Biometrika (2015), **99**, 1, *pp*. 1–17

Printed in Great Britain

Advance Access publication on 31 July 2015

Scalable inference for crossed random effects models

BY O. PAPASPILIOPOULOS

Institució Catalana de Recerca i Estudis Avançats, Ramon Trias Fargas 25-27, Barcelona 08005

omiros.papaspiliopoulos@upf.edu

G.O. ROBERTS

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

gareth.o.roberts@warwick.ac.uk

AND G. ZANELLA

Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Roentgen 1, 20136 Milan, Italy

giacomo.zanella@unibocconi.it

SUMMARY

We develop methodology and complexity theory for Markov chain Monte Carlo algorithms used in inference for crossed random effect models in modern analysis of variance. We consider a plain Gibbs sampler and a simple modification we propose here, a collapsed Gibbs sampler. Under some balancedness assumptions on the data designs and assuming that precision hyperparameters are known, we demonstrate that the plain Gibbs sampler is not scalable, in the sense that its complexity is worse than proportional to the number of parameters and data, but that the collapsed Gibbs sampler is scalable. In simulated and real datasets we show that the explicit convergence rates our theory predicts match remarkably the computable but non-explicit rates in cases where the design assumptions are violated. We also show empirically that the collapsed Gibbs sampler, extended to sample precision hyperparameters, outperforms significantly, often by orders of magnitude, alternative state of the art algorithms. Supplementary material includes some proofs, additional simulations, implementation details and the R code to implement the algorithms considered in the article.

Some key words: Bayesian computation, analysis of variance, Gibbs sampler, spectral gap

1. INTRODUCTION

Crossed random effect models are additive models that relate a response variable to categorical predictors. In the literature they appear under various names, e.g. crossclassified data, variance component models or multiway analysis of variance. They provide the canonical framework for understanding the relative importance of different sources of variation in a data set as argued in Gelman (2005). For the purposes of this article we focus on linear models according to which

$$y_{i_1\cdots i_K} \sim N\left\{a^{(0)} + a^{(1)}_{i_1} + \cdots + a^{(K)}_{i_K}, (n_{i_1\cdots i_K}\tau_0)^{-1}\right\}, \quad i_k = 1, \ldots, I_k, \quad k = 1, \ldots, K \tag{1}$$

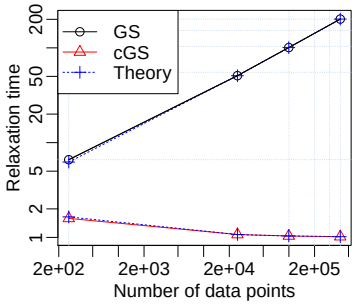
where $a^{(k)}$ is a vector of I_k levels, $a^{(k)}_{i_k}$, corresponding to the k-th categorical factor and $a^{(0)}$ is a global mean with $I_0 = 1$ level. These factors might correspond to both main and interaction effects in categorical data analysis. We work with exchangeable Gaussian random effects, $a^{(k)}_j \sim N(0, 1/\tau_k)$, for $k > 0$, which is by far the most standard choice, although interesting alternative priors exist as in Volfovsky & Hoff (2014). The precision terms $(\tau_0, \ldots, \tau_K)$ are typically unknown and are estimated either using fully Bayesian or optimization-based methods. In the notation of (1) we allow $n_{i_1 \cdots i_K} = 0$, which corresponds to empty cells. The total number of factor levels, hence of regression parameters, is denoted by $p = \sum_{k=0}^{K} I_k$, and the number of observations by $N = \sum_{i_1 \cdots i_K} n_{i_1 \cdots i_K}$. Crossed random effect models adapt naturally to modern high-dimensional but sparse data. For example, they are used in the context of recommender systems where in the simplest setup there are two factors, customers and products, and the response is a rating; the examples in Gao & Owen (2017) are such that $p \ll N \ll I_1 \times I_2$.

We say a data design has balanced levels if the same number of observations are made at each level of each factor, but this number can vary with factor, see Section 2·1 for a mathematical definition. The design has balanced cells if the same number of observations are available for each combination of factor levels, i.e., at each cell of the contingency table defined by the categorical predictors. By construction, a design with balanced cells has also balanced levels.

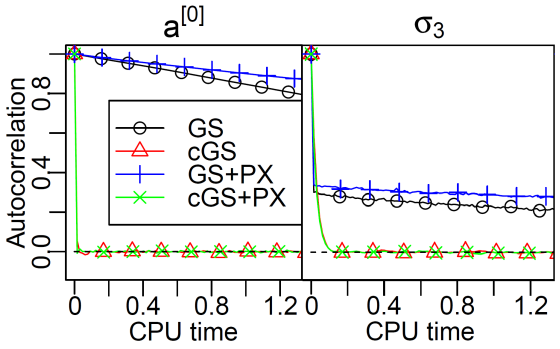
We are interested in scalable methodologies to perform likelihood-based inferences for crossed random effect models. The main computational bottleneck is the need to perform an integration over the high-dimensional space of factors. This is needed either for a fully Bayesian inference that places priors on the precision hypeparameters or for optimization methods using the marginal likelihood. An exact marginalisation is possible in the linear model due to the joint Gaussian distribution of responses and factors. However, this involves matrix operations whose cost is $\mathcal{O}(p^3)$, which can be prohibitively large in modern applications. For example for typical recommender systems we have $I_1 \times I_2 \geq N$ and thus $p \geq \max\{I_1, I_2\} \geq \sqrt{N}$, meaning that the cost of these operations is at least $\mathcal{O}(N^{3/2})$, which can be infeasible for large datasets. See also Section 1 of Gao & Owen (2019) for related discussion and references. Depending on the data design, the precision matrices of the Gaussian distributions involved may be sparse, in which case black-box sparse linear algebra algorithms can be used for the matrix operations in order to reduce the computational cost. However we are not aware of theoretical results that can be applied to context of crossed random effect models guaranteeing that the resulting computational complexity can be reduced to, e.g., $\mathcal{O}(N)$. Additionally, we are interested in exploring methodologies that could be extended to non-Gaussian models, e.g., those with categorical or count observations.

Markov chain Monte Carlo can be used to carry out the integration over factors. A popular and convenient algorithm to update the factors given the precision terms is the Gibbs sampler, which samples the factors $a^{(k)}$ iteratively from their full conditional distributions. Recently, Gao & Owen (2017) showed that, in the special case that $a^{(0)}$ is assumed known, in the context of recommendation where $K = 2$ with balanced cell design, the Gibbs Sampler has complexity $\mathcal{O}(N^{3/2})$. The complexity of a Markov chain Monte Carlo algorithm can be defined as the product of the computational time per iteration and the number of iterations the algorithm needs to mix. The argument in Gao & Owen (2017) suggested that the Gibbs sampler is not scalable for crossed random effects models due to its superlinear cost in the number of observations.

In this article we develop the theory for analysing the complexity of the Gibbs sampler for crossed random effect models under different designs. We propose a small modification of the basic algorithm, the collapsed Gibbs sampler, which we analyse too, and establish rigorously its superior performance and scalability. We obtain explicit results on the spectral gap of the



(a) Relaxation time versus number of datapoints for $K = 2$, $I_1 = I_2$ and $\tau_0 = \tau_1 = \tau_2$. Data are missing at completely random with probability of missingness 0.9. The exact relaxation times and those predicted by our theory are shown.



(b) Autocorrelation function of $a^{(0)}$ and $\sigma_3 = \tau_3^{-1/2}$ for the Gibbs Sampler (GS), collapsed Gibbs Sampler (cGS) and their combinations with parameter expansion (+PX) when applied to the *InstEval* dataset. See Section 6·2 for details on the set-up.

