# SCALABLE INFERENCE FOR CROSSED RANDOM EFFECTS MODELS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Swarnadip Ghosh
May 2022

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Art B. Owen) Principal Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Trevor J. Hastie) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Susan Athey)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Scott W. Linderman)

Approved for the Stanford University Committee on Graduate Studies

# Abstract

Crossed random effects models are widely used in statistics for modeling data with a crossed structure and have found various applications such as recommender systems in electronic commerce. Ignoring the vast tangle of correlation can lead to an analyst being overconfident about the results. The computational costs of estimation and inference of the crossed random effect model limit its use. Scalable inference, which means statistical inference methods with computational cost at most linear in the number of observations and the number of parameters, has been a significant challenge for such models.

Regression models with crossed random effect errors can be very expensive to compute. The cost of both generalized least squares and Gibbs sampling can easily grow as $N^{3/2}$ (or worse) for $N$ observations. We propose a backfitting algorithm to compute a generalized least squares estimate and prove that it costs $O(N)$. A critical part of the proof is in ensuring that the number of iterations required is $O(1)$ which follows from keeping a certain matrix norm below $1-\delta$ for some $\delta > 0$. We extend the method to a generalized linear mixed model for logistic regression. We use backfitting within an iteratively reweighted penalized least squares algorithm. The specific approach is a version of penalized quasi-likelihood due to Schall (1991). A straightforward version of Schall's algorithm would also cost more than $N^{3/2}$ because it requires the trace of the inverse of a large matrix. We approximate that quantity at cost $O(N)$ and prove that this substitution makes an asymptotically negligible difference. Our backfitting algorithm also collapses the fixed effect with one random effect (clubbed backfitting) at a time in a way that is analogous to the collapsed Gibbs sampler of Papaspiliopoulos et al. (2020). We use a symmetric operator that

facilitates efficient covariance computation. We illustrate our method on a real dataset from Stitch Fix. By properly accounting for crossed random effects we show that a naive linear and logistic regression could underestimate sampling variances by several hundred fold.

Next we consider the scalability problem of the Bayesian analogue of the clubbed backfitting algorithm. We analyze the convergence rate of a collapsed Gibbs sampler for such models, and show that in many practical scenarios, the collapsed Gibbs sampler has a bounded relaxation time, and is therefore scalable. The theoretical analysis involves many new tools in approaching this problem, including a connection between relaxation time and autoregression matrix, concentration inequalities, and results from random matrix theory. We show that the effective sample size can be much smaller for a Gibbs sampler compared to the collapsed Gibbs variant.

Finally we develop complexity theory for three different types of coordinate-wise inference algorithm (backfitting, Gibbs sampling and coordinate ascent variational inference (CAVI)) for crossed random effect problem under a stringent sampling model. Our analysis relates the rate of convergence of these to auxiliary random walks on random bipartite graphs. As for the plain vanilla version of these algorithms, we exhibit negative scalability results, whereas for variations, positive ones.

# Acknowledgements

First and foremost, I am profoundly grateful to my advisers, Art B. Owen and Trevor Hastie, for being exceptional advisors, remarkable teachers and amazing people. They were very supportive, encouraging, understanding and considerate throughout my Ph.D. They cared about me both academically and personally. In every way, they have been a strong advocate for me: cheering me on every time we meet, reminding me to make sure I keep my eye on the big picture, and providing valuable career advice every time I need it. They treated me like their own child. When some of my friends in the States went home for Thanksgiving, I never felt alone. I could not have asked for better guides. As a child, my teacher asked me to write the same sentence hundreds of times to improve my handwriting. I absolutely hated it. I feel like I could write one sentence hundreds of times today: Thank you for everything.

I would also like to thank Professor Susan Athey and Scott Linderman for providing insightful comments and invaluable feedback to my thesis work as members of both my oral committee and reading committee. I also thank Professor Jonathan Pritchard for chairing my dissertation committee.

During my years at Stanford, I was very fortunate to collaborate with/learn from many wonderful people, both at Stanford and outside, including Dr. Subhas Banerjee, Dr. Monique T Barakat, Dr. Ruggero Bellio, Dr. Jiming Jiang, Dr. Julia Palacios, Dr. Lorenzo Cappello, Dr. Omiros Papaspiliopoulos, Dr. Cristiano Varin, Dr. Giacomo Zanella and Chenyang Zhong. I would like to thank National Science Foundation for supporting this research under grant IIS-1837931.

It is with deep gratitude that I thank all my professors at Stanford. I thank Professor Stephen Boyd, Sourav Chatterjee, Persi Diaconis, Iain M. Johnstone, George

C. Papanicolaou and David Siegmund. Inspiring teachers and mentors at Indian Statistical Institute continuously encouraged me to work on challenging problems. I am indebted to Professor Bimal Roy, Sourabh Bhattacharya, Subir Bhandari for their support. Indian Statistical Institute gave me a terrific group of friends from whom I learnt a lot. To Rahul Biswas, Partha Sarathi Chowdhury, Ashley Datta, Partha Pratim Ghosh, Subhayan Mondal and Somabha Mukherjee – a hearty thanks.

I could not have imagined navigating through this experience without my cohort. I will never forget the lunches we had together, discussing various topics, even though we saw each other less in recent years. Special thanks to Damian, Maxime, Michael and Suyash. I would also like to thank Bay area seniors who would make my weekends enjoyable : Kinjal, Preetam, Sayantan, Sudip. It would be a crime not to mention Souvik, not only for the long drawn out arguments ranging from world wars to stochastic calculus, but also for being the younger brother. Thank you.

On the personal front, I would like to thank my parents who worked tirelessly over all these years to make sure I had access to the best level of education. My pursuit of higher education outside my country, India, would not have been possible without their support and enthusiasm despite my continued absence. I would not be the person I am without them - for everything they have done, I dedicate this thesis to them. Additionally, I would like to thank my cousin (Saibal Bose) for always being my role model.

I would like to thank an organization that influenced me the most in my life. It is my high school, Ramakrishna Mission Vidyalaya. Not only I benefited from them on the academic front, but also on almost all other aspects of life. They have been doing a phenomenal work since Swami Vivekananda founded it in 1897. Be it providing resources to flood distressed areas or be it uplifting the financially and socially backward people through providing educations to the children. Without them, I would not be writing this today.

Last but not the least, I would like to extend my everlasting gratitude to my best friend and partner, Prathyusha for her unconditional love, support and care. I cannot thank her enough for all that she has done for me. She makes my life beautiful.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The amount of available data in the age of "big data" exceeds our ability to process and extract information from their sources, just as it was over a century ago when most computations had to be handled manually. According to a Forbes article Gil (2021), the amount of data created, captured, copied, and consumed in the world increased from 1.2 trillion gigabytes to 59 trillion gigabytes in the period 2010 to 2020. That is an almost 5,000% growth!! As collected data becomes richer and more complex, it presents new opportunities and challenges to statisticians. It has become more important than ever to develop computationally feasible and statistically efficient methods.

Since the beginning of my PhD, I have been interested in statistical methods that would be useful for a practitioner. I was excited when my advisor Art Owen told me about a problem that the practitioners are facing. Stitch Fix, an online personal styling service, collects data on clients' opinions of items they received, as well as information about the clients and items themselves. The factors are clients and items. This kind of categorical variable is intrinsically different from, say, three types of iris flower. Key differences include:

**1)** the great number of levels,

**2)** a power law or approximate power law frequency for those levels, and

**3)** the constant arrival of previously unseen levels.

The factors are quite useful inputs to prediction models. We think they should often be treated as random effects. Cookies are frequently cleared (Coey and Bailey, 2016) and it is therefore pointless to tune a model about web pages to the specific past cookies. Cookie data is more useful if viewed as a sample from a distribution that will generate the future cookies. It is also reasonable to treat short-lived products and web pages that turn over faster than customers do as random effects. Brad Klingenberg from Stitch Fix mentioned to Art Owen that data analytics team at Stitch Fix wanted to fit mixed models to their data sets, but existing methodology and software could not scale to data sets with millions of observations.

High-dimensional crossed random effect models is the modern canonical framework for understanding the relative importance of different sources of variation in a data set, with applications from political science to recommendation systems (Gelman, 2005; Gao and Owen, 2020; Goplerud, 2021). This thesis is dedicated to study such problems for this fundamental class of models in applied Statistics. Our main goal is to develop methods for mixed models that satisfy the following criteria.

- They **must** be computable in $O(N)$ time where $N$ is the number of observations.
- At most, they **should** produce conservative confidence intervals or variance estimates.
- We **prefer** them to be statistically efficient.

As part of that effort, this thesis covers three interrelated topics based on a series of papers we have written. They can be classified as follows:

I *Linear Crossed Random Effect*

The response $Y_{ij} \in \mathbb{R}$ in a prototypical linear mixed model is modeled by

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + b_j + e_{ij}, \quad 1 \leqslant i \leqslant R, \quad 1 \leqslant j \leqslant C \tag{1.1}$$

Many applications have $1 \ll R, C \ll N \ll RC$. For that problem, the cost to compute the Maximum Likelihood Estimate (MLE) on $N$ data points grows at best like $O(N^{3/2})$.

Here we present an iterative algorithm based on a backfitting approach from Buja et al. (1989). This algorithm converges to the generalized least square (GLS) solution with the cost of each iteration being $O(N)$.

The speed of convergence depends on a certain matrix norm. If the norm remains bounded strictly below 1 as $N \to \infty$, then the number of iterations to convergence is $O(1)$. We show that the matrix norm is $O(1)$ with probability tending to one, under some regularity conditions on observation pattern.

This is based on joint work with Trevor Hastie and Art Owen that appears in our paper Ghosh et al. (2022).

II *Generalized Linear Crossed Random Effect*

In practice in many cases the responses are discrete. For instance, customer might give a binary answer to a question about whether they liked a product or not. Unlike the linear setting where OLS is consistent, the naive logistic regression is inconsistent for generalized linear model with crossed random effect. The extension of maximum likelihood procedures to generalized linear mixed models has been well-studied and implemented in software like Bates et al. (2015). As shown in Bates (2014), such procedures, in addition to requiring strong parametric assumptions requires computation time $O(N^{3/2})$ to compute the estimates.

We model the binary response $Y_{ij} \in \{0, 1\}$ in terms of covariates $x_{ij} \in \mathbb{R}^p$ and row and column random effects $a_i$ and $b_j$ (distributed as independent zero-mean Gaussian random variables) :

$$\Pr(Y_{ij} = 1 \,|\, \boldsymbol{a}, \boldsymbol{b}) = \Pr(Y_{ij} = 1 \,|\, a_i, b_j) = \frac{\exp(x_{ij}^\mathsf{T}\beta + a_i + b_j)}{1 + \exp(x_{ij}^\mathsf{T}\beta + a_i + b_j)}. \tag{1.2}$$

We make suitable adaptations of penalized quasi-likelihood algorithm of Schall (1991) to mixed effects logistic regression models where there are crossed random effects. One adaptation comes from the fact that many of the feature variables in the commerce problem have a strong overlap with the factors. That brings a linear dependence that can slow convergence of iterative methods that alternate between updating random effects and fixed effects. We have devised a 'clubbing' algorithm that collapses the fixed effect with one random effect at a time (which in a way is analogous to the collapsed Gibbs sampler of Papaspiliopoulos et al. (2020)) with a $O(N)$ cost per iteration. This makes simultaneous updates avoiding the problem of slow convergence.

This is based on joint work with Trevor Hastie and Art Owen that appears in our paper Ghosh et al. (2021).

III *Collapsed Gibbs : a Bayesian perspective*

Here we explore Bayesian learning for large scale crossed random effect models. Markov chain Monte Carlo can be used to carry out the integrations over the factors (and the inference more generally). The most popular and convenient algorithm in this context is the Gibbs sampler, which samples the factors iteratively from their full conditional distributions. Gao and Owen (2017) showed that the Gibbs sampler takes order $N^{\frac{1}{2}}$ steps to mix, leading to order $N^{\frac{3}{2}}$ complexity (as the time cost for each iteration is of order $N$).

Recently, Papaspiliopoulos et al. (2020) proposed a collapsed Gibbs sampler for crossed random effects models under the condition that each customer rating equal number of products and each item being rated by equal number of customers. We analyze rigorously the superior performance (in terms of mixing and effective sample size) and scalability of the collapsed Gibbs sampler to data with unbalanced levels, which substantially relaxes the balancedness condition required for the analysis in Papaspiliopoulos et al. (2020). The analysis is based on models with a missingness mechanism, which is a new feature in the Bayesian approach to this problem.

This is based on joint work with Chenyang Zhong (Ghosh and Zhong (2021)).

Part IV connects the two school of thoughts discussed above.

IV *A tale of two cities : unified approach to analyze likelihood based approach and Gibbs sampler*

Both backfitting and Gibbs algorithms are iterative schemes that involve coordinate-wise updates, either in terms of optimization or sampling. Another Bayesian inference that could be of interest in this context is coordinate ascent variational inference (CAVI). Although variational inference is assumed to be faster, the plain vanilla version is also not scalable in this context.

Here we unify two perspectives : frequentist (likelihood based approach) and Bayesian (MCMC based approach). We relate the rate of convergence of the schemes studied in earlier chapters and that of CAVI to auxiliary random walks on random bipartite graphs for any number of factors using a common unifying framework for all three schemes. The theory in this chapter is motivated by scalability problems, however, most of our arguments are non asymptotic in nature. We leverage very recent results for the fast mixing of these random walks to prove the conjecture in Papaspiliopoulos et al. (2020) and establish that the slight modification of these algorithms are scalable.

This is based on joint work with Omiros Papaspiliopoulos and Giacomo Zanella.

# Chapter 2

# Linear Mixed Model with Crossed Random Effect

## 2.1 Introduction

To estimate a regression when the errors have a non-identity covariance matrix, we usually turn first to generalized least squares (GLS). Somewhat surprisingly, GLS proves to be computationally challenging in the very simple setting of the unbalanced crossed random effects models that we study here. For that problem, the cost to compute the GLS estimate on $N$ data points grows at best like $O(N^{3/2})$ under the usual algorithms. If we additionally assume Gaussian errors, then Gao and Owen (2020) show that even evaluating the likelihood one time costs at least a multiple of $N^{3/2}$ (also see Raudenbush (1993)). These costs make the usual algorithms for GLS infeasible for large data sets such as those arising in electronic commerce.

In this chapter, we present an iterative algorithm based on a backfitting approach from Buja et al. (1989). This algorithm is known to converge to the GLS solution. The cost of each iteration is $O(N)$ and so we also study how the number of iterations grows with $N$.

