$ is analogous (with no need to consider a sub-vector of $\delta^{(0)}\boldsymbol{\beta}$ and T'_0 being equal to T_0 itself) and trivial to check. $\qquad\square$

Proof of Theorem 18

Theorem 15 shows that $(\delta^{(p)}\boldsymbol{\gamma}(s))_s$ for $p \in \{0, \ldots, k-1\}$ are k independent Markov chains. Arguing as in the proof of Theorem 17 above for each p we consider $(\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s))_s$ rather than $(\delta^{(p)}\boldsymbol{\gamma}(s))_s$ to avoid working with singular Gaussian random vectors. For any p from 0 to $k-1$, (T) implies that the Q-matrix of $\delta^{(p)}_{T'_p}\boldsymbol{\gamma}$ is tridiagonal (with $k-p$ blocks corresponding to $\delta^{(p)}_{T'_p}\boldsymbol{\gamma}^{(p)}$ up to $\delta^{(p)}_{T'_p}\boldsymbol{\gamma}^{(k-1)}$). Therefore Theorem 5 of [Roberts and Sahu, 1997] implies that the rate of convergence of the Gibbs Sampler in (30) is given by $\rho(\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s)) = \lambda(A_{\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s)})^2$ where $A_{\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s)}$ is the A-matrix defined as in (31) and $\lambda(A_{\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s)})$ is its largest eigenvalue. Then, using (32) from the proof of Theorem 17, it follows $\lambda(A_{\delta^{(p)}_{T'_p}\boldsymbol{\gamma}(s)})^2 = \lambda(C^{(p)} - \mathbb{I}_{k-p})^2$ and thus the rate of convergence of $(\delta^{(p)}\boldsymbol{\gamma}(s))_s$ is given by $\rho(\delta^{(p)}\boldsymbol{\gamma}(s)) = \lambda(C^{(p)} - \mathbb{I}_{k-p})^2$. Noting that $C^{(p+1)} - \mathbb{I}_{k-(p+1)}$ is obtained from $C^{(p)} - \mathbb{I}_{k-p}$ by removing the first row and column, the desired inequality $\rho(\delta^{(p)}\boldsymbol{\gamma}(s)) = \lambda(C^{(p)} - \mathbb{I}_{k-p})^2 \geq \lambda(C^{(p+1)} - \mathbb{I}_{k-(p+1)})^2 = \rho(\delta^{(p+1)}\boldsymbol{\gamma}(s))$ follows by applying the Cauchy interlacing theorem (see e.g. Bhatia, 2013), which states that the eigenvalues of a principal submatrix of a symmetric matrix interlace the original eigenvalues.

Proof of Theorems 9 and 21

Since Theorem 9 is a special case of Theorem 21 we only need to prove the latter. Let Λ be defined as in the statement of Theorem 21. The precision matrix of the hierarchical reparametrization $\Lambda\boldsymbol{\beta}_T$ is given by $D = (\Lambda^T)^{-1}Q\Lambda^{-1}$ where Q is the precision matrix of $\boldsymbol{\beta}_T$. To show that $\Lambda\boldsymbol{\beta}_T$ is made of independent random variables it suffices to show that D is diagonal. In particular we show that $\Lambda^T D \Lambda = Q$ with D being a diagonal matrix with elements defined by (21). Since $\Lambda^T D \Lambda$ is symmetric, we can consider $(\Lambda^T D \Lambda)_{rt}$ assuming $t \succeq r$ without loss of generality. Since D is diagonal and $\Lambda = (\lambda_{ss'})_{s,s' \in T}$ is non-zero only for $s' \in \{s, pa(s)\}$ we have

$$(\Lambda^T D \Lambda)_{rt} = \sum_{r' \in T} \sum_{s \in T} \lambda_{sr} D_{sr'} \lambda_{r't}$$
$$= \sum_{r' \in T} \lambda_{r'r} D_{r'r'} \lambda_{r't} = D_{rr}\lambda_{rt} + \sum_{r' \in ch(r)} \lambda_{r'r} D_{r'r'} \lambda_{r't}.$$

Therefore $(\Lambda^T D \Lambda)_{rt} = 0$ unless $r = t$ or $t = pa(r)$. In the latter case $(\Lambda^T D \Lambda)_{r\,pa(r)} = D_{rr}\lambda_{rpa(r)} = D_{rr}\frac{Q_{r\,pa(r)}}{D_{rr}} = Q_{r\,pa(r)}$. In the former case (i.e. $r = t$), $(\Lambda^T D \Lambda)_{rr} = D_{rr} + \sum_{r' \in ch(r)} \lambda_{r'r}^2 D_{r'r'} = Q_{rr}$, where we used (21) for D_{rr}. It follows $\Lambda^T D \Lambda = Q$ as desired.