Fig. 1. Comparison between the Gibbs Sampler (GS) and its collapsed version (cGS).

algorithms in Section 5 and analyse their computational cost in the Appendix. The essence of the methodology we develop in this article is shown in Figure 1(a), details of which are given in Section 6·1. The figure highlights different aspects of our results. We consider modern big data asymptotic regimes where both the number of parameters and observations grow. The relaxation time of the Gibbs sampler and the collapsed Gibbs sampler, see Section 4·2 for definition and interpretation of relaxation time, can be computed numerically and are plotted versus the size of the datasets. It is evident the slowing down of the Gibbs sampler and the improvement of the collapsed Gibbs sampler with increasing data sizes. Our theory is not applicable in these cases since the resultant designs, which have been generated randomly, do not have balanced levels, still the rate that our theory predicts matches remarkably the correct rates. We obtain comparable results in a well-known real dataset of student evaluations with 5 factors in Section 6·2.

The complexity theory developed in the first part of the article relies on assuming the precision hyperparameters being known. However, the collapsed Gibbs sampler methodology can be easily extended to provide an algorithm for fully Bayesian inference when precision hyperparameters are given prior distributions. In Section 6·2 we compare numerically the performances of the resulting sampler with state of the art Markov chain Monte Carlo algorithms, such as parameter expansion and Hamiltonian Monte Carlo. We find that the collapsed Gibbs sampler we propose has far superior performance, with an improvement in effective sample size per unit of computation time ranging from 1 to 3 orders of magnitude depending on the parameter under consideration. Figure 1(b) highlights some of the results for a dataset analyzed in Section 6·2. The collapsed Gibbs Sampler provides a dramatic decrease in autocorrelation compared to the plain Gibbs Sampler, for both factors and precision terms, while no significant benefit is obtained from parameter expansion.

The theory is based upon a multigrid decomposition of the Markov chain generated by the sampler, which allows us to identify the slowest mixing components, and capitalises on existing theory for the convergence of Gaussian Markov chains. The multigrid decomposition of a Markov chains is a powerful theoretical tool for studying its spectral gap, since it provides its decomposition into independent processes. Identifying such decomposition is a kind of art; a previously successful example is in Zanella & Roberts (2017) in the context of multilevel nested linear models.

2. DECOMPOSITIONS OF THE POSTERIOR DISTRIBUTION

2·1. *Notation*

The statistical model we work with is described in (1). In accordance with standard practice Gaussian priors are used for the factor levels, $a_j^{(k)} \sim N(0, 1/\tau_k)$, and an improper prior for the global mean, $p(a^{(0)}) \propto 1$. When convenient we write $a_{i_0}^{(0)}$, which is is the same as $a^{(0)}$. We allow $n_{i_1 \cdots i_K} = 0$, which corresponds to empty cells in the contingency table defined by the outer product of the categorical factors. With n we denote the data incidence array, a multidimensional array with elements $n_{i_1 \cdots i_K}$. Two-dimensional marginal tables extracted from the data incidence matrix are denoted by $n^{(l,k)}$ and have elements $n_{i_l, i_k}^{(l,k)}$, which is the total number of observations on level i_l of factor l and i_k of factor k; margins of this table are denoted by $n^{(k)}$, and are vectors of size I_k with elements $n_j^{(k)}$, which is the total number of observations with level j on the k-th factor. By definition $\sum_j n_j^{(k)} = N$, where N is the total number of observations. A data design has balanced levels if $n_j^{(k)} = N/I_k$ for every k and j, and balanced cells if $n_{i_1 \cdots i_K} = N/\prod_k I_k$ for all combinations of factor levels.

Averages of vectors are denoted by an overline, e.g., $\bar{a}^{(k)}$; weighted averages are denoted by a tilde, e.g., $\tilde{y} = \sum_{i_1 \cdots i_K} y_{i_1 \cdots i_K} n_{i_1 \cdots i_K} / N$. The vector of all factor averages is denoted by $\bar{a}$, the first element of which is trivially $a^{(0)}$. We use $a^{(-k)}$ to denote the vector of all factor levels except those of $a^{(k)}$ and $a_{-j}^{(k)}$ to denote the vector of all levels of factor k except the j-th level; a denotes the vector of all levels of all factors. We define δ to be a residual operator that when applied to a vector returns the difference of its elements from their sample average, e.g., $\delta a^{(k)}$ has elements $a_j^{(k)} - \bar{a}^{(k)}$ and is referred to as the factor's level increments; δa denotes the vector of all such increments, except $\delta a^{(0)}$ which is 0 trivially.

The law of a random variable X is denoted by $\mathcal{L}(X)$, e.g., $\mathcal{L}(a_j^{(k)}) = N(0, 1/\tau_k)$, and that of X conditionally on Y by $\mathcal{L}(X \mid Y)$. When a joint distribution has been specified for X and other random variables, $\mathcal{L}\{X \mid \cdot\}$ denotes the full conditional distribution of X conditionally on the rest. In the following sections up to Section 6 we assume the precision terms to be fixed without explicitly writing the conditioning on τ in all expressions.

2·2. *Full conditional distributions*

Fairly standard Bayesian linear model calculations yield that the conditional distribution of $a^{(0)}$ given all other parameters and data is

$$\mathcal{L}\left\{a^{(0)} \mid \cdot\right\} = N\left\{\tilde{y} - \frac{\sum_k \sum_i a_i^{(k)} n_i^{(k)}}{N}, (N\tau_0)^{-1}\right\}. \tag{2}$$

With balanced levels this simplifies to

$$\mathcal{L}\left\{a^{(0)} \mid \cdot\right\} = N\left\{\tilde{y} - \sum_k \bar{a}^{(k)}, (N\tau_0)^{-1}\right\}. \tag{3}$$

Similarly we obtain that for $k > 0$

$$\mathcal{L}\left\{a_j^{(k)} \mid \cdot\right\} = N\left\{\frac{n_j^{(k)}\tau_0}{n_j^{(k)}\tau_0 + \tau_k}\left(\tilde{y}_j^{(k)} - a^{(0)} - \frac{\sum_{l \neq k, l \neq 0} \sum_i a_i^{(l)} n_{j,i}^{(k,l)}}{n_j^{(k)}}\right), (n_j^{(k)}\tau_0 + \tau_k)^{-1}\right\}, \tag{4}$$

where $\tilde{y}_j^{(k)}$ is the weighted average of all observations for which their level on factor k is j. With balanced cells this simplifies to

$$\mathcal{L}\left\{a_j^{(k)} \mid \cdot\right\} = N\left\{\frac{N\tau_0}{N\tau_0 + I_k\tau_k}\left(\bar{y} - \sum_{l\neq k}\bar{a}^{(l)}\right), I_k(N\tau_0 + I_k\tau_k)^{-1}\right\}. \tag{5}$$

2·3. *Factorisations*

In balanced levels designs the posterior distribution of regression parameters admits certain factorisation, which are collected together in the following Proposition. Throughout the paper, we use products of laws to denote independence.