The crossed random effects model we consider has

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + b_j + e_{ij}, \quad 1 \leqslant i \leqslant R, \quad 1 \leqslant j \leqslant C \tag{2.1}$$

for random effects $a_i$ and $b_j$ and an error $e_{ij}$ with a fixed effects regression parameter $\beta \in \mathbb{R}^p$ for the covariates $x_{ij} \in \mathbb{R}^p$. We assume that $a_i \overset{\text{iid}}{\sim} (0, \sigma_A^2)$, $b_j \overset{\text{iid}}{\sim} (0, \sigma_B^2)$, and $e_{ij} \overset{\text{iid}}{\sim} (0, \sigma_E^2)$ are all independent. It is thus a mixed effects model in which the random portion has a crossed structure. The GLS estimate is also the maximum likelihood estimate (MLE), when $a_i$, $b_j$ and $e_{ij}$ are Gaussian. Because we assume that $p$ is fixed as $N$ grows, we often leave $p$ out of our cost estimates, giving instead the complexity in $N$.

The GLS estimate $\hat{\beta}_{\text{GLS}}$ for crossed random effects can be efficiently estimated if all $R \times C$ values are available. Our motivating examples involve ratings data where $R$ people rate $C$ items and then it is usual that the data are very unbalanced with a haphazard observational pattern in which only $N \ll R \times C$ of the $(x_{ij}, Y_{ij})$ pairs are observed. The crossed random effects setting is significantly more difficult than a hierarchical model with just $a_i + e_{ij}$ but no $b_j$ term. Then the observations for index $j$ are 'nested within' those for each level of index $i$. The result is that the covariance matrix of all observed $Y_{ij}$ values has a block diagonal structure allowing GLS to be computed in $O(N)$ time.

Hierarchical models are very well suited to Bayesian computation (Gelman and Hill, 2006). Crossed random effects are a much greater challenge. Gao and Owen (2017) find that the Gibbs sampler can take $O(N^{1/2})$ iterations to converge to stationarity, with each iteration costing $O(N)$ leading once again to $O(N^{3/2})$ cost. For more examples where the costs of solving equations versus sampling from a covariance attain the same rate see Goodman and Sokal (1989) and Roberts and Sahu (1997). As further evidence of the difficulty of this problem, the Gibbs sampler was one of nine MCMC algorithms that Gao and Owen (2017) found to be unsatisfactory. Furthermore, Bates et al. (2015) removed the `mcmcsamp` function from the R package lme4 because it was considered unreliable even for the problem of sampling the posterior distribution of the parameters from previously fitted models, and even for those with random effects variances near zero.

Papaspiliopoulos et al. (2020) present an exception to the high cost of a Bayesian approach for crossed random effects. They propose a collapsed Gibbs sampler that can potentially mix in $O(1)$ iterations. To prove this rate, they make an extremely stringent assumption that every index $i = 1, \ldots, R$ appears in the same number $N/C$ of observed data points and similarly every $j = 1, \ldots, C$ appears in $N/R$ data points. Such a condition is tantamount to requiring a designed experiment for the data and it is much stronger than what their algorithm seems to need in practice. Under that condition their mixing rate asymptotes to a quantity $\rho_{\mathrm{aux}}$, described in our discussion section, that in favorable circumstances is $O(1)$. They find empirically that their sampler has a cost that scales well in many data sets where their balance condition does not hold.

In this chapter we study an iterative linear operation, known as backfitting, for GLS. Each iteration costs $O(N)$. The speed of convergence depends on a certain matrix norm of that iteration, which we exhibit below. If the norm remains bounded strictly below 1 as $N \to \infty$, then the number of iterations to convergence is $O(1)$. We are able to show that the matrix norm is $O(1)$ with probability tending to one, under conditions where the number of observations per row (or per column) is random and even the expected row or column counts may vary, though in a narrow range. While this is a substantial weakening of the conditions in Papaspiliopoulos et al. (2020), it still fails to cover many interesting cases. Like them, we find empirically that our algorithm scales much more broadly than under the conditions for which scaling is proved.

We suspect that the computational infeasibility of GLS leads many users to use ordinary least squares (OLS) instead. OLS has two severe problems. First, it is ***inefficient*** with $\mathrm{var}(\hat\beta_{\mathrm{OLS}})$ larger than $\mathrm{var}(\hat\beta_{\mathrm{GLS}})$. This is equivalent to OLS ignoring some possibly large fraction of the information in the data. Perhaps more seriously, OLS is ***naive***. It produces an estimate of $\mathrm{var}(\hat\beta_{\mathrm{OLS}})$ that can be too small by a large factor. That amounts to overestimating the quantity of information behind $\hat\beta_{\mathrm{OLS}}$, also by a potentially large factor.

The naivete of OLS can be countered by using better variance estimates. One can bootstrap it by resampling the row and column entities as in Owen (2007). There is also a version of Huber-White variance estimation for this case in econometrics. See for instance Cameron et al. (2011). While these methods counter the naivete of OLS, the inefficiency of OLS remains.

The method of moments algorithm in Gao and Owen (2020) gets consistent asymptotically normal estimates of $\beta$, $\sigma_A^2$, $\sigma_B^2$ and $\sigma_E^2$. It produces a GLS estimate $\hat\beta$ that is more efficient than OLS but still not fully efficient because it accounts for correlations due to only one of the two crossed random effects. While inefficient, it is not naive because its estimate of $\mathrm{var}(\hat\beta)$ properly accounts for variance due to $a_i$, $b_j$ and $e_{ij}$.

In this chapter we get a GLS estimate $\hat\beta$ that takes account of all three variance components, making it efficient. We also provide an estimate of $\mathrm{var}(\hat\beta)$ that accounts for all three components, so our estimate is not naive. Our algorithm requires consistent estimates of the variance components $\sigma_A^2$, $\sigma_B^2$ and $\sigma_E^2$ in computing $\hat\beta$ and $\widehat{\mathrm{var}}(\hat\beta)$. We use the method of moments estimators from Gao and Owen (2017) that can be computed in $O(N)$ work. By Gao and Owen (2017, Theorem 4.2), these estimates of $\sigma_A^2$, $\sigma_B^2$ and $\sigma_E^2$ are asymptotically uncorrelated and each of them has the same asymptotic variance it would have had were the other two variance components equal to zero. It is not known whether they are optimally estimated, much less optimal subject to an $O(N)$ cost constraint. The variance component estimates are known to be asymptotically normal (Gao, 2017).

The rest of this chapter is organized as follows. Section 3.2 introduces our notation and assumptions for missing data. Section 2.3 presents the backfitting algorithm from Buja et al. (1989). That algorithm was defined for smoothers, but we are able to cast the estimation of random effect parameters as a special kind of smoother. Section 2.4 proves our result about backfitting being convergent with a probability tending to one as the problem size increases. Section 2.5 shows numerical measures of the matrix norm of the backfitting operator. It remains bounded below and away from one under more conditions than our theory shows. We find that even one iteration of the lmer function in lme4 package Bates et al. (2015) has a cost that grows like $N^{3/2}$ in one setting and like $N^{2.1}$ in another, sparser one. The backfitting algorithm has cost $O(N)$ in both of these cases. Section 2.6 illustrates our GLS algorithm on some data provided to us by Stitch Fix, Inc. (2011). These are customer ratings of items of clothing on a ten point scale. Section 3.7 has a discussion of these results. An appendix contains some regression output for the Stitch Fix data.

## 2.2 Missingness

We adopt the notation from Gao and Owen (2020). We let $Z_{ij} \in \{0,1\}$ take the value 1 if $(x_{ij}, Y_{ij})$ is observed and 0 otherwise, for $i = 1,\dots,R$ and $j = 1,\dots,C$. In many of the contexts we consider, the missingness is not at random and is potentially informative. Handling such problems is outside the scope of this chapter, apart from a brief discussion in Section 3.7. It is already a sufficient challenge to work without informative missingness.

The matrix $Z \in \{0,1\}^{R\times C}$, with elements $Z_{ij}$ has $N_{i\bullet} = \sum_{j=1}^{C} Z_{ij}$ observations in 'row $i$' and $N_{\bullet j} = \sum_{i=1}^{R} Z_{ij}$ observations in 'column $j$'. We often drop the limits of summation so that $i$ is always summed over $1,\dots,R$ and $j$ over $1,\dots,C$. When we need additional symbols for row and column indices we use $r$ for rows and $s$ for columns. The total sample size is $N = \sum_i \sum_j Z_{ij} = \sum_i N_{i\bullet} = \sum_j N_{\bullet j}$.

There are two co-observation matrices, $Z^\mathsf{T} Z$ and $ZZ^\mathsf{T}$. Here $(Z^\mathsf{T} Z)_{js} = \sum_i Z_{ij} Z_{is}$ gives the number of rows in which data from both columns $j$ and $s$ were observed, while $(ZZ^\mathsf{T})_{ir} = \sum_j Z_{ij} Z_{rj}$ gives the number of columns in which data from both rows $i$ and $r$ were observed.

In our regression models, we treat $Z_{ij}$ as nonrandom. We are conditioning on the actual pattern of observations in our data. When we study the rate at which our backfitting algorithm converges, we consider $Z_{ij}$ drawn at random. That is, the analyst is solving a GLS conditionally on the pattern of observations and missingness, while we study the convergence rates that analyst will see for data drawn from a missingness mechanism defined in Section 2.4.2.

If we place all of the $Y_{ij}$ into a vector $\mathcal{Y} \in \mathbb{R}^N$ and $x_{ij}$ compatibly into a matrix $\mathcal{X} \in \mathbb{R}^{N\times p}$, then the naive and inefficient OLS estimator is

$$\hat\beta_{\mathrm{OLS}} = (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y}. \tag{2.2}$$

This can be computed in $O(Np^2)$ work. We prefer to use the GLS estimator

$$\hat\beta_{\mathrm{GLS}} = (\mathcal{X}^\mathsf{T}\mathcal{V}^{-1}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{V}^{-1}\mathcal{Y}, \tag{2.3}$$

where $\mathcal{V} \in \mathbb{R}^{N\times N}$ contains all of the $\operatorname{cov}(Y_{ij}, Y_{rs})$ in an ordering compatible with $\mathcal{X}$ and $\mathcal{Y}$. A naive algorithm costs $O(N^3)$ to solve for $\hat{\beta}_{\mathrm{GLS}}$. It can actually be solved through a Cholesky decomposition of an $(R+C)\times(R+C)$ matrix (Searle et al., 2006). That has cost $O(R^3 + C^3)$. Now $N \leqslant RC$, with equality only for completely observed data. Therefore $\max(R, C) \geqslant \sqrt{N}$, and so $R^3 + C^3 \geqslant N^{3/2}$. When the data are sparsely enough observed it is possible that $\min(R, C)$ grows more rapidly than $N^{1/2}$. In a numerical example in Section 2.5 we have $\min(R, C)$ growing like $N^{0.70}$. In a hierarchical model, with $a_i$ but no $b_j$ we would find $\mathcal{V}$ to be block diagonal and then $\hat{\beta}_{\mathrm{GLS}}$ could be computed in $O(N)$ work.

It has been known since Strassen (1969) that systems of equations can be solved more quickly than cubic time. Despite that, current software is still dominated by cubic time algorithms. Also none of the known solutions are quadratic and so in our setting the cost would be at least a multiple of $(R+C)^{2+\gamma}$ for some $\gamma > 0$ and hence not $O(N)$.

We can write our crossed effects model as

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{Z}_A\boldsymbol{a} + \mathcal{Z}_B\boldsymbol{b} + \boldsymbol{e} \tag{2.4}$$

for matrices $\mathcal{Z}_A \in \{0,1\}^{N\times R}$ and $\mathcal{Z}_B \in \{0,1\}^{N\times C}$. The $i$'th column of $\mathcal{Z}_A$ has ones for all of the $N$ observations that come from row $i$ and zeroes elsewhere. The definition of $\mathcal{Z}_B$ is analogous. The observation matrix can be written $Z = \mathcal{Z}_A^{\mathsf{T}}\mathcal{Z}_B$. The vector $\boldsymbol{e}$ has all $N$ values of $e_{ij}$ in compatible order. Vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ contain the row and column random effects $a_i$ and $b_j$. In this notation

$$\mathcal{V} = \mathcal{Z}_A\mathcal{Z}_A^{\mathsf{T}}\sigma_A^2 + \mathcal{Z}_B\mathcal{Z}_B^{\mathsf{T}}\sigma_B^2 + I_N\sigma_E^2, \tag{2.5}$$

where $I_N$ is the $N \times N$ identity matrix.

Our main computational problem is to get a value for $\mathcal{U} = \mathcal{V}^{-1}\mathcal{X} \in \mathbb{R}^{N\times p}$. To do that we iterate towards a solution $\boldsymbol{u} \in \mathbb{R}^N$ of $\mathcal{V}\boldsymbol{u} = \boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^N$ is one of the $p$ columns of $\mathcal{X}$. After that, finding

$$\hat{\beta}_{\mathrm{GLS}} = (\mathcal{X}^{\mathsf{T}}\mathcal{U})^{-1}(\mathcal{Y}^{\mathsf{T}}\mathcal{U})^{\mathsf{T}} \tag{2.6}$$

is not expensive, because $\mathcal{X}^\mathsf{T}\mathcal{U} \in \mathbb{R}^{p\times p}$ and we suppose that $p$ is not large.

If the data ordering in $\mathcal{Y}$ and elsewhere sorts by index $i$, breaking ties by index $j$, then $\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T} \in \{0,1\}^{N\times N}$ is a block matrix with $R$ blocks of ones of size $N_{i\bullet}\times N_{i\bullet}$ along the diagonal and zeroes elsewhere. The matrix $\mathcal{Z}_B\mathcal{Z}_B^\mathsf{T}$ will not be block diagonal in that ordering. Instead $P\mathcal{Z}_B\mathcal{Z}_B^\mathsf{T}P^\mathsf{T}$ will be block diagonal with $N_{\bullet j}\times N_{\bullet j}$ blocks of ones on the diagonal, for a suitable $N\times N$ permutation matrix $P$.