PROPOSITION 1. *For balanced levels designs*

$$\mathcal{L}\left\{\bar{a}, \delta a \mid y\right\} = \mathcal{L}\left\{\bar{a} \mid y\right\}\mathcal{L}\left\{\delta a \mid y\right\},$$

and

$$\mathcal{L}\left\{\bar{a}^{(-0)} \mid y\right\} = \prod_{k=1}^{K}\mathcal{L}\left\{\bar{a}^{(k)} \mid y\right\}. \tag{6}$$

For balanced cells designs we have further

$$\mathcal{L}\left\{\delta a \mid y\right\} = \prod_k \mathcal{L}\left\{\delta a^{(k)} \mid y\right\}.$$

The factorisation in (6) is particularly relevant to the collapsed Gibbs sampler we introduce later in the article. A sketch of the proof is the following. For the first factorisation, directly from (4) with the assumption of balanced levels we obtain that

$$\mathcal{L}\left\{\bar{a}^{(k)} \mid y, a^{(-k)}, \delta a^{(k)}\right\} = N\left\{\frac{N\tau_0}{N\tau_0 + I_k\tau_k}\left(\tilde{y} - a^{(0)} - \sum_{l\neq k}\bar{a}^{(l)}\right), (N\tau_0 + I_k\tau_k)^{-1}\right\}. \tag{7}$$

We use the fact that global and local Markovian properties are equivalent, see, e.g., Section 3 of Besag (1974). This yields the first independence statement in the proposition. The proof of (6) follows by similar arguments using $\mathcal{L}\left\{\bar{a}^{(-0,-k)} \mid y\right\} = N(0, (I_k\tau_k)^{-1})$. The third factorisation is argued in the same way noting that (5) implies that

$$\mathcal{L}\left\{\delta a^{(k)} \mid y, a^{(-k)}, \delta a^{(-k)}\right\} = N\left\{(0,\ldots,0)^{\mathrm{T}}, (N\tau_0 + I_k\tau_k)^{-1}\left(I_k\mathbb{I}_{I_k} - \mathbb{H}_{I_k}\right)\right\},$$

where $\mathbb{I}_{I_k}$ denotes the $I_k \times I_k$ identity matrix and $\mathbb{H}_{I_k}$ the $I_k \times I_k$ matrix with each entry equal to 1.

3. GIBBS SAMPLERS FOR INFERENCE

We consider two main algorithms in this paper. The first is a block Gibbs sampler that updates in a single block the levels of a given factor conditional on everything else. Due to the dependence structure in the model, the levels of a given factor conditional on the rest are independent, hence the sampling is done separately for each factor level, i.e., iteratively from $\mathcal{L}\left\{a_{i_k}^{(k)} \mid \cdot\right\}$,

for $i_k = 1, \ldots, I_k$, and $k = 0, \ldots, K$; these distributions are specified in Section 2·2. We refer to this algorithm as the Gibbs sampler, although it should be understood that it is just one implementation of the scheme.

We also consider a collapsed Gibbs sampler that samples from $\mathcal{L}\left\{a^{(-0)} \mid y\right\}$, i.e., the algorithm that is obtained by first analytically integrating out the global mean $a^{(0)}$, and then sampling in blocks the levels of each of the remaining factors. In practice, we implement this algorithm by sampling iteratively from $\mathcal{L}\left\{a^{(0)}, a^{(k)} \mid \cdot\right\}$, for $k = 1, \ldots, K$. Updating $a^{(0)}$ together with each block is equivalent to integrating it out before sampling starts, in the sense that the resulting transition kernel for $a^{(-0)}$ is the same, with the only difference being whether the values of $a^{(0)}$ are stored or not. In our implementation we first sample $\mathcal{L}\left\{a^{(0)} \mid y, a^{(-0,-k)}\right\}$, and then $\mathcal{L}\left\{a_{i_k}^{(k)} \mid \cdot\right\}$ for $i_k = 1, \ldots, I_k$ as in the original Gibbs Sampler. The implementation of the collapsed Gibbs sampler relies on the following result.

PROPOSITION 2. *Denoting* $s_j^{(k)} = n_j^{(k)}\tau_0/(\tau_k + n_j^{(k)}\tau_0)$, *then*

$$\mathcal{L}\left\{a^{(0)} \mid y, a^{(-0,-k)}\right\} = N\left\{\frac{1}{\sum_j s_j^{(k)}} \sum_j s_j^{(k)} \left(\tilde{y}_j^{(k)} - \frac{\sum_{l \neq k} \sum_i a_i^{(l)} n_{j,i}^{(k,l)}}{n_j^{(k)}}\right), \frac{1}{\tau_k \sum_j s_j^{(k)}}\right\}. \tag{8}$$

The reason why we prefer to present the collapsed Gibbs sampler in this way where $a^{(0)}$ is updated together with each block, is because our preferred version is still realisable in more elaborate models, e.g., generalised linear crossed random effects models. In such extensions exact sampling from $\mathcal{L}\left\{a^{(0)}, a^{(k)} \mid \cdot\right\}$ might not be feasible, but a Metropolis-Hastings step can be used instead. Additionally, it requires a minimal modification of the Gibbs sampler code to implement, as shown in the supplementary material.

4. COMPLEXITY OF MARKOV CHAIN MONTE CARLO

4·1. *Notation*

For the stochastic processes generated by Markov chain Monte Carlo the time index corresponds to iteration, which is generically denoted by t, and it is included in parentheses, e.g., $x(t)$; in such a case the stochastic process over T iterations is denoted by $\{x(t)\}_{t=1}^{T}$; we write $\{x(t)\}$ when $T = \infty$; we write $\{(x, z)(t)\}$ to denote a stochastic process that at each time t takes as value the vector composed by $x(t)$ and $z(t)$. We say that the stochastic process $\{x(t)\}$ is a timewise transformation of another $\{y(t)\}$ if there is a function ϕ such that $x(t) = \phi\{y(t)\}$ for all t.

4·2. *Spectral gap and relaxation time*

In this article we focus on $L^2(\pi)$ convergence, which relies on functional analytic concepts, a very high level description of which are given below. For a given target distribution π defined on a state space $\mathcal{X}$, we define $L^2(\pi)$ to be the space of complex-valued functions that are square-integrable with respect to π. We define the inner product in this space such that the associated norm of a function $f : \mathcal{X} \to C$ is $\|f\|^2 = \int_{\mathcal{X}} |f(x)|^2 \pi(dx)$. For a Markov chain $\{x(t)\}$ defined on $\mathcal{X}$ with transition kernel P that is invariant with respect to π, we view P as an integral operator on $L^2(\pi)$ and denote its spectrum by S. We say that P converges geometrically fast to π in $L^2(\pi)$ norm, also known as operator norm, if and only if its geometric rate of convergence, defined as $\sup_{\lambda \in S} |\lambda|$, is less than 1. The spectral gap of P is defined as the difference between 1 and the

geometric rate of convergence, hence a Markov chain converges in $L^2(\pi)$ norm if and only it has positive spectral gap. In the remainder of the paper we refer to the geometric rate of convergence simply as rate of convergence for brevity.

Define the relaxation time of a Markov chain as the reciprocal of its spectral gap. This can be interpreted as the number of iterations needed to subsample the Markov chain so that the resultant draws are roughly independent of each other. The complexity of a Markov chain Monte Carlo algorithm can be defined as the product of the relaxation time and the cost per iteration.

4·3. *The spectral gap of the Gibbs sampler on Gaussian distributions*

The Markov chain $\{x(t)\}$ generated by a Gibbs Sampler targeting a Gaussian multivariate distribution $N(\mu, \Sigma)$ is a Gaussian autoregressive process evolving as $x(t+1) \mid x(t) \sim N(Bx(t) + b, \Sigma - B\Sigma B^{\mathrm{T}})$, see for example Lemma 1 in Roberts & Sahu (1997). The details of the Gibbs sampler, such as the order that its components are updated or blocked together, are reflected in the precise form of B. This representation implies that the rate of convergence of the Gibbs sampler is $\rho(B)$, the largest absolute eigenvalue of the matrix B, see Theorem 1 of Roberts & Sahu (1997). This characterisation of the $L^2(\pi)$ rate of convergence has provided useful insights into the performance of the Gibbs sampler and has lead to much more efficient modifications of the basic algorithm, see for example Papaspiliopoulos et al. (2003, 2007). However, in high-dimensional scenarios it is often very challenging to compute $\rho(B)$ explicitly as a function of the important parameters of the model, such as p and N in the crossed effects models considered here. Hence as a tool for understanding the complexity of the Gibbs sampler in difficult problems this approach has limited scope. In this article we will make it useful by combining it with the multigrid decomposition developed below, which collapses the problem to studying the spectral gaps of two Gaussian subchains, $\{\bar{a}(t)\}$ and $\{\delta a(t)\}$, that turn out to be amenable to direct analysis.

5. Complexity analysis for crossed random effect models

5·1. *Multigrid decomposition of the Gibbs samplers*

The results we derive in this paper stem from the following result, the proof of which is given in the Appendix.

THEOREM 1. *(Multigrid decomposition) Let $\{a(t)\}$ be the Markov chain generated either by the Gibbs sampler or the collapsed Gibbs sampler for balanced levels designs. Then, the timewise transformations $\{\bar{a}(t)\}$ and $\{\delta a(t)\}$ obtained from $\{a(t)\}$ are each a Markov chain and they are independent of each other.*

A crucial point here is that the independence of $\bar{a}$ and δa under the posterior distribution, see Proposition 1, does not imply that the corresponding chains $\{\bar{a}(t)\}$ and $\{\delta a(t)\}$ are independent of each other. The following very simple example makes this point clear. Consider a Gibbs sampler that targets a bivariate Gaussian for (x, y) with correlation ρ and standard Gaussian marginals. Then the transformation x and $z = y - \rho x$ orthogonalises the target, but the corresponding stochastic processes $\{x(t)\}$ and $\{z(t)\}$ obtained by timewise transformation of the original chain $\{(x, y)(t)\}$ are not independent Markov chains, see, e.g., the cross-correlogram in Figure 1 of the supplementary material. Although this is a toy example, there are many instances where an independence factorisation of the target distribution does not require that of the MCMC algorithm adopted. In the following sections we use Theorem 1 in conjunction with the theory from Section 4·3 to characterise the complexity of the two samplers.

5·2. *Timewise transformations and convergence of Markov chains*

The multigrid decomposition in Theorem 1 identifies two timewise transformations of the Markov chain $\{a(t)\}$ produced by either of the algorithms considered in this article, each of which evolves independently of each other as a Markov chain. We can relate the rate of convergence of the Markov chains involved in this decomposition using the following two technical lemmata that are proved in the supplementary material.

LEMMA 1. *Let $\{x(t)\}$ be a Markov chain with invariant distribution π and $\{y(t)\}$ be a timewise transformation given by $y(t) = \phi(x(t))$, where ϕ is an injective function. Then $\{y(t)\}$ is a Markov chain with the same rate of convergence as $\{x(t)\}$.*

Lemma 1 is similar in spirit to Theorem 6 of Johnson & Geyer (2012), where it is shown that various properties of Markov chains are preserved under one-to-one timewise transformations.

LEMMA 2. *Let $\{x(t)\}$ be a Markov chain with state space $\mathcal{X}_1 \times \mathcal{X}_2$ and target distribution $\pi_1 \otimes \pi_2$. If the stochastic processes $\{x_1(t)\}$ and $\{x_2(t)\}$ obtained by projection on the $\mathcal{X}_1$ and $\mathcal{X}_2$ components are two independent Markov chains, then the rate of convergence of $\{x(t)\}$ equals the supremum between the rates of convergence of $\{x_1(t)\}$ and $\{x_2(t)\}$.*

Therefore, for balanced levels designs the rate of convergence of the Markov chain $\{a(t)\}$, generated either by the Gibbs sampler or the collapsed Gibbs sampler, is the larger of the rates of the two chains $\{\bar{a}(t)\}$ and $\{\delta a(t)\}$. In the remainder of Section 5 we analyse these two chains using the theory summarised in Section 4·3 and use the results to characterize the complexity of the Gibbs sampler and its collapsed version.

5·3. *Complexity analysis for balanced cells designs*

The following result characterises the rate of convergence of one of the two timewise transformations involved in the multigrid decomposition.

PROPOSITION 3. *For balanced levels designs the rate of convergence of the Markov chain $\{\bar{a}(t)\}$ defined in Theorem 1 equals $\max_k \frac{N\tau_0}{N\tau_0 + I_k\tau_k}$ for the Gibbs Sampler and 0 for the collapsed Gibbs Sampler, and this rate is the same for any order that the different blocks are updated.*

Proof. For the Gibbs Sampler, the subchain $\{\bar{a}(t)\}$ is a Gaussian Gibbs Sampler, with $(K+1)$ one-dimensional components. We can explicitly work out that its autoregressive matrix B takes the form

$$B = \left(\begin{array}{c|c} 0 & -1 \ \dots \ -1 \\ \hline 0 & \\ \vdots & \quad L \\ 0 & \end{array}\right) \tag{9}$$

where L is a $K \times K$ lower triangular matrix with diagonal elements equal to $(r_1, \dots, r_K)$, with

$$r_k = \frac{N\tau_0}{N\tau_0 + I_k\tau_k}\,. \tag{10}$$

The correctness of (9) is shown in the supplementary material verifying directly that $\mathbb{E}\{\bar{a}(t+1) \mid \bar{a}(t)\} = B\bar{a}(t) + b$ with an induction argument.

Since L is a lower triangular matrix its spectrum coincides with its diagonal elements $(r_1, \dots, r_K)$. For each $k = 1, \dots, K$, let $v^{(k)}$ be the eigenvector with eigenvalue r_k. It is easy to

check that the $(K+1)$-dimensional vector $w^{(k)} = (-r_k^{-1} \sum_{\ell=1}^{K} v_\ell^{(k)}, v_1^{(k)}, \ldots, v_K^{(k)})$ is an eigenvector of B with eigenvalue r_k. Thus $(r_1, \ldots, r_k)$ are also eigenvalues of B. Finally note that $(1, 0, \ldots, 0)$ is an eigenvector of B with eigenvalue 0. With these ingredients the proof of the claim for the Gibbs sampler follows immediately.

For the collapsed Gibbs Sampler, $\bar{a}(t)$ is obtained from $\bar{a}(t-1)$ by simulating $\bar{a}^{(k)}(t)$ from

$$\mathcal{L}\left\{\bar{a}^{(k)}(t) \mid y, \bar{a}^{(1)}(t), \ldots, \bar{a}^{(k-1)}(t), \bar{a}^{(k+1)}(t-1), \ldots, \bar{a}^{(K)}(t-1)\right\},$$

for $k = 1, \ldots, K$. By Proposition 1, this procedure produces independent and identically distributed draws from $\mathcal{L}\left\{\bar{a}^{(-0)} \mid y\right\}$, or equivalently $\mathcal{L}\left\{\bar{a} \mid y\right\}$ if $\bar{a}^{(0)}$ is jointly updated with $\bar{a}^{(k)}$.

These rates do not depend on the order that the different components are updated. This is trivially true for the collapsed Gibbs since the components are independent. For the Gibbs sampler the argument is as follows. The Gibbs Sampler rate of convergence is invariant with respect to cyclic permutations of the order of update of the components, see e.g. Roberts & Sahu (1997, p.297). Thus we can always assume $a^{(0)}$ to be the first component to be updated. Then the result follows by relabeling the components $a^{(1)}$ to $a^{(K)}$ according to their update order and replicating the argument developed in the previous paragraphs. □

The main result of this section follows rather easily from Proposition 3.