## 2.3 Backfitting algorithms

Our first goal is to develop computationally efficient ways to solve the GLS problem (2.6) for the linear mixed model (2.4). We use the backfitting algorithm that Hastie and Tibshirani (1990) and Buja et al. (1989) use to fit additive models. We write $\mathcal{V}$ in (2.5) as $\sigma_E^2\left(\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T}/\lambda_A + \mathcal{Z}_B\mathcal{Z}_B^\mathsf{T}/\lambda_B + I_N\right)$ with $\lambda_A = \sigma_E^2/\sigma_A^2$ and $\lambda_B = \sigma_E^2/\sigma_B^2$, and define $\mathcal{W} = \sigma_E^2\mathcal{V}^{-1}$. Then the GLS estimate of $\beta$ is

$$\hat\beta_{\mathrm{GLS}} = \arg\min_\beta (\mathcal{Y}-\mathcal{X}\beta)^\mathsf{T}\mathcal{W}(\mathcal{Y}-\mathcal{X}\beta) = (\mathcal{X}^\mathsf{T}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{W}\mathcal{Y} \tag{2.7}$$

and $\mathrm{cov}(\hat\beta_{\mathrm{GLS}}) = \sigma_E^2(\mathcal{X}^\mathsf{T}\mathcal{W}\mathcal{X})^{-1}$.

It is well known (e.g., Robinson (1991)) that we can obtain $\hat\beta_{\mathrm{GLS}}$ by solving the following penalized least-squares problem

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \|\mathcal{Y}-\mathcal{X}\beta-\mathcal{Z}_A\boldsymbol{a}-\mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{2.8}$$

Then $\hat\beta = \hat\beta_{\mathrm{GLS}}$ and $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are the best linear unbiased prediction (BLUP) estimates of the random effects. This derivation works for any number of factors, but it is instructive to carry it through initially for one.

### 2.3.1 One factor

For a single factor, we simply drop the $\mathcal{Z}_B\boldsymbol{b}$ term from (2.4) to get

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{Z}_A\boldsymbol{a} + \boldsymbol{e}.$$

Then $\mathcal{V} = \mathrm{cov}(\mathcal{Z}_A \boldsymbol{a} + \boldsymbol{e}) = \sigma_A^2 \mathcal{Z}_A \mathcal{Z}_A^\mathsf{T} + \sigma_E^2 I_N$, and $\mathcal{W} = \sigma_E^2 \mathcal{V}^{-1}$ as before. The penalized least squares problem is to solve

$$\min_{\beta, \boldsymbol{a}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_A \boldsymbol{a}\|^2 + \lambda_A \|\boldsymbol{a}\|^2. \tag{2.9}$$

We show the details as we need them for a later derivation.

The normal equations from (2.9) yield

$$\mathbf{0} = \mathcal{X}^\mathsf{T}(\mathcal{Y} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A \hat{\boldsymbol{a}}), \quad \text{and} \tag{2.10}$$

$$\mathbf{0} = \mathcal{Z}_A^\mathsf{T}(\mathcal{Y} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A \hat{\boldsymbol{a}}) - \lambda_A \hat{\boldsymbol{a}}. \tag{2.11}$$

Solving (3.18) for $\hat{\boldsymbol{a}}$ and multiplying the solution by $\mathcal{Z}_A$ yields

$$\mathcal{Z}_A \hat{\boldsymbol{a}} = \mathcal{Z}_A (\mathcal{Z}_A^\mathsf{T} \mathcal{Z}_A + \lambda_A I_R)^{-1} \mathcal{Z}_A^\mathsf{T} (\mathcal{Y} - \mathcal{X}\hat{\beta}) \equiv \mathcal{S}_A (\mathcal{Y} - \mathcal{X}\hat{\beta}),$$

for an $N \times N$ ridge regression "smoother matrix" $\mathcal{S}_A$. As we explain below this smoother matrix implements shrunken within-group means. Then substituting $\mathcal{Z}_A \hat{\boldsymbol{a}}$ into equation (3.17) yields

$$\hat{\beta} = (\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_A)\mathcal{X})^{-1} \mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_A)\mathcal{Y}. \tag{2.12}$$

Using the Sherman-Morrison-Woodbury (SMW) identity, one can show that $\mathcal{W} = I_N - \mathcal{S}_A$ and hence $\hat{\beta}$ above equals $\hat{\beta}_{\mathrm{GLS}}$ from (2.7). This is not in itself a new discovery; see for example Robinson (1991) or Hastie and Tibshirani (1990) (Section 5.3.3).

To compute the solution in (2.12), we need to compute $\mathcal{S}_A \mathcal{Y}$ and $\mathcal{S}_A \mathcal{X}$. The heart of the computation in $\mathcal{S}_A \mathcal{Y}$ is $(\mathcal{Z}_A^\mathsf{T} \mathcal{Z}_A + \lambda_A I_R)^{-1} \mathcal{Z}_A^\mathsf{T} \mathcal{Y}$. But $\mathcal{Z}_A^\mathsf{T} \mathcal{Z}_A = \mathrm{diag}(N_{1\bullet}, N_{2\bullet}, \ldots, N_{R\bullet})$ and we see that all we are doing is computing an $R$-vector of shrunken means of the elements of $\mathcal{Y}$ at each level of the factor $A$; the $i$th element is $\sum_j Z_{ij} Y_{ij} / (N_{i\bullet} + \lambda_A)$. This involves a single pass through the $N$ elements of $Y$, accumulating the sums into $R$ registers, followed by an elementwise scaling of the $R$ components. Then premultiplication by $\mathcal{Z}_A$ simply puts these $R$ shrunken means back into an $N$-vector in the appropriate positions. The total cost is $O(N)$. Likewise $\mathcal{S}_A\mathcal{X}$ does the same separately for each of the columns of $\mathcal{X}$. Hence the entire computational cost for (2.12) is $O(Np^2)$, the same order as regression on $\mathcal{X}$.

What is also clear is that the indicator matrix $\mathcal{Z}_A$ is not actually needed here; instead all we need to carry out these computations is the factor vector $f_A$ that records the level of factor $A$ for each of the $N$ observations. In the R language (R Core Team, 2015) the following pair of operations does the computation:

```
hat_a = tapply(y,fA,sum)/(table(fA)+lambdaA)
hat_y = hat_a[fA]
```

where `fA` is a categorical variable (factor) $f_A$ of length $N$ containing the row indices $i$ in an order compatible with $Y \in \mathbb{R}^N$ (represented as `y`) and `lambdaA` is $\lambda_A = \sigma_A^2/\sigma_E^2$.

### 2.3.2 Two factors

With two factors we face the problem of incompatible block diagonal matrices discussed in Section 3.2. Define $\mathcal{Z}_G = (\mathcal{Z}_A\!:\!\mathcal{Z}_B)$ ($R{+}C$ columns), $\mathcal{D}_\lambda = \mathrm{diag}(\lambda_A I_R, \lambda_B I_C)$, and $\boldsymbol{g}^\mathsf{T} = (\boldsymbol{a}^\mathsf{T}, \boldsymbol{b}^\mathsf{T})$. Then solving (3.12) is equivalent to

$$\min_{\beta,\boldsymbol{g}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_G\boldsymbol{g}\|^2 + \boldsymbol{g}^\mathsf{T}\mathcal{D}_\lambda\boldsymbol{g}. \tag{2.13}$$

A derivation similar to that used in the one-factor case gives

$$\hat\beta = H_{\mathrm{GLS}}\mathcal{Y} \quad \text{for} \quad H_{\mathrm{GLS}} = (\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_G)\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_G), \tag{2.14}$$

where the hat matrix $H_{\mathrm{GLS}}$ is written in terms of a smoother matrix

$$\mathcal{S}_G = \mathcal{Z}_G(\mathcal{Z}_G^\mathsf{T}\mathcal{Z}_G + \mathcal{D}_\lambda)^{-1}\mathcal{Z}_G^\mathsf{T}. \tag{2.15}$$

We can again use SMW to show that $I_N - \mathcal{S}_G = \mathcal{W}$ and hence the solution $\hat\beta = \hat\beta_{\mathrm{GLS}}$ in (2.7). But in applying $\mathcal{S}_G$ we do not enjoy the computational simplifications that occurred in the one factor case, because

$$\mathcal{Z}_G^\mathsf{T}\mathcal{Z}_G = \begin{pmatrix} \mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A & \mathcal{Z}_A^\mathsf{T}\mathcal{Z}_B \\ \mathcal{Z}_B^\mathsf{T}\mathcal{Z}_A & \mathcal{Z}_B^\mathsf{T}\mathcal{Z}_B \end{pmatrix} = \begin{pmatrix} \mathrm{diag}(N_{i\bullet}) & Z \\ Z^\mathsf{T} & \mathrm{diag}(N_{\bullet j}) \end{pmatrix},$$

where $Z \in \{0,1\}^{R\times C}$ is the observation matrix which has no special structure. Therefore we need to invert an $(R+C)\times(R+C)$ matrix to apply $\mathcal{S}_G$ and hence to solve (2.14), at a cost of at least $O(N^{3/2})$ (see Section 3.2).

Rather than group $\mathcal{Z}_A$ and $\mathcal{Z}_B$, we keep them separate, and develop an algorithm to apply the operator $\mathcal{S}_G$ efficiently. Consider a generic response vector $\mathcal{R}$ (such as $\mathcal{Y}$ or a column of $\mathcal{X}$) and the optimization problem

$$\min_{\boldsymbol{a},\boldsymbol{b}} \|\mathcal{R} - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{2.16}$$

Using $\mathcal{S}_G$ defined at (2.15) in terms of the indicator variables $\mathcal{Z}_G \in \{0,1\}^{N\times(R+C)}$ it is clear that the fitted values are given by $\widehat{\mathcal{R}} = \mathcal{S}_G\mathcal{R}$. Solving (3.34) would result in two blocks of estimating equations similar to equations (3.17) and (3.18). These can be written

$$\begin{aligned} \mathcal{Z}_A\hat{\boldsymbol{a}} &= \mathcal{S}_A(\mathcal{R} - \mathcal{Z}_B\hat{\boldsymbol{b}}), \quad \text{and} \\ \mathcal{Z}_B\hat{\boldsymbol{b}} &= \mathcal{S}_B(\mathcal{R} - \mathcal{Z}_A\hat{\boldsymbol{a}}), \end{aligned} \tag{2.17}$$

where $\mathcal{S}_A = \mathcal{Z}_A(\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A + \lambda_A I_R)^{-1}\mathcal{Z}_A^\mathsf{T}$ is again the ridge regression smoothing matrix for row effects and similarly $\mathcal{S}_B = \mathcal{Z}_B(\mathcal{Z}_B^\mathsf{T}\mathcal{Z}_B + \lambda_B I_C)^{-1}\mathcal{Z}_B^\mathsf{T}$ the smoothing matrix for column effects. We solve these equations iteratively by block coordinate descent, also known as backfitting. The iterations converge to the solution of (3.34) (Buja et al., 1989; Hastie and Tibshirani, 1990).

It is evident that $\mathcal{S}_A, \mathcal{S}_B \in \mathbb{R}^{N\times N}$ are both symmetric matrices. It follows that the limiting smoother $\mathcal{S}_G$ formed by combining them is also symmetric. See Hastie and Tibshirani (1990, page 120). We will need this result later for an important computational shortcut.

Here the simplifications we enjoyed in the one-factor case once again apply. Each step applies its operator to a vector (the terms in parentheses on the right hand side in (2.17)). For both $\mathcal{S}_A$ and $\mathcal{S}_B$ these are simply the shrunken-mean operations described for the one-factor case, separately for factor $A$ and $B$ each time. As before, we do not need to actually construct $\mathcal{Z}_B$, but simply use a factor $f_B$ that records the level of factor $B$ for each of the $N$ observations.

The above description holds for a generic response $\mathcal{R}$; we apply that algorithm to $\mathcal{Y}$ and each column of $\mathcal{X}$ to obtain the quantities $\mathcal{S}_G\mathcal{X}$ and $\mathcal{S}_G\mathcal{Y}$ that we need to compute $H_{\text{GLS}}\mathcal{Y}$ in (2.14). Importantly, it is very amenable to parallel computation in tools such as MapReduce (Dean and Ghemawat (2008)) or Apache Spark (Zaharia et al. (2010)). Now solving (2.14) is $O(Np^2)$ plus a negligible $O(p^3)$ cost. These computations deliver $\hat{\beta}_{\text{GLS}}$; if the BLUP estimates $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are also required, the same algorithm can be applied to the response $\mathcal{Y} - \mathcal{X}\hat{\beta}_{\text{GLS}}$, retaining the $\boldsymbol{a}$ and $\boldsymbol{b}$ at the final iteration. We can also write

$$\text{cov}(\hat{\beta}_{\text{GLS}}) = \sigma_E^2(\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_G)\mathcal{X})^{-1}. \tag{2.18}$$

It is also clear that we can trivially extend this approach to accommodate any number of factors.

### 2.3.3 Centered operators

The matrices $\mathcal{Z}_A$ and $\mathcal{Z}_B$ both have row sums all ones, since they are factor indicator matrices ("one-hot encoders"). This creates a nontrivial intersection between their column spaces, and that of $\mathcal{X}$ since we always include an intercept, that can cause backfitting to converge more slowly. In this section we show how to counter this intersection of column spaces to speed convergence. We work with this two-factor model

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{2.19}$$

*Lemma* 2.3.1. If $\mathcal{X}$ in model (2.19) includes a column of ones (intercept), and $\lambda_A > 0$ and $\lambda_B > 0$, then the solutions for $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy $\sum_{i=1}^R a_i = 0$ and $\sum_{j=1}^C b_j = 0$.

*Proof.* It suffices to show this for one factor and with $\mathcal{X} = \mathbf{1}$. The objective is now

$$\min_{\beta,\boldsymbol{a}} \|\mathcal{Y} - \mathbf{1}\beta - \mathcal{Z}_A\boldsymbol{a}\|^2 + \lambda_A\|\boldsymbol{a}\|^2. \tag{2.20}$$

Notice that for any candidate solution $(\beta, \{a_i\}_1^R)$, the alternative solution $(\beta+c, \{a_i - c\}_1^R)$ leaves the loss part of (2.20) unchanged, since the row sums of $\mathcal{Z}_A$ are all one. Hence if $\lambda_A > 0$, we would always improve $\boldsymbol{a}$ by picking $c$ to minimize the penalty term $\sum_{i=1}^R (a_i - c)^2$, or $c = (1/R)\sum_{i=1}^R a_i$. □

It is natural then to solve for $\boldsymbol{a}$ and $\boldsymbol{b}$ with these constraints enforced, instead of waiting for them to simply emerge in the process of iteration.