THEOREM 2. *For balanced cells designs, the relaxation time of the Gibbs Sampler is* $1 + \max_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$*, and that of the collapsed Gibbs Sampler is 1, i.e. the sampler produces independent and identically distributed draws from the target, and these rates do not depend on the order that different components are updated.*

Proof of Theorem 2. Let $\{a(t)\}$ be the Markov chain generated by the Gibbs Sampler or its collapsed version. Lemma 1 implies that $\{(\bar{a}, \delta a)(t)\}$ is a Markov chain with the same rate of convergence as $\{a(t)\}$. Thus, by means of Theorem 1 and Lemma 2, the rate of convergence of $\{a(t)\}$ equals the maximum between the rate of convergence of $\{\bar{a}(t)\}$ and the one of $\{\delta a(t)\}$. Proposition 1 implies that $\{\delta a(t)\}$ performs independent sampling from $\mathcal{L}\left\{\delta a \mid y\right\}$ and thus its rate of convergence is 0 and the rate of convergence of $\{a(t)\}$ equals the one of $\{\bar{a}(t)\}$. To conclude, Proposition 3 and the definition of relaxation times as reciprocal of the spectral gap imply the statement to be proved. □

The theorem completely characterises the relaxation time of the Gibbs sampler and the collapsed Gibbs sampler for balanced cells designs. Considering the computational cost of the algorithms, we find that each of the algorithms requires an $\mathcal{O}(N)$ computation at initialisation to precompute data averages. In the Appendix we show that both algorithms have the same cost per iteration, which is proportional to the number of parameters, p. Therefore, the collapsed Gibbs sampler is an $\mathcal{O}(p)$ implementation of exact sampling from the posterior.

We now consider asymptotic regimes. The more classical asymptotic regime, which we will refer to as infill asymptotics, keeps the number of factors and levels fixed, hence K and p fixed, and increases the number of observations per cell, hence N grows. The other more modern asymptotic regime, which we will refer to as outfill asymptotics, increases p with N, e.g. considering the observations per cell bounded and increasing the number of levels and/or factors. It is this type of asymptotic that it is more interesting in recommendation applications.

Regardless of the asymptotic regime considered the relaxation time of the collapsed Gibbs sampler is $\mathcal{O}(1)$. On the other hand, that of the Gibbs sampler depends on the regime considered. In infill asymptotics Theorem 2 implies that the relaxation time of the algorithm is $\mathcal{O}(N)$. An intuition for this deterioration of the algorithm with increasing data size can be obtained by

considering the analysis of non-centered parameterisations for hierarchical models in Section 2 of Papaspiliopoulos et al. (2007); the parameterisation of the crossed effect model is non-centred and the infill asymptotics regime makes the data increasingly informative per random effect, hence we should anticipate the deterioration. Therefore, in this regime the complexity of both algorithms is $\mathcal{O}(N)$ but in practice the collapsed will be much more efficient. In outfill asymptotics, both N and the number of factor levels I_k's are growing, hence by Theorem 2 the relaxation time of the Gibbs sampler is no worse than $\mathcal{O}(N)$ but no better than $\mathcal{O}(N^{1-1/K})$. The lower bound on the relaxation time can be deduced from the balanced cells design assumption, which implies $\prod_{k=1}^{K} I_k \leq N$ and $\min_k I_k \leq N^{1/K}$; the bound is achievable when $I_1 = \cdots = I_K$. On the other hand, the number of parameters can grow as different powers of N. For example, if the number of levels for all but one factor are fixed and those of the remaining factor are increasing, e.g. fixed number of customers and increasing number of products, then p is $\mathcal{O}(N)$ and the relaxation time of the Gibbs sampler is also $\mathcal{O}(N)$, resulting in a Gibbs Sampler complexity of $\mathcal{O}(N^2)$, whereas the collapsed Gibbs sampler is $\mathcal{O}(N)$.

5·4. *Complexity analysis for balanced levels designs*

The strategy for obtaining complexity results for balanced levels designs is the same as for balanced cells and Proposition 3 is again instrumental. However, in this case the analysis is much more complicated since the second timewise transformation, $\{\delta a(t)\}$, no longer samples independently from its invariant distribution; in fact its invariant distribution does not factorise as in the case of balanced cells. Nonetheless, Lemma 2 and Proposition 3 imply immediately the following lower bound on the relaxation time of the Gibbs sampler.

THEOREM 3. *For balanced levels designs, the relaxation time of the Gibbs Sampler is at least* $1 + \max_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$.

From Proposition 3 we also know that the rate of the collapsed Gibbs sampler is that of $\{\delta a(t)\}$. Therefore, obtaining explicit rates of convergence for $\{\delta a(t)\}$ is the step needed for characterising the relaxation time of both algorithms in balanced levels designs. We are able to do this for $K = 2$ in Proposition 4 below. Our theory is based on an auxiliary process $\{i(t)\}$ with discrete state space $\{1, \ldots, I_1\} \times \{1, \ldots, I_2\}$ that evolves according to a two component Gibbs Sampler, iteratively updating $i_1 \mid i_2$ and $i_2 \mid i_1$, with invariant distribution $p(i_1, i_2) = n_{i_1 i_2}/N$.

PROPOSITION 4. *For balanced levels designs with $K = 2$, the rate of convergence of the Markov chain $\{\delta a(t)\}$ is*

$$\frac{N\tau_0}{N\tau_0 + I_1\tau_1} \frac{N\tau_0}{N\tau_0 + I_2\tau_2} \rho_{aux} \,,$$

where ρ_{aux} is the rate of convergence of the auxiliary Gibbs sampler $\{i(t)\}$.

Proof. The chain $\{\delta a(t)\}$ is a two-component Gibbs Sampler that alternates updates from $\mathcal{L}\left\{\delta a^{(1)} \mid y, \delta a^{(2)}\right\}$ and $\mathcal{L}\left\{\delta a^{(2)} \mid y, \delta a^{(1)}\right\}$. Thus, $\{\delta a^{(1)}(t)\}$ is marginally a Markov chain and its rate of convergence equals the one of $\{\delta a(t)\}$, see e.g. Roberts & Rosenthal (2001). Let B_1 and B_2 defined by $\mathbb{E}[\delta a^{(1)} \mid \delta a^{(2)}, y] = B_1 \delta a^{(2)} + b_1$ and $\mathbb{E}[\delta a^{(2)} \mid \delta a^{(1)}, y] = B_2 \delta a^{(1)} + b_2$. It is then a simple computation that $\delta a^{(1)}(t)$ is a Gaussian autoregressive process with autoregression matrix $B_1 B_2$. Since for balanced levels design it holds $\frac{n_j^{(k)}\tau_0}{n_j^{(k)}\tau_0 + \tau_k} = r_k$, it can be deduced from (4) and (7) that $B_1 = -r_1 P_1$, where P_1 is a $I_1 \times I_2$ matrix being the transition kernel of the update $i_2|i_1$ of the auxiliary process. Similarly, one can show $B_2 = -r_2 P_2$, where P_2 is a $I_2 \times I_1$ matrix being the transition kernel of the update $i_1|i_2$ of the auxiliary process. Hence, the

autoregressive matrix of $\delta a^{(1)}(t)$ is $r_1 r_2 P_1 P_2$, where $P_1 P_2$ is the transition kernel of the auxiliary Gibbs sampler $\{i(t)\}$. Consequently, the spectrum of the autoregressive matrix is $r_1 r_2 \lambda_i$, where λ_i are the eigenvalues of $P_1 P_2$. The largest $|\lambda_i|$ is of course 1 since $P_1 P_2$ is a stochastic matrix. However, since $\delta a^{(1)}$ is constrained to have zero sum, by Lemma 3 in the Appendix the rate of convergence of $\delta a^{(1)}(t)$ is not given by the largest modulus eigenvalue of the autoregressive matrix, but the largest modulus eigenvalue whose eigenvector has zero sum, i.e., we need to consider only the subspace orthogonal to the vector of 1's. Therefore, the rate of convergence of $\{\delta a(t)\}$ equals $r_1 r_2$ times the second largest modulus eigenvalue of $P_1 P_2$, which is ρ_{aux} by definition. □

With Proposition 4 in place, the main result of this Section on the relaxation time of the algorithms follows immediately.