*Theorem* 2.3.1. Consider the generic optimization problem

$$\min_{\boldsymbol{a}} \|\mathcal{R} - \mathcal{Z}_A\boldsymbol{a}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 \quad \text{subject to } \sum_{i=1}^R a_i = 0. \tag{2.21}$$

Define the partial sum vector $\mathcal{R}^+ = \mathcal{Z}_A^\mathsf{T}\mathcal{R}$ with components $\mathcal{R}_i^+ = \sum_j Z_{ij}\mathcal{R}_{ij}$, and let

$$w_i = \frac{(N_{i\bullet} + \lambda)^{-1}}{\sum_r (N_{r\bullet} + \lambda)^{-1}}.$$

Then the solution $\hat{\boldsymbol{a}}$ is given by

$$\hat{a}_i = \frac{\mathcal{R}_i^+ - \sum_r w_r\mathcal{R}_r^+}{N_{i\bullet} + \lambda_A}, \quad i = 1, \ldots, R. \tag{2.22}$$

Moreover, the fit is given by

$$\mathcal{Z}_A\hat{\boldsymbol{a}} = \tilde{\mathcal{S}}_A\mathcal{R},$$

where $\tilde{\mathcal{S}}_A$ is a symmetric operator.

The computations are a simple modification of the non-centered case.

*Proof.* Let $M$ be an $R \times R$ orthogonal matrix with first column $\mathbf{1}/\sqrt{R}$. Then $\mathcal{Z}_A\boldsymbol{a} = \mathcal{Z}_A MM^\mathsf{T}\boldsymbol{a} = \tilde{\mathcal{G}}\tilde{\boldsymbol{\gamma}}$ for $\tilde{\mathcal{G}} = \mathcal{Z}_A M$ and $\tilde{\boldsymbol{\gamma}} = M^\mathsf{T}\boldsymbol{a}$. Reparametrizing in this way leads to the equivalent problem

$$\min_{\tilde{\boldsymbol{\gamma}}} \|\mathcal{R} - \tilde{\mathcal{G}}\tilde{\boldsymbol{\gamma}}\|^2 + \lambda_A\|\tilde{\boldsymbol{\gamma}}\|^2, \quad \text{subject to } \tilde{\gamma}_1 = 0. \tag{2.23}$$

To solve (2.23), we simply drop the first column of $\tilde{\mathcal{G}}$. Let $\mathcal{G} = \mathcal{Z}_A Q$ where $Q$ is the matrix $M$ omitting the first column, and $\boldsymbol{\gamma}$ the corresponding subvector of $\tilde{\boldsymbol{\gamma}}$ having $R-1$ components. We now solve

$$\min_{\tilde{\boldsymbol{\gamma}}} \|\mathcal{R} - \mathcal{G}\boldsymbol{\gamma}\|^2 + \lambda_A\|\tilde{\boldsymbol{\gamma}}\|^2 \tag{2.24}$$

with no constraints, and the solution is $\hat{\boldsymbol{\gamma}} = (\mathcal{G}^\mathsf{T}\mathcal{G} + \lambda_A I_{R-1})^{-1}\mathcal{G}^\mathsf{T}\mathcal{R}$. The fit is given by $\mathcal{G}\hat{\boldsymbol{\gamma}} = \mathcal{G}(\mathcal{G}^\mathsf{T}\mathcal{G} + \lambda_A I_{R-1})^{-1}\mathcal{G}^\mathsf{T}\mathcal{R} = \tilde{\mathcal{S}}_A\mathcal{R}$, and $\tilde{\mathcal{S}}_A$ is clearly a symmetric operator.

To obtain the simplified expression for $\hat{\boldsymbol{a}}$, we write

$$\begin{aligned}
\mathcal{G}\hat{\gamma} &= \mathcal{Z}_A Q(Q^\mathsf{T}\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A Q + \lambda_A I_{R-1})^{-1}Q^\mathsf{T}\mathcal{Z}_A^\mathsf{T}\mathcal{R} \\
&= \mathcal{Z}_A Q(Q^\mathsf{T} D Q + \lambda_A I_{R-1})^{-1}Q^\mathsf{T}\mathcal{R}^+ \\
&= \mathcal{Z}_A\hat{\boldsymbol{a}},
\end{aligned} \tag{2.25}$$

with $D = \mathrm{diag}(N_{i\bullet})$. We write $H = Q(Q^\mathsf{T}DQ + \lambda_A I_{R-1})^{-1}Q^\mathsf{T}$ and $\tilde{Q} = (D + \lambda_A I_R)^{\frac{1}{2}}Q$, and let

$$\tilde{H} = (D + \lambda_A I_R)^{\frac{1}{2}} H (D + \lambda_A I_R)^{\frac{1}{2}} = \tilde{Q}(\tilde{Q}^\mathsf{T}\tilde{Q})^{-1}\tilde{Q}^\mathsf{T}. \tag{2.26}$$

Now (2.26) is a projection matrix in $\mathbb{R}^R$ onto a $R-1$ dimensional subspace. Let $\tilde{q} = (D + \lambda_A I_R)^{-\frac{1}{2}}\mathbf{1}$. Then $\tilde{q}^\mathsf{T}\tilde{Q} = \mathbf{0}$, and so

$$\tilde{H} = I_R - \frac{\tilde{q}\tilde{q}^\mathsf{T}}{\|\tilde{q}\|^2}.$$

Unraveling this expression we get

$$H = (D + \lambda_A I_R)^{-1} - (D + \lambda_A I_R)^{-1}\frac{\mathbf{1}\mathbf{1}^\mathsf{T}}{\mathbf{1}^\mathsf{T}(D + \lambda_A I_R)^{-1}\mathbf{1}}(D + \lambda_A I_R)^{-1}.$$

With $\hat{\boldsymbol{a}} = H\mathcal{R}^+$ in (2.25), this gives the expressions for each $\hat{a}_i$ in (2.22). Finally, $\tilde{\mathcal{S}}_A = \mathcal{Z}_A H \mathcal{Z}_A^\mathsf{T}$ is symmetric. $\square$

### 2.3.4 Covariance matrix for $\hat{\beta}_{\mathrm{GLS}}$ with centered operators

In Section 2.3.2 we saw in (2.18) that we get a simple expression for $\mathrm{cov}(\hat{\beta}_{\mathrm{GLS}})$. This simplicity relies on the fact that $I_N - \mathcal{S}_G = \mathcal{W} = \sigma_E^2 \mathcal{V}^{-1}$, and the usual cancellation occurs when we use the sandwich formula to compute this covariance. When we backfit with our centered smoothers we get a modified residual operator $I_N - \widetilde{\mathcal{S}}_G$ such that the analog of (2.14) still gives us the required coefficient estimate:

$$\hat{\beta}_{\mathrm{GLS}} = (\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{Y}. \tag{2.27}$$

However, $I_N - \widetilde{\mathcal{S}}_G \neq \sigma_E^2 \mathcal{V}^{-1}$, and so now we need to resort to the sandwich formula $\mathrm{cov}(\hat{\beta}_{\mathrm{GLS}}) = H_{\mathrm{GLS}} \mathcal{V} H_{\mathrm{GLS}}^\mathsf{T}$ with $H_{\mathrm{GLS}}$ from (2.14). Expanding this we find that $\mathrm{cov}(\hat{\beta}_{\mathrm{GLS}})$ equals

$$(\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G) \cdot \mathcal{V} \cdot (I_N - \widetilde{\mathcal{S}}_G)\mathcal{X}(\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{X})^{-1}.$$

While this expression might appear daunting, the computations are simple. Note first that while $\hat{\beta}_{\mathrm{GLS}}$ can be computed via $\tilde{\mathcal{S}}_G\mathcal{X}$ and $\tilde{\mathcal{S}}_G\mathcal{Y}$ this expression for $\mathrm{cov}(\hat{\beta}_{\mathrm{GLS}})$ also involves $\mathcal{X}^\mathsf{T}\tilde{\mathcal{S}}_G$. When we use the centered operator from Theorem 2.3.1 we get a symmetric matrix $\tilde{\mathcal{S}}_G$. Let $\widetilde{\mathcal{X}} = (I_N - \tilde{\mathcal{S}}_G)\mathcal{X}$, the residual matrix after backfitting each column of $\mathcal{X}$ using these centered operators. Then because $\widetilde{\mathcal{S}}_G$ is symmetric, we have

$$\begin{aligned} \hat{\beta}_{\mathrm{GLS}} &= (\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}})^{-1}\widetilde{\mathcal{X}}^\mathsf{T}\mathcal{Y}, \quad \text{and} \\ \mathrm{cov}(\hat{\beta}_{\mathrm{GLS}}) &= (\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}})^{-1}\widetilde{\mathcal{X}}^\mathsf{T} \cdot \mathcal{V} \cdot \widetilde{\mathcal{X}}(\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}})^{-1}. \end{aligned} \tag{2.28}$$

Since $\mathcal{V} = \sigma_E^2\left(\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T}/\lambda_A + \mathcal{Z}_B\mathcal{Z}_B^\mathsf{T}/\lambda_B + I_N\right)$ (two low-rank matrices plus the identity), we can compute $\mathcal{V} \cdot \widetilde{\mathcal{X}}$ very efficiently, and hence also the covariance matrix in (2.28). The entire algorithm is summarized in Section 2.6.3.

## 2.4 Convergence of the matrix norm

In this section we prove a bound on the norm of the matrix that implements backfitting for our random effects $\boldsymbol{a}$ and $\boldsymbol{b}$ and show how this controls the number of iterations required. In our algorithm, backfitting is applied to $\mathcal{Y}$ as well as to each non-intercept column of $\mathcal{X}$ so we do not need to consider the updates for $\mathcal{X}\hat{\beta}$. It is useful to take account of intercept adjustments in backfitting, by the centerings described in Section 2.3 because the space spanned by $a_1, \dots, a_R$ intersects the space spanned by $b_1, \dots, b_C$ because both include an intercept column of ones.

In backfitting we alternate between adjusting $\boldsymbol{a}$ given $\boldsymbol{b}$ and $\boldsymbol{b}$ given $\boldsymbol{a}$. At any iteration, the new $\boldsymbol{a}$ is an affine function of the previous $\boldsymbol{b}$ and then the new $\boldsymbol{b}$ is an affine function of the new $\boldsymbol{a}$. This makes the new $\boldsymbol{b}$ an affine function of the previous $\boldsymbol{b}$. We will study that affine function to find conditions where the updates converge. If the $\boldsymbol{b}$ updates converge, then so must the $\boldsymbol{a}$ updates.

Because the updates are affine they can be written in the form

$$\boldsymbol{b} \leftarrow M\boldsymbol{b} + \eta$$

for $M \in \mathbb{R}^{C\times C}$ and $\eta \in \mathbb{R}^C$. We iterate this update and it is convenient to start with $\boldsymbol{b} = \mathbf{0}$. We already know from Buja et al. (1989) that this backfitting will converge. However, we want more. We want to avoid having the number of iterations required grow with $N$. We can write the solution $\boldsymbol{b}$ as

$$\boldsymbol{b} = \eta + \sum_{k=1}^{\infty} M^k \eta,$$

and in computations we truncate this sum after $K$ steps producing an error $\sum_{k>K} M^k\eta$. We want $\sup_{\eta\neq 0} \|\sum_{k>K} M^k\eta\|/\|\eta\| < \epsilon$ to hold with probability tending to one as the sample size increases for any $\epsilon$, given sufficiently large $K$. For this it suffices to have the spectral radius $\lambda_{\max}(M) < 1 - \delta$ hold with probability tending to one for some $\delta > 0$.

Now for any $1 \leqslant p \leqslant \infty$ we have

$$\lambda_{\max}(M) \leqslant \|M\|_p \equiv \sup_{\boldsymbol{x}\in\mathbb{R}^C\setminus\{\boldsymbol{0}\}} \frac{\|M\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}.$$

The explicit formula

$$\|M\|_1 \equiv \sup_{\boldsymbol{x}\in\mathbb{R}^C\setminus\{\boldsymbol{0}\}} \frac{\|M\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} = \max_{1\leqslant s\leqslant C} \sum_{j=1}^{C} |M_{js}|$$

makes the matrix $L_1$ matrix norm very tractable theoretically and so that is the one we study. We look at this and some other measures numerically in Section 2.5.

### 2.4.1 Updates

Recall that $Z \in \{0,1\}^{R\times C}$ describes the pattern of observations. In a model with no intercept, centering the responses and then taking shrunken means as in (2.17) would yield these updates

$$a_i \leftarrow \frac{\sum_s Z_{is}(Y_{is} - b_s)}{N_{i\bullet} + \lambda_A} \quad \text{and} \quad b_j \leftarrow \frac{\sum_i Z_{ij}(Y_{ij} - a_i)}{N_{\bullet j} + \lambda_B}.$$

The update from the old $\boldsymbol{b}$ to the new $\boldsymbol{a}$ and then to the new $\boldsymbol{b}$ takes the form $\boldsymbol{b} \leftarrow M\boldsymbol{b} + \eta$ for $M = M^{(0)}$ where

$$M^{(0)}_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_i \frac{Z_{is}Z_{ij}}{N_{i\bullet} + \lambda_A}.$$

This update $M^{(0)}$ alternates shrinkage estimates for $\boldsymbol{a}$ and $\boldsymbol{b}$ but does no centering. We don't exhibit $\eta$ because it does not affect the convergence speed.