THEOREM 4. *For balanced levels designs with $K = 2$, the rate of convergence of the Gibbs Sampler and the collapsed Gibbs Sampler are given by, respectively,*

$$\max\left\{\frac{N\tau_0}{N\tau_0 + I_1\tau_1}, \frac{N\tau_0}{N\tau_0 + I_2\tau_2}\right\}, \qquad \frac{N\tau_0}{N\tau_0 + I_1\tau_1}\frac{N\tau_0}{N\tau_0 + I_2\tau_2}\rho_{aux},$$

where ρ_{aux} is the rate of convergence of the auxiliary Gibbs sampler $\{i(t)\}$ with invariant distribution $p(i_1, i_2) = n_{i_1 i_2}/N$.

Note that if the design is in fact balanced cells, the rates given in Theorem 4 match those of Theorem 3, as they should, since $\rho_{aux} = 0$ in this case.

A corollary to this Theorem is that the relaxation time of the Gibbs sampler is $1 + \max_{k=1,2} \frac{N\tau_0}{I_k\tau_k}$ and that of the collapsed Gibbs sampler is no larger than $1 + \min\{\frac{N\tau_0}{I_1\tau_1}, \frac{N\tau_0}{I_2\tau_2}, T_{aux}\}$, where T_{aux} is the relaxation time of the auxiliary process $\{i(t)\}$. An implication of this is that the collapsed Gibbs Sampler is never slower than the standard Gibbs Sampler and it is has good mixing both when the amount of data per level is low and high. To see this, note first the ratios N/I_1 and N/I_2 coincide with the number of datapoints per column and row, respectively, in the data incidence matrix with entries $n_{i_1 i_2}$ and thus their value increases as the amount of data per level increases. On the contrary the relaxation time T_{aux} of the auxiliary process $\{i(t)\}$ tends to decrease as the amount of data per level increases because the latter corresponds to adding more edges in the conditional independence graph, hence larger connectivity in the state space of the auxiliary process. Unfortunately, it is not true in general that the minimum across $\frac{N\tau_0}{I_1\tau_1}$, $\frac{N\tau_0}{I_2\tau_2}$ and T_{aux} is uniformly bounded over N. Consider for example a design where users and items are split into two communities of equal size, and users inside each community have rated all items from their community and no item from the other community. In this case the random walk $\{i(t)\}$ is reducible. Therefore $T_{aux} = \infty$ and, provided both N/I_1 and N/I_2 go to infinity, the relaxation time of the collapsed Gibbs Sampler diverges as N goes to infinity.

We now address the case of number of factors $K > 2$ that Theorem 4 does not cover. A conjecture we make in this paper is that $1 + \max_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$ is the relaxation time of the Gibbs sampler also for $K > 2$. We have experimented numerically quite extensively, since for specific examples we can compute the relaxation time by computing numerically the largest eigenvalue of an explicit matrix, and we have not been able to find a counter-example. The missing step for a generic result would be to show that $\{\delta a(t)\}$ always mixes faster than $\{\bar{a}(t)\}$. Such a result would also immediately prove, due to Proposition 3, that the collapsed Gibbs sampler has lower relaxation time than the Gibbs sampler for arbitrary number of factors for balanced levels designs. On the other hand, numerical experimentation has also showed that certain extensions

of Theorem 4 are not true. We know that the convergence rate of the collapsed Gibbs sampler can be larger than $\prod_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$ when $K > 2$; we also know that the rate will depend on the order that the different components are updated. We return to these points in the Discussion.

We close the section with some asymptotic considerations on the complexity. The following arguments assume that the relaxation time of the Gibbs sampler is the conjectured $1 + \max_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$; we will not consider the collapsed Gibbs sampler in the following considerations since we do not have conjecture for its rate when $K > 2$. The asymptotic behaviour of the Gibbs sampler relaxation time depends on the regime under consideration as it was for balanced cells designs. The relaxation time can be as bad as $\mathcal{O}(N)$, for example if the number of levels of at least one factor is fixed as N grows; it can be $\mathcal{O}(N^{1-1/K})$ in the regime where $I_1 = \ldots = I_K$ and $N = \mathcal{O}(I_1^K)$; but it can also be $\mathcal{O}(1)$ in the sparse observation regime where $N = I_1 = \ldots = I_2$. The Appendix discuss the computational cost per iteration, which for these designs can grow quadratically with the number of parameters, as opposed to linearly in the case of balanced cells. In terms of its growth with the observations, this can be $\mathcal{O}(1)$, in infill asymptotics regimes where the number of levels of factors does not grow with N; it can be $\mathcal{O}(N^{2/K})$ when $I_1 = \ldots = I_K$ and $N = \mathcal{O}(I_1^K)$; but it can also be $\mathcal{O}(N)$ in the sparse regime $N = I_1 = \ldots = I_2$. Connecting now to the observation in Gao & Owen (2017), we obtain that for $K = 2$ when $N = \mathcal{O}(I_1^2)$ and $I_1 = I_2$, the complexity of the Gibbs sampler is $\mathcal{O}(N^{3/2})$, hence the algorithm is not scalable.

6. SIMULATION STUDIES

6·1. *Simulated data with missingness completely at random*

First we consider simulated data with $K = 2$ and $I_1 = I_2$. We assume data to be missing completely at random, where for each combination of factors we observe a datapoint, i.e., $n_{i_1 i_2} = 1$, with probability 0.1 independently of the rest, and otherwise we have a missing observation, i.e., $n_{i_1 i_2} = 0$. Since the relaxation time of the samplers under consideration does not depend on the the value of the observations y, but only on their presence or absence, we can set $y_{i_1 i_2} = 0$ without affecting the computed convergence rates. In this context our theory does not apply directly because the designs under consideration are not balanced in general. However, we can still compute numerically the convergence rate of the Gibbs Sampler and its collapsed version in the context of known precisions, using the results discussed in Section 4·3, to explore to which extent the qualitative findings of our theory still apply. Figure 1(a) displays the behaviour of the relaxation time of the Gibbs Sampler and its collapsed version in an outfill asymptotic regime, where both the number of datapoints and factor levels increase. For the simulations we fixed the precision terms τ_k to 1 and take I_1 in the set $\{50, 500, 1000, 2000\}$. The results suggest that the relaxation time of the Gibbs Sampler diverges with N, while the relaxation time of its collapsed version converges to 1 as N increases. This is consistent with the theoretical results of previous section. In fact, we can compare the relaxation times that we computed numerically with the theoretical values computed as if the design were balanced levels, which of course it is not here. The figure shows an extremely close match, which showcases the use of our theory beyond the specific designs that have facilitated the analysis. This suggests that the theory previously developed is relevant beyond cases that strictly satisfy balanced levels. Since the cost per iteration of both samplers is $\mathcal{O}(N)$, the results in Figure 1(a) suggest that, for the asymptotic regime considered in this section, the computational complexity of the Gibbs Sampler is $\mathcal{O}(N^{3/2})$ and the one of the collapsed Gibbs Sampler is $\mathcal{O}(N)$.

6·2. *ETH Instructor Evaluations dataset*

We now consider a real dataset containing university lecture evaluations by students at ETH Zurich. The dataset is freely available from the R package *lme4* (Bates et al., 2015) under the name *InstEval*. It contains 73421 observations, each corresponding to a score ranging from 1 to 5, assigned to a lecture together with 6 factors potentially impacting such score, such as identity of the student giving the rating or department that offers the course. See the *lme4* help material for more details on the dataset. We fit model (1) to the *InstEval* dataset. Following the notation in (1), we have $N = 73421$, $K = 6$ and $(I_1, \ldots, I_K) = (2972, 1128, 4, 6, 2, 14)$. Clearly, a categorical response calls for a generalised linear model extension of (1), however the point of this analysis is to test the algorithms, and (1) is not an outright unreasonable model to fit for this dataset.