In the presence of an intercept, we know that $\sum_i a_i = 0$ should hold at the solution and we can impose this simply and very directly by centering the $a_i$, taking

$$a_i \leftarrow \frac{\sum_s Z_{is}(Y_{is} - b_s)}{N_{i\bullet} + \lambda_A} - \frac{1}{R}\sum_{r=1}^{R} \frac{\sum_s Z_{rs}(Y_{rs} - b_s)}{N_{r\bullet} + \lambda_A}, \quad \text{and}$$

$$b_j \leftarrow \frac{\sum_i Z_{ij}(Y_{ij} - a_i)}{N_{\bullet j} + \lambda_B}.$$

The intercept estimate will then be $\hat{\beta}_0 = (1/C)\sum_j b_j$ which we can subtract from $b_j$ upon convergence. This iteration has the update matrix $M^{(1)}$ with

$$M^{(1)}_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r \frac{Z_{rs}(Z_{rj} - N_{\bullet j}/R)}{N_{r\bullet} + \lambda_A} \tag{2.29}$$

after replacing a sum over $i$ by an equivalent one over $r$. In practice, we prefer to use the weighted centering from Section 2.3.3 to center the $a_i$ because it provides a symmetric smoother $\tilde{\mathcal{S}}_G$ that supports computation of $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$. While it is more complicated to analyze it is easily computable and it satisfies the optimality condition in Theorem 2.3.1. The algorithm is for a generic response $\mathcal{R} \in \mathbb{R}^N$ such as $\mathcal{Y}$ or a column of $\mathcal{X}$. Let us illustrate it for the case $\mathcal{R} = \mathcal{Y}$. We begin with vector of $N$ values $Y_{ij} - b_j$ and so $Y_i^+ = \sum_s Z_{is}(Y_{is} - b_s)$. Then $w_i = (N_{i\bullet} + \lambda_A)^{-1}/\sum_r (N_{r\bullet} + \lambda_A)^{-1}$ and the updated $a_r$ is

$$\frac{Y_r^+ - \sum_i w_i Y_i^+}{N_{r\bullet} + \lambda_A} = \frac{\sum_s Z_{rs}(Y_{rs} - b_s) - \sum_i w_i \sum_s Z_{is}(Y_{is} - b_s)}{N_{r\bullet} + \lambda_A}.$$

Using shrunken averages of $Y_{ij} - a_i$, the new $b_j$ are

$$b_j = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r Z_{rj}\left(Y_{rj} - \frac{\sum_s Z_{rs}(Y_{rs} - b_s) - \sum_i w_i \sum_s Z_{is}(Y_{is} - b_s)}{N_{r\bullet} + \lambda_A}\right).$$

Now $\boldsymbol{b} \leftarrow M\boldsymbol{b} + \eta$ for $M = M^{(2)}$, where

$$M^{(2)}_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r \frac{Z_{rj}}{N_{r\bullet} + \lambda_A}\left(Z_{rs} - \frac{\sum_i \frac{Z_{is}}{N_{i\bullet} + \lambda_A}}{\sum_i \frac{1}{N_{i\bullet} + \lambda_A}}\right). \tag{2.30}$$

Our preferred algorithm applies the optimal update from Theorem 2.3.1 to both $\boldsymbol{a}$ and $\boldsymbol{b}$ updates. With that choice we do not need to decide beforehand which random effects to center and which to leave uncentered to contain the intercept. We call the corresponding matrix $M^{(3)}$. Our theory below analyzes $\|M^{(1)}\|_1$ and $\|M^{(2)}\|_1$ which have simpler expressions than $\|M^{(3)}\|_1$.

Update $M^{(0)}$ uses symmetric smoothers for $A$ and $B$. Both are shrunken averages. The naive centering update $M^{(1)}$ uses a non-symmetric smoother $\mathcal{Z}_A(I_R - \mathbf{1}_R\mathbf{1}_R^\mathsf{T}/R)(\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A + \lambda_A I_R)^{-1}\mathcal{Z}_A^\mathsf{T}$ on the $a_i$ with a symmetric smoother on $b_j$ and hence it does not generally produce a symmetric smoother needed for efficient computation of $\widehat{\mathrm{cov}}(\hat\beta_{\mathrm{GLS}})$. The update $M^{(2)}$ uses two symmetric smoothers, one optimal and one a simple shrunken mean. The update $M^{(3)}$ takes the optimal smoother for both $A$ and $B$. Thus both $M^{(2)}$ and $M^{(3)}$ support efficient computation of $\widehat{\mathrm{cov}}(\hat\beta_{\mathrm{GLS}})$. A subtle point is that these symmetric smoothers are matrices in $\mathbb{R}^{N\times N}$ while the matrices $M^{(k)} \in \mathbb{R}^{C\times C}$ are not symmetric.

### 2.4.2 Model for $Z_{ij}$

We will state conditions on $Z_{ij}$ under which both $\|M^{(1)}\|_1$ and $\|M^{(2)}\|_1$ are bounded below 1 with probability tending to one, as the problem size grows. We need the following exponential inequalities.

*Lemma* 2.4.1. If $X \sim \mathrm{Bin}(n,p)$, then for any $t \geqslant 0$,

$$\begin{aligned}\Pr(X \geqslant np + t) &\leqslant \exp(-2t^2/n), \quad \text{and}\\ \Pr(X \leqslant np - t) &\leqslant \exp(-2t^2/n).\end{aligned}$$

*Proof.* This follows from Hoeffding's theorem. □

*Lemma* 2.4.2. Let $X_i \sim \mathrm{Bin}(n,p)$ for $i = 1,\dots,m$, not necessarily independent. Then for any $t \geqslant 0$,

$$\begin{aligned}\Pr\Bigl(\max_{1\leqslant i\leqslant m} X_i \geqslant np + t\Bigr) &\leqslant m\exp(-2t^2/n), \quad \text{and}\\ \Pr\Bigl(\min_{1\leqslant i\leqslant m} X_i \leqslant np - t\Bigr) &\leqslant m\exp(-2t^2/n).\end{aligned}$$

*Proof.* This is from the union bound applied to Lemma 4.3.1. □

Here is our sampling model. We index the size of our problem by $S \to \infty$. The sample size $N$ will satisfy $\mathbb{E}(N) \geqslant S$. The number of rows and columns in the data set are

$$R = S^{\rho} \quad \text{and} \quad C = S^{\kappa}$$

respectively, for positive numbers $\rho$ and $\kappa$. Because our application domain has $N \ll RC$, we assume that $\rho + \kappa > 1$. We ignore that $R$ and $C$ above are not necessarily integers.

In our model, $Z_{ij} \sim \mathrm{Bern}(p_{ij})$ independently with

$$\frac{S}{RC} \leqslant p_{ij} \leqslant \Upsilon \frac{S}{RC} \quad \text{for} \quad 1 \leqslant \Upsilon < \infty. \tag{2.31}$$

That is $1 \leqslant p_{ij} S^{\rho+\kappa-1} \leqslant \Upsilon$. Letting $p_{ij}$ depend on $i$ and $j$ allows the probability model to capture stylistic preferences affecting the missingness pattern in the ratings data.

### 2.4.3 Bounds for row and column size

Letting $X \preccurlyeq Y$ mean that $X$ is stochastically smaller than $Y$, we know that

$$\begin{aligned}
\mathrm{Bin}(R, S^{1-\rho-\kappa}) &\preccurlyeq N_{\bullet j} \preccurlyeq \mathrm{Bin}(R, \Upsilon S^{1-\rho-\kappa}), \quad \text{and} \\
\mathrm{Bin}(C, S^{1-\rho-\kappa}) &\preccurlyeq N_{i\bullet} \preccurlyeq \mathrm{Bin}(C, \Upsilon S^{1-\rho-\kappa}).
\end{aligned}$$

By Lemma 4.3.1, if $t \geqslant 0$, then

$$\begin{aligned}
\Pr(N_{i\bullet} \geqslant S^{1-\rho}(\Upsilon + t)) &\leqslant \Pr\big(\mathrm{Bin}(C, \Upsilon S^{1-\rho-\kappa}) \geqslant S^{1-\rho}(\Upsilon + t)\big) \\
&\leqslant \exp(-2(S^{1-\rho}t)^2/C) \\
&= \exp(-2S^{2-\kappa-2\rho}t^2).
\end{aligned}$$

Therefore if $2\rho + \kappa < 2$, we find using using Lemma 2.4.2 that

$$\Pr\big(\max_i N_{i\bullet} \geqslant S^{1-\rho}(\Upsilon + \epsilon)\big) \leqslant S^{\rho} \exp(-2S^{2-\kappa-2\rho}\epsilon^2) \to 0$$

for any $\epsilon > 0$. Combining this with an analogous lower bound,

$$\lim_{S\to\infty} \Pr\bigl((1-\epsilon)S^{1-\rho} \leqslant \min_i N_{i\bullet} \leqslant \max_i N_{i\bullet} \leqslant (\Upsilon+\epsilon)S^{1-\rho}\bigr) = 1. \tag{2.32}$$

Likewise, if $\rho + 2\kappa < 2$, then for any $\epsilon > 0$,

$$\lim_{S\to\infty} \Pr\bigl((1-\epsilon)S^{1-\kappa} \leqslant \min_j N_{\bullet j} \leqslant \max_j N_{\bullet j} \leqslant (\Upsilon+\epsilon)S^{1-\kappa}\bigr) = 1. \tag{2.33}$$

### 2.4.4 Interval arithmetic

We will replace $N_{i\bullet}$ and other quantities by intervals that asymptotically contain them with probability one and then use interval arithmetic in order to streamline some of the steps in our proofs. For instance,

$$N_{i\bullet} \in [(1-\epsilon)S^{1-\rho}, (\Upsilon+\epsilon)S^{1-\rho}] = [1-\epsilon, \Upsilon+\epsilon] \times S^{1-\rho} = [1-\epsilon, \Upsilon+\epsilon] \times \frac{S}{R}$$

holds simultaneously for all $1 \leqslant i \leqslant R$ with probability tending to one as $S \to \infty$. In interval arithmetic,

$$[A,B] + [a,b] = [a+A, b+B] \quad \text{and} \quad [A,B] - [a,b] = [A-b, B-a].$$

If $0 < a \leqslant b < \infty$ and $0 < A \leqslant B < \infty$, then

$$[A,B] \times [a,b] = [Aa, Bb] \quad \text{and} \quad [A,B]/[a,b] = [A/b, B/a].$$

Similarly, if $a < 0 < b$ and $X \in [a,b]$, then $|X| \in [0, \max(|a|,|b|)]$. Our arithmetic operations on intervals yield new intervals guaranteed to contain the results obtained using any members of the original intervals. We do not necessarily use the smallest such interval.

### 2.4.5 Co-observation

Recall that the co-observation matrices are $Z^{\mathsf{T}}Z \in \{0,1\}^{C\times C}$ and $ZZ^{\mathsf{T}} \in \{0,1\}^{R\times R}$. If $s \neq j$, then

$$\mathrm{Bin}\Big(R, \frac{S^2}{R^2C^2}\Big) \preccurlyeq (Z^{\mathsf{T}}Z)_{sj} \preccurlyeq \mathrm{Bin}\Big(R, \frac{\Upsilon^2 S^2}{R^2C^2}\Big).$$

That is $\mathrm{Bin}(S^{\rho}, S^{2-2\rho-2\kappa}) \preccurlyeq (Z^{\mathsf{T}}Z)_{sj} \preccurlyeq \mathrm{Bin}(S^{\rho}, \Upsilon^2 S^{2-2\rho-2\kappa})$. For $t \geqslant 0$,

$$\begin{aligned}\Pr\Big(\max_s \max_{j\neq s}(Z^{\mathsf{T}}Z)_{sj} \geqslant (\Upsilon^2+t)S^{2-\rho-2\kappa}\Big) &\leqslant \frac{C^2}{2}\exp(-(tS^{2-\rho-2\kappa})^2/R)\\ &= \frac{C^2}{2}\exp(-t^2S^{4-3\rho-4\kappa}).\end{aligned}$$

If $3\rho+4\kappa<4$ then

$$\begin{aligned}&\Pr\Big(\max_s \max_{j\neq s}\,(Z^{\mathsf{T}}Z)_{sj} \geqslant (\Upsilon^2+\epsilon)S^{2-\rho-2\kappa}\Big) \to 0, \quad \text{and}\\ &\Pr\Big(\min_s \min_{j\neq s}\,(Z^{\mathsf{T}}Z)_{sj} \leqslant (1-\epsilon)S^{2-\rho-2\kappa}\Big) \to 0,\end{aligned}$$

for any $\epsilon > 0$.

### 2.4.6 Asymptotic bounds for $\|M\|_1$

Here we prove upper bounds for $\|M^{(k)}\|_1$ for $k=1,2$ of equations (2.29) and (2.30), respectively. The bounds depend on $\Upsilon$ and there are values of $\Upsilon>1$ for which these norms are bounded strictly below one, with probability tending to one.

*Theorem* 2.4.1. Let $Z_{ij}$ follow the model from Section 2.4.2 with $\rho,\kappa\in(0,1)$, that satisfy $\rho+\kappa>1$, $2\rho+\kappa<2$ and $3\rho+4\kappa<4$. Then for any $\epsilon>0$,

$$\Pr\big(\|M^{(1)}\|_1 \leqslant \Upsilon^2-\Upsilon^{-2}+\epsilon\big)\to 1, \quad \text{and} \tag{2.34}$$

$$\Pr\big(\|M^{(2)}\|_1 \leqslant \Upsilon^2-\Upsilon^{-2}+\epsilon\big)\to 1 \tag{2.35}$$

as $S\to\infty$.

*Proof.* Without loss of generality we assume that $\epsilon<1$. We begin with (2.35). Let

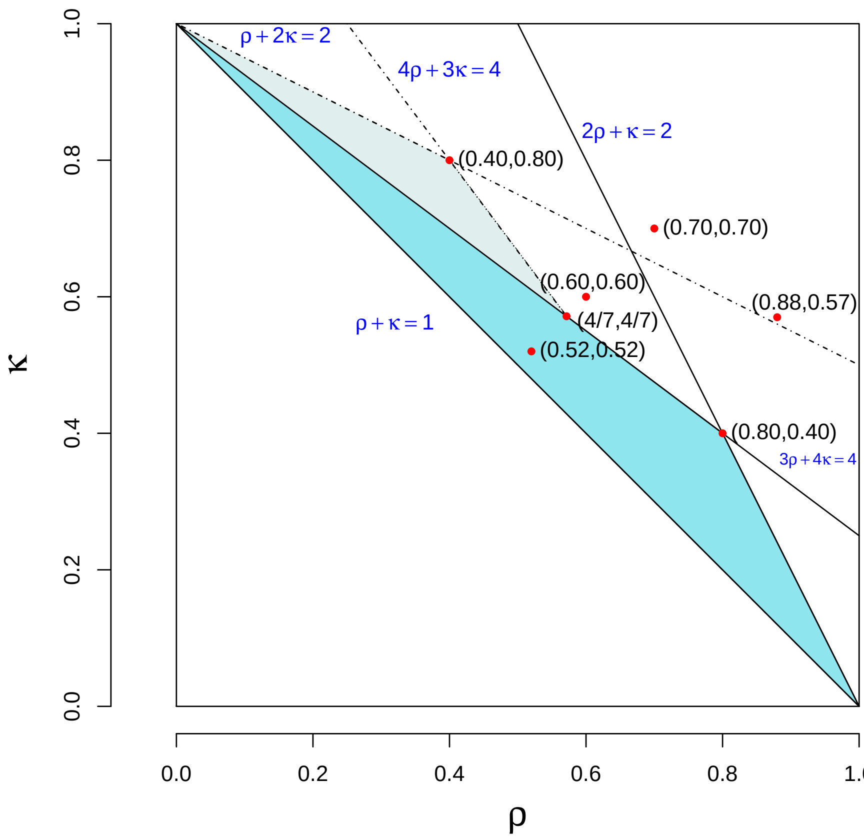



Figure 2.1: The large shaded triangle is the domain of interest $\mathcal{D}$ for Theorem 2.4.1. The smaller shaded triangle shows a region where the analogous update to $\boldsymbol{a}$ would have acceptable norm. The points marked are the ones we look at numerically, including $(0.88, 0.57)$ which corresponds to the Stitch Fix data in Section 2.6.

$M = M^{(2)}$. When $j \neq s$,

$$M_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r \frac{Z_{rj}}{N_{r\bullet} + \lambda_A}(Z_{rs} - \bar{Z}_{\bullet s}), \quad \text{for}$$

$$\bar{Z}_{\bullet s} = \sum_i \frac{Z_{is}}{N_{i\bullet} + \lambda_A} \Big/ \sum_i \frac{1}{N_{i\bullet} + \lambda_A}.$$

Although $|Z_{rs} - \bar{Z}_{\bullet s}| \leqslant 1$, replacing $Z_{rs} - \bar{Z}_{\bullet s}$ by one does not prove to be sharp enough for our purposes.

Every $N_{r\bullet} + \lambda_A \in S^{1-\rho}[1-\epsilon, \Upsilon+\epsilon]$ with probability tending to one and so

$$\begin{aligned}\frac{\bar{Z}_{\bullet s}}{N_{\bullet j}+\lambda_B}\sum_r \frac{Z_{rj}}{N_{r\bullet}+\lambda_A} &\in \frac{\bar{Z}_{\bullet s}}{N_{\bullet j}+\lambda_B}\sum_r \frac{Z_{rj}}{[1-\epsilon,\Upsilon+\epsilon]S^{1-\rho}}\\ &\subseteq [1-\epsilon,\Upsilon+\epsilon]^{-1}\bar{Z}_{\bullet s}S^{\rho-1}.\end{aligned}$$

Similarly

$$\begin{aligned}\bar{Z}_{\bullet s} &\in \frac{\sum_i Z_{is}[1-\epsilon,\Upsilon+\epsilon]^{-1}}{R[1-\epsilon,\Upsilon+\epsilon]^{-1}} \subseteq \frac{N_{\bullet s}}{R}[1-\epsilon,\Upsilon+\epsilon][1-\epsilon,\Upsilon+\epsilon]^{-1}\\ &\subseteq S^{1-\rho-\kappa}[1-\epsilon,\Upsilon+\epsilon]^2[1-\epsilon,\Upsilon+\epsilon]^{-1}\end{aligned}$$

and so

$$\frac{\bar{Z}_{\bullet s}}{N_{\bullet j}+\lambda_B}\sum_r \frac{Z_{rj}}{N_{r\bullet}+\lambda_A} \in S^{-\kappa}\frac{[1-\epsilon,\Upsilon+\epsilon]^2}{[1-\epsilon,\Upsilon+\epsilon]^2} \subseteq \frac{1}{C}\Big[\Big(\frac{1-\epsilon}{\Upsilon+\epsilon}\Big)^2, \Big(\frac{\Upsilon+\epsilon}{1-\epsilon}\Big)^2\Big]. \tag{2.36}$$

Next using bounds on the co-observation counts,

$$\frac{1}{N_{\bullet j}+\lambda_B}\sum_r \frac{Z_{rj}Z_{rs}}{N_{r\bullet}+\lambda_A} \in \frac{S^{\rho+\kappa-2}(Z^\mathsf{T}Z)_{sj}}{[1-\epsilon,\Upsilon+\epsilon]^2} \subseteq \frac{1}{C}\frac{[1-\epsilon,\Upsilon^2+\epsilon]}{[1-\epsilon,\Upsilon+\epsilon]^2}. \tag{2.37}$$

Combining (2.36) and (2.37)

$$M_{js} \in \frac{1}{C}\Big[\frac{1-\epsilon}{(\Upsilon+\epsilon)^2} - \Big(\frac{\Upsilon+\epsilon}{1-\epsilon}\Big)^2, \frac{\Upsilon^2+\epsilon}{1-\epsilon} - \Big(\frac{1-\epsilon}{\Upsilon+\epsilon}\Big)^2\Big]$$

For any $\epsilon' > 0$ we can choose $\epsilon$ small enough that

$$M_{js} \in C^{-1}[\Upsilon^{-2} - \Upsilon^2 - \epsilon', \Upsilon^2 - \Upsilon^{-2} + \epsilon']$$

and then $|M_{js}| \leqslant (\Upsilon^2 - \Upsilon^{-2} + \epsilon')/C$.

Next, arguments like the preceding give $|M_{jj}| \leqslant (1-\epsilon')^{-2}(\Upsilon+\epsilon')S^{\rho-1} \to 0$. Then with probability tending to one,

$$\sum_j |M_{js}| \leqslant \Upsilon^2 - \Upsilon^{-2} + 2\epsilon'.$$

This bound holds for all $s \in \{1, 2, \ldots, C\}$, establishing (2.35).

The proof of (2.34) is similar. The quantity $\bar{Z}_{\bullet s}$ is replaced by $(1/R)\sum_i Z_{is}/(N_{i\bullet} + \lambda_A)$. $\square$

It is interesting to find the largest $\Upsilon$ with $\Upsilon^2 - \Upsilon^{-2} \leqslant 1$. It is $((1+5^{1/2})/2)^{1/2} \doteq 1.27$.

## 2.5 Convergence and computation

In this section we make some computations on synthetic data following the probability model from Section 2.4. First we study the norms of our update matrix $M^{(2)}$ which affects the number of iterations to convergence. In addition to $\|\cdot\|_1$ covered in Theorem 2.4.1 we also consider $\|\cdot\|_2$, $\|\cdot\|_\infty$ and $\lambda_{\max}(\cdot)$. Then we compare the cost to compute $\hat{\beta}_{\mathrm{GLS}}$ by our backfitting method with that of lmer (Bates et al., 2015).

The problem size is indexed by $S$. Indices $i$ go from 1 to $R = \lceil S^\rho \rceil$ and indices $j$ go from 1 to $C = \lceil S^\kappa \rceil$. Reasonable parameter values have $\rho, \kappa \in (0, 1)$ with $\rho + \kappa > 1$. Theorem 2.4.1 applies when $2\rho + \kappa < 2$ and $3\rho + 4\kappa < 4$. Figure 4.1 depicts this triangular domain of interest $\mathcal{D}$. There is another triangle $\mathcal{D}'$ where a corresponding update for $\boldsymbol{a}$ would satisfy the conditions of Theorem 2.4.1. Then $\mathcal{D} \cup \mathcal{D}'$ is a non-convex polygon of five sides. Figure 4.1 also shows $\mathcal{D}' \setminus \mathcal{D}$ as a second triangular region. For points $(\rho, \kappa)$ near the line $\rho + \kappa = 1$, the matrix $Z$ will be mostly ones unless $S$ is very large. For points $(\rho, \kappa)$ near the upper corner $(1, 1)$, the matrix $Z$ will be extremely sparse with each $N_{i\bullet}$ and $N_{\bullet j}$ having nearly a Poisson distribution with mean between 1 and $\Upsilon$. The fraction of potential values that have been observed is $O(S^{1-\rho-\kappa})$.

Given $p_{ij}$, we generate our observation matrix via $Z_{ij} \overset{\text{ind}}{\sim} \mathrm{Bern}(p_{ij})$. These probabilities are first generated via $p_{ij} = U_{ij} S^{1-\rho-\kappa}$ where $U_{ij} \overset{\text{iid}}{\sim} \mathbb{U}[1, \Upsilon]$ and $\Upsilon$ is the largest value for which $\Upsilon^2 - \Upsilon^{-2} \leqslant 1$. For small $S$ and $\rho + \kappa$ near 1 we can get some values $p_{ij} > 1$ and in that case we take $p_{ij} = 1$.

The following $(\rho, \kappa)$ combinations are of interest. First, $(4/5, 2/5)$ is the closest vertex of the domain of interest to the point $(1, 1)$. Second, $(2/5, 4/5)$ is outside the domain of interest for the $\boldsymbol{b}$ but within the domain for the analogous $\boldsymbol{a}$ update. Third, among points with $\rho = \kappa$, the value $(4/7, 4/7)$ is the farthest one from the origin that is in the domain of interest. We also look at some points on the 45 degree line that are outside the domain of interest because the sufficient conditions in Theorem 2.4.1 might not be necessary.

In our matrix norm computations we took $\lambda_A = \lambda_B = 0$. This completely removes shrinkage and will make it harder for the algorithm to converge than would be the case for the positive $\lambda_A$ and $\lambda_B$ that hold in real data. The values of $\lambda_A$ and $\lambda_B$ appear in expressions $N_{i\bullet} + \lambda_A$ and $N_{\bullet j} + \lambda_B$ where their contribution is asymptotically negligible, so conservatively setting them to zero will nonetheless be realistic for large data sets.

We sample from the model multiple times at various values of $S$ and plot $\|M^{(2)}\|_1$ versus $S$ on a logarithmic scale. Figure 2.2 shows the results. We observe that $\|M^{(2)}\|_1$ is below 1 and decreasing with $S$ for all the examples $(\rho, \kappa) \in \mathcal{D}$. This holds also for $(\rho, \kappa) = (0.60, 0.60) \notin \mathcal{D}$. We chose that point because it is on the convex hull of $\mathcal{D} \cup \mathcal{D}'$.

The point $(\rho, \kappa) = (0.40, 0.80) \notin \mathcal{D}$. Figure 2.2 shows large values of $\|M^{(2)}\|_1$ for this case. Those values increase with $S$, but remain below 1 in the range considered. This is a case where the update from $\boldsymbol{a}$ to $\boldsymbol{a}$ would have norm well below 1 and decreasing with $S$, so backfitting would converge. We do not know whether $\|M^{(2)}\|_1 > 1$ will occur for larger $S$.

The point $(\rho, \kappa) = (0.70, 0.70)$ is not in the domain $\mathcal{D}$ covered by Theorem 2.4.1 and we see that $\|M^{(2)}\|_1 > 1$ and generally increasing with $S$ as shown in Figure 2.3. This does not mean that backfitting must fail to converge. Here we find that $\|M^{(2)}\|_2 < 1$ and generally decreases as $S$ increases. This is a strong indication that the number of backfitting iterations required will not grow with $S$ for this $(\rho, \kappa)$ combination. We cannot tell whether $\|M^{(2)}\|_2$ will decrease to zero but that is what appears to happen.

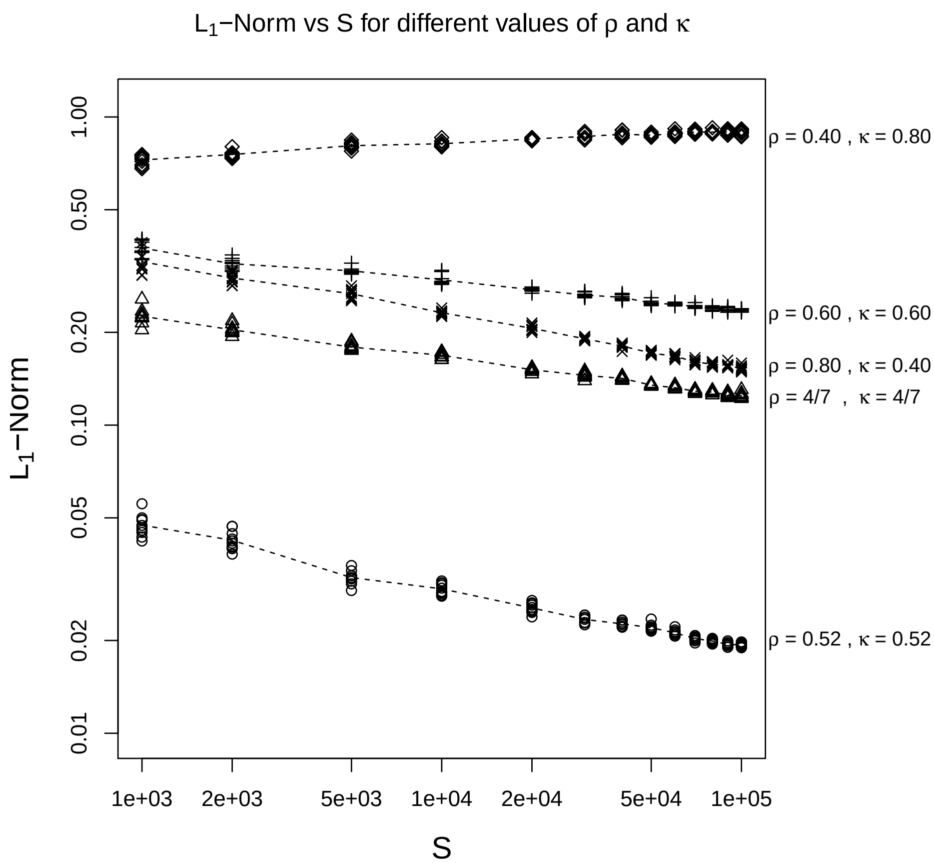



Figure 2.2: Norm $\|M^{(2)}\|_1$ of centered update matrix versus problem size $S$ for different $(\rho, \kappa)$.

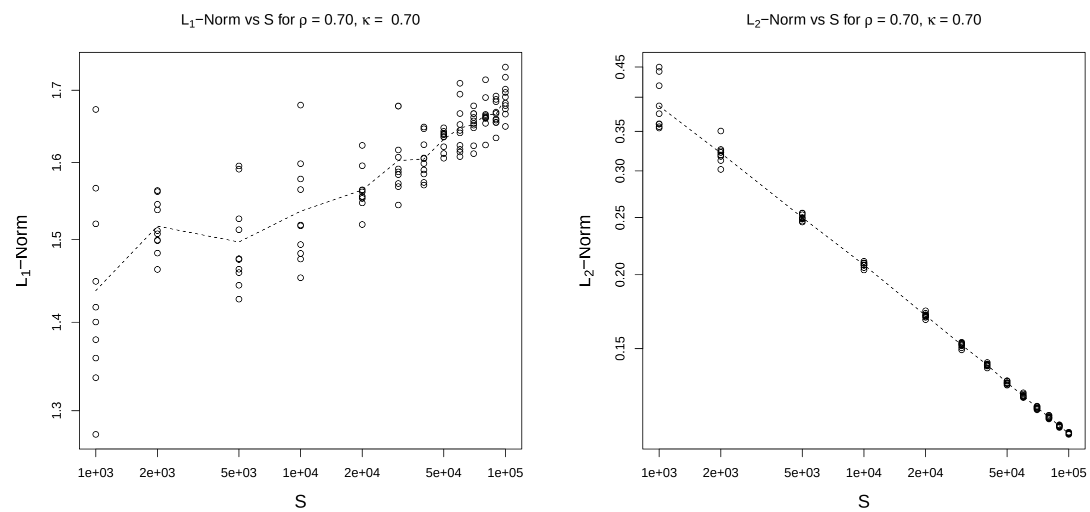



Figure 2.3: The left panel shows $\|M^{(2)}\|_1$ versus $S$. The right panel shows $\|M^{(2)}\|_2$ versus $S$ with a logarithmic vertical scale. Both have $(\rho, \kappa) = (0.7, 0.7)$.