First we consider the case where the values of the precisions $\{\tau_k\}$ are assumed to be known. As in Section 6·1, we compute the true relaxation times numerically, and compare them to the theoretical predictions implied by Theorem 4. The resulting values for various combination of factors are reported in the supplementary Material. The relaxation time of the collapsed Gibbs sampler is up to three orders of magnitude smaller than the one of the Gibbs sampler. In all cases considered, the theoretical predictions matched closely the actual values computed numerically.

Next consider the case of unknown precisions, where the hyperparameters τ_k are given a prior distribution and the posterior of interest is the joint distribution of a and $\tau = (\tau_0, \ldots, \tau_K)$. We consider two popular prior specifications for τ, the first being a flat prior $p(\tau_k^{-1/2}) \propto 1$ and the second a half-Cauchy prior $\tau_k^{-1/2} \sim \text{Cauchy}^+(0, 1)$, see Gelman (2006) and Polson et al. (2012) for a discussion. Under both prior specifications, we consider five Markov chain Monte Carlo schemes. The first two schemes alternate sampling τ from the conditional distribution $\mathcal{L}\{\tau \mid a\}$ and updating a with the Gibbs Sampler and its collapsed version, respectively. In the flat prior case, the exact update $\tau \sim \mathcal{L}\{\tau \mid a\}$ is straightforward, while in the half-Cauchy case the latter is replaced by a Metropolis-Hastings update. The third and fourth schemes add parameter expansion (Liu & Wu, 1999; Meng & Van Dyk, 1999). See the supplementary material for full details on the implementation of these first four schemes. Finally, the fifth scheme is the No U-Turn sampler (Hoffman & Gelman, 2014), a state-of-the-art Hamiltonian Monte Carlo scheme implemented in the *R* package *RStan* (Stan Development Team, 2018). In order to avoid potential issues related to using flat priors with a very low number of factor levels, we excluded the factor with only two levels from the analysis, resulting in $K = 5$ and $(I_1, \ldots, I_K) = (2972, 1128, 4, 6, 14)$.

Table 1 reports runtimes for the five schemes together with effective sample sizes. It can be seen that the first four schemes have similar runtimes, but the ones using the collapsed methodology proposed in this paper induce a much faster mixing compared to the others. In this example the use of parameter expansion has little effect on mixing, giving some improvement in the flat prior case and some more deterioration in the half-Cauchy case. Hamiltonian Monte Carlo has a cost per iteration that is two orders of magnitude larger than the other schemes, resulting in the lowest effective sample sizes per unit of computation time. The supplementary material contains figures displaying autocorrelation functions versus CPU time for the Gibbs samplers, again showcasing the much improved mixing of the collapsed one.

Finally, to obtain a higher level sense of the practicality of the approach we pursue in this article, we also fit the same crossed effect model in a frequentist fashion using the *R* package *lme4*, which took 40.9 seconds to run. All computations were performed on the same desktop computer with 16GB of RAM and an $i7$ Intel processor. It is worth noting that the first four schemes were directly implemented using a high level language such as R, so we would expect

Scheme	time per 1000 iter.	Effective Sample Size / time (1/s) $(a^{(0)}, \bar{a}^{(1)}, \bar{a}^{(2)}, \bar{a}^{(3)}, \bar{a}^{(4)}, \bar{a}^{(5)})$	$(\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$
GS	13.2s	(0.07, 11.0, 2.12, 0.16, 0.21, 0.87)	(60.9, 15.9, 36.8, 3.56, 2.53, 2.14)
GS+PX	13.5s	(0.06, 10.7, 2.02, 0.08, 0.11, 0.95)	(59.6, 41.2, 43.9, 0.85, 0.58, 2.33)
HMC	1112.6s	(0.11, 0.78, 0.19, 0.08, 0.25, 0.68)	(0.99, 0.51, 0.99, 0.10, 0.41, 0.18)
cGS	14.2s	(65.9, 42.1, 18.1, 70.5, 62.3, 35.0)	(55.1, 14.7, 34.5, 17.7, 33.9, 2.51)
cGS+PX	14.4s	(62.5, 44.1, 19.9, 62.9, 63.2, 34.6)	(55.1, 38.0, 41.9, 19.2, 33.0, 2.96)
GS	15.0s	(0.07, 9.87, 1.86, 0.16, 0.23, 0.85)	(51.6, 14.1, 31.5, 1.33, 3.84, 2.19)
GS+PX	13.8s	(0.07, 10.9, 1.85, 0.07, 0.18, 0.84)	(56.9, 39.9, 43.6, 0.80, 0.70, 1.71)
HMC	1186.9s	(0.08, 0.30, 0.08, 0.10, 0.08, 0.24)	(0.79, 0.27, 0.55, 0.18, 0.14, 0.09)
cGS	15.5s	(74.0, 39.0, 17.7, 77.6, 57.2, 31.6)	(51.8, 13.7, 30.1, 4.84, 15.0, 2.40)
cGS+PX	14.6s	(59.1, 43.7, 18.8, 63.2, 61.6, 33.2)	(54.7, 38.0, 40.4, 0.72, 0.92, 1.88)

Table 1. *Comparison of sampling schemes on the* InstEval *data. The first five lines refer to flat priors for $\sigma_k = 1/\sqrt{\tau_k}$, the second five lines to half-Cauchy ones. GS and cGS refer to the Gibbs Sampler and the collapsed version with precision updates, while +PX indicates combination with the parameter expanded methodology. HMC referes to the* RStan *implementation of the No U-Turn sampler. Numbers are averaged over 10 runs of 10000 iterations for each scheme, discarding the first 1000 samples as burn-in.*

significant further speed-ups by using a low-level language and use of distributed computing for the precomputations needed for the Gibbs samplers.

7. DISCUSSION

There are many directions this work can move forward. First we highlight the two that are most imminent. One is to investigate the conjecture made in Section 5·4 that the relaxation time of the Gibbs sampler for balanced levels designs is $1 + \max_{k=1,\ldots,K} \frac{N\tau_0}{I_k\tau_k}$. If this is true we also obtain that the collapsed Gibbs sampler has always smaller rate for such designs. The other is to obtain a characterisation of the rate of the collapsed Gibbs sampler for such designs when $K > 2$. From numerical experimentation we know that the natural extension of the expression of Theorem 4 is not true for $K > 2$, hence a different line of attack is needed.

Finally, we mention two important possible directions of future research that would help providing a clearer picture about the scalablity of likelihood-based inferences for crossed effect models. The first is to provide theoretical understanding regarding the extent to which sparse linear algebra methods can reduce the computational complexity of the matrix operations involved in the exact marginalizations of the space of factors. The second is to develop rigorous complexity results for the case of fully Bayesian inferences with unknown variances.

ACKNOWLEDGEMENT

The authors would like to acknowledge helpful discussions with Art B. Owen. Papaspiliopoulos was supported by the Spanish Ministry of Economics via a research grant. Roberts was supported by the UK Engineering and Physical Sciences Research Council. Zanella was supported by the European Research Council and by the Ministry of Education, Universities and Research.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Lemmas 1, 2 and 3, additional figures and details on the simulation study in Section 6·2 and R code to implement the MCMC schemes under consideration.