We consistently find in our computations that $\lambda_{\max}(M^{(2)}) \leqslant \|M^{(2)}\|_2 \leqslant \|M^{(2)}\|_1$. The first of these inequalities must necessarily hold. For a symmetric matrix $M$ we know that $\lambda_{\max}(M) = \|M\|_2$ which is then necessarily no larger than $\|M\|_1$. Our update matrices are nearly symmetric but not perfectly so. We believe that explains why their $L_2$ norms are close to their spectral radius and also smaller than their $L_1$ norms. While the $L_2$ norms are empirically more favorable than the $L_1$ norms, they are not amenable to our theoretical treatment.

We believe that backfitting will have a spectral radius well below 1 for more cases than we can as yet prove. In addition to the previous figures showing matrix norms as $S$ increases for certain special values of $(\rho, \kappa)$ we have computed contour maps of those norms over $(\rho, \kappa) \in [0, 1]$ for $S = 10{,}000$. See Figure 2.4.

To compare the computation times for algorithms we generated $Z_{ij}$ as above and also took $x_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I_7)$ plus an intercept, making $p = 8$ fixed effect parameters. Although backfitting can run with $\lambda_A = \lambda_B = 0$, lmer cannot do so for numerical reasons. So we took $\sigma_A^2 = \sigma_B^2 = 1$ and $\sigma_E^2 = 1$ corresponding to $\lambda_A = \lambda_B = 1$. The cost per iteration does not depend on $Y_{ij}$ and hence not on $\beta$ either. We used $\beta = 0$.

Figure 2.5 shows computation times for one single iteration when $(\rho, \kappa) = (0.52, 0.52)$

and when $(\rho, \kappa) = (0.70, 0.70)$. The time to do one iteration in lmer grows roughly like $N^{3/2}$ in the first case. For the second case, it appears to grow at the even faster rate of $N^{2.1}$. Solving a system of $S^\kappa \times S^\kappa$ equations would cost $S^{3\kappa} = S^{2.1} = O(N^{2.1})$, which explains the observed rate. This analysis would predict $O(N^{1.56})$ for $\rho = \kappa = 0.52$ but that is only minimally different from $O(N^{3/2})$. These experiments were carried out in R on a computer with the macOS operating system, 16 GB of memory and an Intel i7 processor. Each backfitting iteration entails solving (2.17) along with the fixed effects.

The cost per iteration for backfitting follows closely to the $O(N)$ rate predicted by the theory. OLS only takes one iteration and it is also of $O(N)$ cost. In both of these cases $\|M^{(2)}\|_2$ is bounded away from one so the number of backfitting iterations does not grow with $S$. For $\rho = \kappa = 0.52$, backfitting took 4 iterations to converge for the smaller values of $S$ and 3 iterations for the larger ones. For $\rho = \kappa = 0.70$, backfitting took 6 iterations for smaller $S$ and 4 or 5 iterations for larger $S$. In each case our convergence criterion was a relative change of $10^{-8}$ as described in Section 2.6.3. Further backfitting to compute BLUPs $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ given $\hat{\beta}_{\mathrm{GLS}}$ took at most 5 iterations for $\rho = \kappa = 0.52$ and at most 10 iterations for $\rho = \kappa = 0.7$. In the second example, lme4 did not reach convergence in our time window so we ran it for just 4 iterations to measure its cost per iteration.

## 2.6 Example: ratings from Stitch Fix

We illustrate backfitting for GLS on some data from Stitch Fix. Stitch Fix sells clothing. They mail their customers a sample of items. The customers may keep and purchase any of those items that they want, while returning the others. It is valuable to predict the extent to which a customer will like an item, not just whether they will purchase it. Stitch Fix has provided us with some of their client ratings data. It was anonymized, void of personally identifying information, and as a sample it does not reflect their total numbers of clients or items at the time they provided it. It is also from 2015. While it does not describe their current business, it is a valuable data set for illustrative purposes.

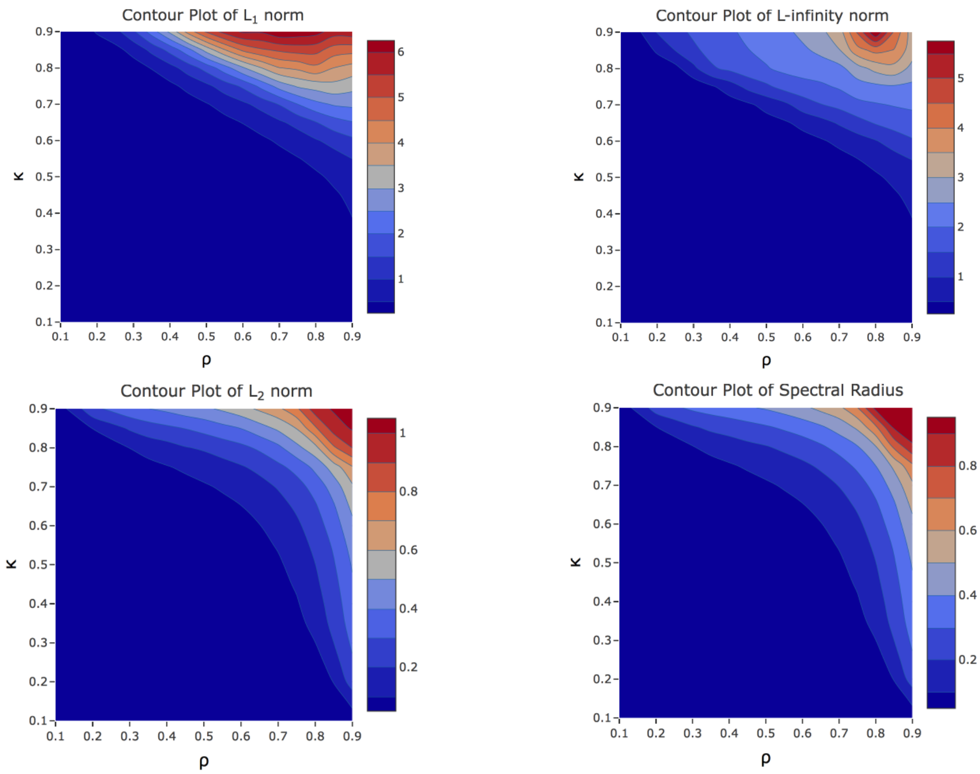
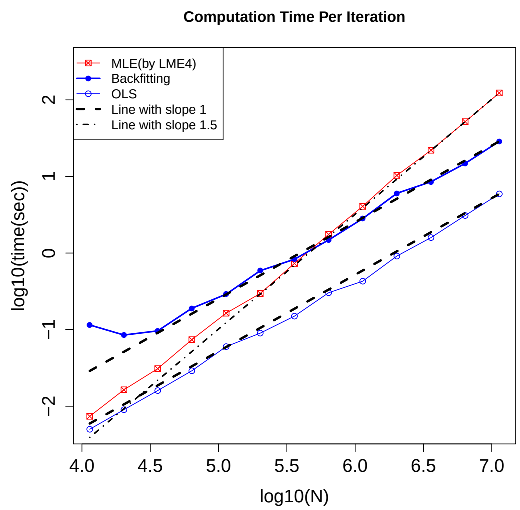



Figure 2.4: Numerically computed matrix norms for $M^{(2)}$ using $S = 10{,}000$. The color code varies with the subfigures.

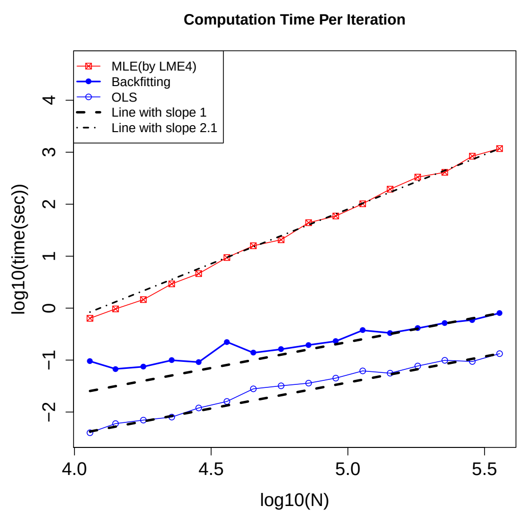
The sample sizes for this data are as follows. We received $N = 5{,}000{,}000$ ratings by $R = 762{,}752$ customers on $C = 6{,}318$ items. These values of $R$ and $C$ correspond to the point $(0.88, 0.57)$ in Figure 4.1. Thus $C/N \doteq 0.00126$ and $R/N \doteq 0.153$. The data are not dominated by a single row or column because $\max_i N_{i\bullet}/N \doteq 9 \times 10^{-6}$ and $\max_j N_{\bullet j}/N \doteq 0.0143$. The data are sparse because $N/(RC) \doteq 0.001$.

## 2.6.1 An illustrative linear model

The response $Y_{ij}$ is a rating on a ten point scale of the satisfaction of customer $i$ with item $j$. The data come with features about the clients and items. In a business setting one would fit and compare possibly dozens of different regression models to understand the data. Our purpose here is to study large scale GLS and compare it to ordinary least squares (OLS) and so we use just one model, not necessarily one that




(a) $(\rho, \kappa) = (0.52, 0.52)$




(b) $(\rho, \kappa) = (0.70, 0.70)$

Figure 2.5: Time for one iteration versus the number of observations, $N$ at two points $(\rho, \kappa)$. The cost for lmer is roughly $O(N^{3/2})$ in the top panel and $O(N^{2.1})$ in the bottom panel. The costs for OLS and backfitting are $O(N)$.

we would have settled on. For that purpose we use the same model that was used in Gao and Owen (2020). It is not chosen to make OLS look as bad as possible. Instead it is potentially the first model one might look at in a data analysis. For client $i$ and item $j$,

$$\begin{aligned}Y_{ij} = \beta_0 &+ \beta_1 \text{match}_{ij} + \beta_2 \mathbb{I}\{\text{client edgy}\}_i + \beta_3 \mathbb{I}\{\text{item edgy}\}_j \\ &+ \beta_4 \mathbb{I}\{\text{client edgy}\}_i * \mathbb{I}\{\text{item edgy}\}_j + \beta_5 \mathbb{I}\{\text{client boho}\}_i \\ &+ \beta_6 \mathbb{I}\{\text{item boho}\}_j + \beta_7 \mathbb{I}\{\text{client boho}\}_i * \mathbb{I}\{\text{item boho}\}_j \\ &+ \beta_8 \text{material}_j + a_i + b_j + e_{ij}.\end{aligned}$$

Here $\text{material}_j$ is a categorical variable that is implemented via indicator variables for each type of material other than the baseline. Following Gao and Owen (2020), we chose 'Polyester', the most common material, as the baseline. Some customers and some items were given the adjective 'edgy' in the data set. Another adjective was 'boho', short for 'Bohemian'. The variable $\text{match}_{ij} \in [0, 1]$ is an estimate of the probability that the customer keeps the item, made before the item was sent. The match score is a prediction from a baseline model and is not representative of all algorithms used at Stitch Fix. All told, the model has $p = 30$ parameters.

### 2.6.2 Estimating the variance parameters

We use the method of moments method from Gao and Owen (2020) to estimate $\theta^\mathsf{T} = (\sigma_A^2, \sigma_B^2, \sigma_E^2)$ in $O(N)$ computation. That is in turn based on the method that Gao and Owen (2017) use in the intercept only model where $Y_{ij} = \mu + a_i + b_j + e_{ij}$. For that model they set

$$\begin{aligned}U_A &= \sum_i \sum_j Z_{ij}\Bigl(Y_{ij} - \frac{1}{N_{i\bullet}} \sum_{j'} Z_{ij'} Y_{ij'}\Bigr)^2, \\ U_B &= \sum_j \sum_i Z_{ij}\Bigl(Y_{ij} - \frac{1}{N_{\bullet j}} \sum_{i'} Z_{i'j} Y_{i'j}\Bigr)^2, \quad \text{and} \\ U_E &= N \sum_{ij} Z_{ij}\Bigl(Y_{ij} - \frac{1}{N} \sum_{i'j'} Z_{i'j'} Y_{i'j'}\Bigr)^2.\end{aligned}$$

These are, respectively, sums of within row sums of squares, sums of within column sums of squares and a scaled overall sum of squares. Straightforward calculations show that

$$\begin{aligned}
\mathbb{E}(U_A) &= \bigl(\sigma_B^2 + \sigma_E^2\bigr)(N - R),\\
\mathbb{E}(U_B) &= \bigl(\sigma_A^2 + \sigma_E^2\bigr)(N - C), \quad \text{and}\\
\mathbb{E}(U_E) &= \sigma_A^2\Bigl(N^2 - \sum_i N_{i\bullet}^2\Bigr) + \sigma_B^2\Bigl(N^2 - \sum_j N_{\bullet j}^2\Bigr) + \sigma_E^2(N^2 - N).
\end{aligned}$$

By matching moments, we can estimate $\theta$ by solving the $3 \times 3$ linear system

$$\begin{pmatrix} 0 & N-R & N-R \\ N-C & 0 & N-C \\ N^2 - \Sigma N_i^2 & N^2 - \Sigma N_j^2 & N^2 - N \end{pmatrix} \begin{pmatrix} \sigma_A^2 \\ \sigma_B^2 \\ \sigma_E^2 \end{pmatrix} = \begin{pmatrix} U_A \\ U_B \\ U_E \end{pmatrix}$$

for $\theta$.

Following Gao and Owen (2017) we note that $\eta_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\beta = a_i + b_j + e_{ij}$ has the same parameter $\theta$ as $Y_{ij}$ have. We then take a consistent estimate of $\beta$, in this case $\hat\beta_{\mathrm{OLS}}$ that Gao and Owen (2017) show is consistent for $\beta$, and define $\hat\eta_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat\beta_{\mathrm{OLS}}$. We then estimate $\theta$ by the above method after replacing $Y_{ij}$ by $\hat\eta_{ij}$. For the Stitch Fix data we obtained $\hat\sigma_A^2 = 1.14$ (customers), $\hat\sigma_B^2 = 0.11$ (items) and $\hat\sigma_E^2 = 4.47$.