APPENDIX

Proof of Theorem 1. For concreteness and without affecting the validity of the argument we assume that the algorithm updates factors and their levels in ascending order, i.e., first simulates $a^{(0)}$, then $a_1^{(1)}$, $a_2^{(1)}$, and so on and so forth. We first establish the result for the Gibbs sampler,i.e., part 1. Note that due to the conditional independence structure the algorithm can be equivalently represented as one that samples in blocks according to the conditional laws $\mathcal{L}\left\{a^{(k)} \mid y, a^{(-k)}\right\}$. For each iteration t, each such draw, $a^{(k)}(t)$ can be transformed to $\bar{a}^{(k)}(t)$ and $\delta a^{(k)}(t)$. Proposition 1 establishes that the

$$\begin{aligned}
&\mathcal{L}\left\{\bar{a}^{(k)}(t), \delta a^{(k)}(t) \mid y, a^{(0)}(t), \ldots, a^{(k-1)}(t), a^{(k+1)}(t-1), \ldots, a^{(K)}(t-1)\right\} = \\
&\mathcal{L}\left\{\bar{a}^{(k)}(t) \mid y, \bar{a}^{(0)}(t), \ldots, \bar{a}^{(k-1)}(t), \bar{a}^{(k+1)}(t-1), \ldots, \bar{a}^{(K)}(t-1)\right\} \times \qquad (11) \\
&\mathcal{L}\left\{\delta a^{(k)}(t) \mid y, \delta a^{(1)}(t), \ldots, \delta a^{(k-1)}(t), \delta a^{(k+1)}(t-1), \ldots, \delta a^{(K)}(t-1)\right\} .
\end{aligned}$$

Appealing to the equivalence of local and global Markov properties, as in Section 3 of Besag (1974), we obtain that the processes $\{\bar{a}(t)\}$ and $\{\delta a(t)\}$, obtained as functions of $\{a(t)\}$, are each a Markov chain with respect to its own filtration, and independent of each other.

The collapsed Gibbs Sampler case is analogous. Here the sampler iterates the updates of $\mathcal{L}\left\{a^{(k)} \mid y, a^{(-0,-k)}\right\}$ for $k = 1, \ldots, K$. It can be easily deduced from Proposition 1 that $\mathcal{L}\left\{\bar{a}^{(-0)}, \delta a \mid y\right\} = \mathcal{L}\left\{\bar{a}^{(-0)} \mid y\right\} \mathcal{L}\left\{\delta a \mid y\right\}$. Therefore, transforming each draw $a^{(k)}(t)$ to $\bar{a}^{(k)}(t)$ and $\delta a^{(k)}(t)$, we obtain

$$\begin{aligned}
&\mathcal{L}\left\{\bar{a}^{(k)}(t), \delta a^{(k)}(t) \mid y, a^{(1)}(t), \ldots, a^{(k-1)}(t), a^{(k+1)}(t-1), \ldots, a^{(K)}(t-1)\right\} = \\
&\mathcal{L}\left\{\bar{a}^{(k)}(t) \mid y, \bar{a}^{(1)}(t), \ldots, \bar{a}^{(k-1)}(t), \bar{a}^{(k+1)}(t-1), \ldots, \bar{a}^{(K)}(t-1)\right\} \times \\
&\mathcal{L}\left\{\delta a^{(k)}(t) \mid y, \delta a^{(1)}(t), \ldots, \delta a^{(k-1)}(t), \delta a^{(k+1)}(t-1), \ldots, \delta a^{(K)}(t-1)\right\} .
\end{aligned}$$

It follows that the processes $\{\bar{a}^{(-0)}(t)\}$ and $\{\delta a(t)\}$, obtained as functions of $\{a(t)\}$, are each a Markov chain with respect to its own filtration, and independent of each other. □

LEMMA 3. *Let $\{x(t)\}$ be a d-dimensional gaussian AR(1) process with $\mathbb{E}[x(t+1) \mid x(t)] = Bx(t) + b$, for some fixed b, and stationary distribution $N(\mu, \Sigma)$ concentrated on the hyperplane $\sum_i x_i = 0$ and Σ of rank $d-1$. Then the rate of convergence of $\{x(t)\}$ equals the largest modulus eigenvalue of B whose eigenvector has zero sum.*

Cost per iteration of the Gibbs Sampler and its collapsed version

In order to implement the Gibbs Sampler, the computation of the one and two-dimensional marginals $\{n^{(k)}\}$ and $\{n^{(l,k)}\}$ of the data incidence table are required, as well as the computation of the weighted averages $\{\tilde{y}_j^{(k)}\}$ of the data. Such precomputation needs to be performed only once and requires $\mathcal{O}(N)$ operations in general. Then, at each iteration of the Gibbs Sampler the update of $a^{(0)}$ and each $a_j^{(k)}$ can be accomplished in $\mathcal{O}(\sum_l I_l)$ and $\mathcal{O}(\sum_{l \neq k} I_l)$ operations using

(2) and (4), respectively, resulting in a total of $\mathcal{O}(\sum_k I_k \sum_{l\neq k} I_l)$ operations for each Gibbs sweep. The latter can be as bad as $\mathcal{O}(p^2)$, where $p = \sum_k I_k$ is the number of parameters and its relationship with N depends on the asymptotic regime under consideration.

For the collapsed Gibbs Sampler one needs to additionaly precompute $\{s^{(k)}\}$ defined in Proposition 2, which can be done in $\mathcal{O}(p)$ operations given $\{n^{(k)}\}$. Therefore the collapsed Gibbs Sampler has a precomputation cost of order $\mathcal{O}(N)$, similarly to the standard Gibbs Sampler. Moreover, the updates of $a^{(0)}$ from (8) for $k = 1, \ldots, K$ require $\mathcal{O}(\sum_k I_k \sum_{l\neq k} I_l)$ operations altogether, which is at most $\mathcal{O}(p^2)$. Thus the collapsed Gibbs Sampler has also the same cost per iteration of the standard Gibbs Sampler.

In the balanced cells case, the only precomputation required is the one of $\{\tilde{y}_j^{(k)}\}$, which has $\mathcal{O}(N)$ cost. Also, each Gibbs or collapsed Gibbs sweep can be accomplished in $\mathcal{O}(p)$ operations, rather than $\mathcal{O}(p^2)$, using (3), (5) and the version of (8) for balanced cells.

REFERENCES

BATES, D., MÄCHLER, M., BOLKER, B. & WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36**, 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

GAO, K. & OWEN, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics* **11**, 1235–1296.

GAO, K. & OWEN, A. (2019). Estimation and inference for very large linear mixed effects models. *Statistical Sinica, In press. arXiv preprint arXiv:1610.08088* .

GELMAN, A. (2005). Analysis of variance—why it is more important than ever. *Ann. Statist.* **33**, 1–53. With discussions and a rejoinder by the author.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

HOFFMAN, M. D. & GELMAN, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623.

JOHNSON, L. T. & GEYER, C. J. (2012). Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics* **40**, 3050–3076.

LIU, J. S. & WU, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.

MENG, X.-L. & VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301–320.

PAPASPILIOPOULOS, O., ROBERTS, G. O. & SKÖLD, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*. Oxford Univ. Press, New York, pp. 307–326. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

PAPASPILIOPOULOS, O., ROBERTS, G. O. & SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science* , 59–73.

POLSON, N. G., SCOTT, J. G. et al. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis* **7**, 887–902.

ROBERTS, G. O. & ROSENTHAL, J. S. (2001). Markov Chains and De-initializing Processes. *Scandinavian Journal of Statistics* **28**, 489–504.

ROBERTS, G. O. & SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 291–317.

STAN DEVELOPMENT TEAM (2018). RStan: the R interface to Stan. R package version 2.17.3.

VOLFOVSKY, A. & HOFF, P. D. (2014). Hierarchical array priors for ANOVA decompositions of cross-classified data. *Ann. Appl. Stat.* **8**, 19–47.

ZANELLA, G. & ROBERTS, G. (2017). Analysis of the gibbs sampler for gaussian hierarchical models via multigrid decomposition. *arXiv preprint arXiv:1703.06098* .

[*Received December* 2015. *Revised December* 2014]