### 2.6.3 Computing $\hat\beta_{\mathrm{GLS}}$

The estimated coefficients $\hat\beta_{\mathrm{GLS}}$ and their standard errors are presented in a table in the appendix A. Open-source R code at `https://github.com/G28Sw/backfit_code` does these computations. Here is a concise description of the algorithm we used:

1) Compute $\hat\beta_{\mathrm{OLS}}$ via (2.2).
2) Get residuals $\hat\eta_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat\beta_{\mathrm{OLS}}$.
3) Compute $\hat\sigma_A^2$, $\hat\sigma_B^2$ and $\hat\sigma_E^2$ by the method of moments on $\hat\eta_{ij}$.
4) Compute $\widetilde{\mathcal{X}} = (I_N - \widetilde{\mathcal{S}}_G)\mathcal{X}$ using doubly centered backfitting $M^{(3)}$.
5) Compute $\hat\beta_{\mathrm{GLS}}$ by (2.28).

6) If we want BLUPs $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ backfit $\mathcal{Y} - \mathcal{X}\hat{\beta}_{\mathrm{GLS}}$ to get them.
7) Compute $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$ by plugging $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$ and $\hat{\sigma}^2_E$ into $\mathcal{V}$ at (2.28).

Stage $k$ of backfitting provides $(\tilde{\mathcal{S}}_G\mathcal{X})^{(k)}$. We iterate until

$$\frac{\|(\tilde{\mathcal{S}}_G\mathcal{X})^{(k+1)} - (\tilde{\mathcal{S}}_G\mathcal{X})^{(k)}\|_F^2}{\|(\tilde{\mathcal{S}}_G\mathcal{X})^{(k)}\|_F^2} < \epsilon$$

where $\|\cdot\|_F$ is the Frobenius norm (root mean square of all elements). Our numerical results use $\epsilon = 10^{-8}$.

When we want $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$ then we need to use a backfitting strategy with a symmetric smoother $\tilde{\mathcal{S}}_G$. This holds for $M^{(0)}$, $M^{(2)}$ and $M^{(3)}$ but not $M^{(1)}$. After computing $\hat{\beta}_{\mathrm{GLS}}$ one can return to step 2, form new residuals $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}_{\mathrm{GLS}}$ and continue through steps 3–7. We have seen small differences from doing this.

### 2.6.4 Quantifying inefficiency and naivete of OLS

In the introduction we mentioned two serious problems with the use of OLS on crossed random effects data. The first is that OLS is naive about correlations in the data and this can lead it to severely underestimate the variance of $\hat{\beta}$. The second is that OLS is inefficient compared to GLS by the Gauss-Markov theorem. Let $\hat{\beta}_{\mathrm{OLS}}$ and $\hat{\beta}_{\mathrm{GLS}}$ be the OLS and GLS estimates of $\beta$, respectively. We can compute their corresponding variance estimates $\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})$ and $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS}})$. We can also find $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})$, the variance under our GLS model of the linear combination of $Y_{ij}$ values that OLS uses. This section explore them graphically.

We can quantify the naivete of OLS via the ratios $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})/\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})$ for $j = 1, \ldots, p$. Figure 2.6 plots these values. They range from 1.75 to 345.28 and can be interpreted as factors by which OLS naively overestimates its sample size. The largest and second largest ratios are for material indicators corresponding to 'Modal' and 'Tencel', respectively. These appear to be two names for the same product with Tencel being a trademarked name for Modal fibers (made from wood). We can also identify the linear combination of $\hat{\beta}_{\mathrm{OLS}}$ for which OLS is most naive. We maximize the ratio $x^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})x/x^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})x$ over $x \neq 0$. The resulting maximal ratio

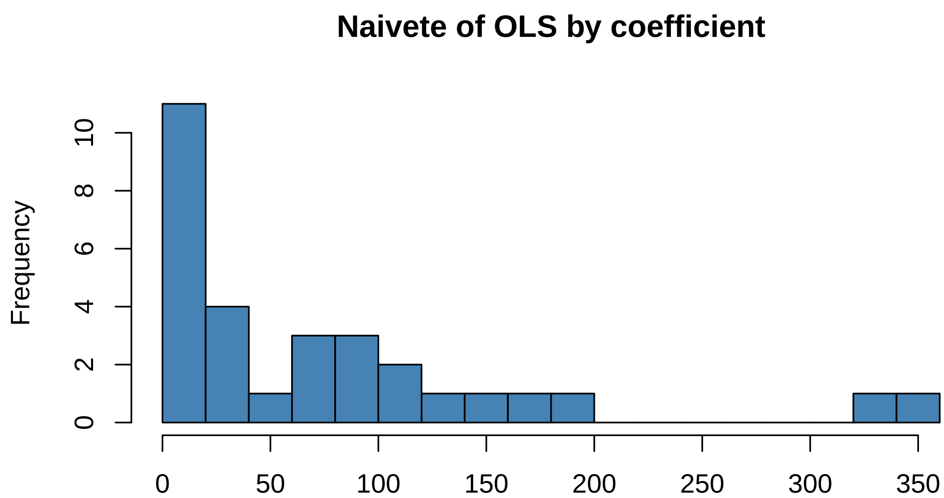



Figure 2.6: OLS naivete $\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{OLS},j})/\widehat{\text{cov}}_{\text{OLS}}(\hat{\beta}_{\text{OLS},j})$ for coefficients $\beta_j$ in the Stitch Fix data.

is the largest eigenvalue of

$$\widehat{\text{cov}}_{\text{OLS}}(\hat{\beta}_{\text{OLS}})^{-1}\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{OLS}})$$

and it is about 361 for the Stitch Fix data.

We can quantify the inefficiency of OLS via the ratio $\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{OLS},j})/\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{GLS},j})$ for $j = 1, \ldots, p$. Figure 2.7 plots these values. They range from just over 1 to 50.6 and can be interpreted as factors by which using OLS reduces the effective sample size. There is a clear outlier: the coefficient of the match variable is very inefficiently estimated by OLS. The second largest inefficiency factor is for the intercept term. The most inefficient linear combination of $\hat{\beta}$ reaches a variance ratio of 52.6, only slightly more inefficient than the match coefficient alone.

The variables for which OLS is more naive tend to also be the variables for which it is most inefficient. Figure 3.5 plots these quantities against each other for the 30 coefficients in our model.

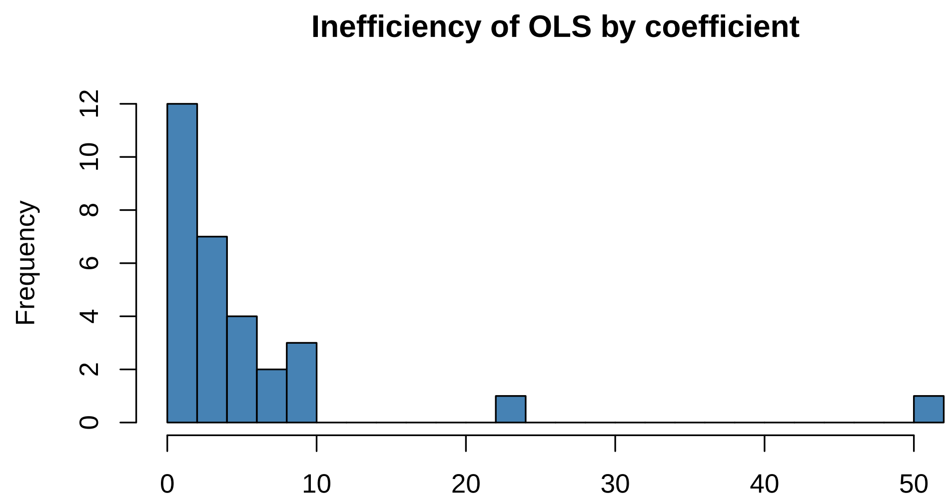



Figure 2.7: OLS inefficiency $\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{OLS},j})/\widehat{\text{cov}}_{\text{GLS}}(\hat{\beta}_{\text{GLS},j})$ for coefficients $\beta_j$ in the Stitch Fix data.

### 2.6.5 Convergence speed of backfitting

The Stitch Fix data have row and column sample sizes that are much more uneven than our sampling model for $Z$ allows. Accordingly we cannot rely on Theorem 2.4.1 to show that backfitting must converge rapidly for it.

The sufficient conditions in that theorem may not be necessary and we can compute our norms and the spectral radius on the update matrices for the Stitch Fix data using some sparse matrix computations. Here $Z \in \{0,1\}^{762{,}752\times 6318}$, so $M^{(k)} \in \mathbb{R}^{6318\times 6318}$ for $k \in \{0,1,2,3\}$. The results are

$$\begin{pmatrix} \|M^{(0)}\|_1 & \|M^{(0)}\|_2 & |\lambda_{\max}(M^{(0)})| \\ \|M^{(1)}\|_1 & \|M^{(1)}\|_2 & |\lambda_{\max}(M^{(1)})| \\ \|M^{(2)}\|_1 & \|M^{(2)}\|_2 & |\lambda_{\max}(M^{(2)})| \\ \|M^{(3)}\|_1 & \|M^{(3)}\|_2 & |\lambda_{\max}(M^{(3)})| \end{pmatrix} = \begin{pmatrix} 31.9525 & 1.4051 & 0.64027 \\ 11.2191 & 0.4512 & 0.33386 \\ 8.9178 & 0.4541 & 0.33407 \\ 9.2143 & 0.4546 & 0.33377 \end{pmatrix}.$$

All the updates have spectral radius comfortably below one. The centered updates

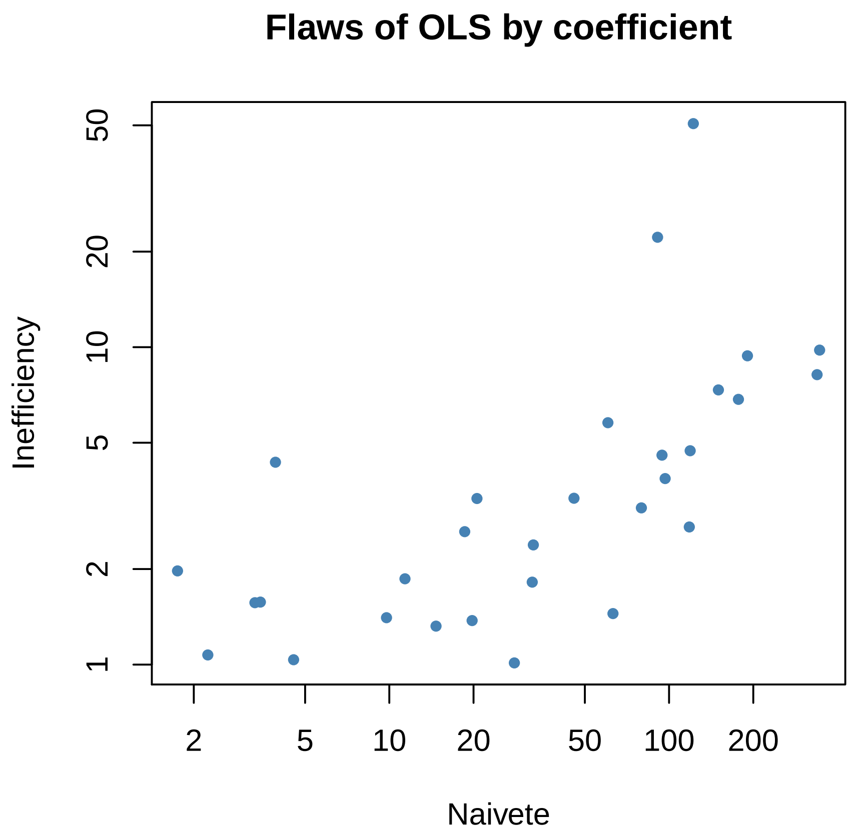



Figure 2.8: Inefficiency vs naivete for OLS coefficients in the Stitch Fix data.

have $L_2$ norm below one but the uncentered update does not. Their $L_2$ norms are somewhat larger than their spectral radii because those matrices are not quite symmetric. The two largest eigenvalue moduli for $M^{(0)}$ are 0.6403 and 0.3337 and the centered updates have spectral radii close to the second largest eigenvalue of $M^{(0)}$. This is consistent with an intuitive explanation that the space spanned by a column of $N$ ones that is common to the columns spaces of $\mathcal{Z}_A$ and $\mathcal{Z}_B$ is the biggest impediment to $M^{(0)}$ and that all three centering strategies essentially remove it. The best spectral radius is for $M^{(3)}$, which employs two principled centerings, although in this data set it made little difference. Our backfitting algorithm took 8 iterations when applied to $\mathcal{X}$ and 12 more to compute the BLUPs. We used a convergence threshold of $10^{-8}$.

## 2.7 Discussion

We have shown that the cost of our backfitting algorithm is $O(N)$ under strict conditions that are nonetheless much more general than having $N_{i\bullet} = N/C$ for all $i = 1, \dots, R$ and $N_{\bullet j} = N/R$ for all $j = 1, \dots, C$ as in Papaspiliopoulos et al. (2020). As in their setting, the backfitting algorithm scales empirically to much more general problems than those for which rapid convergence can be proved. Our contour map of the spectral radius of the update matrix $M$ shows that this norm is well below 1 over many more $(\rho, \kappa)$ pairs that our theorem covers. The difficulty in extending our approach to those settings is that the spectral radius is a much more complicated function of the observation matrix $Z$ than the $L_1$ norm is.

Theorem 4 of Papaspiliopoulos et al. (2020) has the rate of convergence for their collapsed Gibbs sampler for balanced data. It involves an auxiliary convergence rate $\rho_{\text{aux}}$ defined as follows. Consider the Gibbs sampler on $(i, j)$ pairs where given $i$ a random $j$ is chosen with probability $Z_{ij}/N_{i\bullet}$ and given $j$ a random $i$ is chosen with probability $Z_{ij}/N_{\bullet j}$. That Markov chain has invariant distribution $Z_{ij}/N$ on $(i, j)$ pairs and $\rho_{\text{aux}}$ is the rate at which the chain converges. In our notation

$$\rho_{\text{PRZ}} = \frac{N\sigma_A^2}{N\sigma_A^2 + R\sigma_E^2} \times \frac{N\sigma_B^2}{N\sigma_B^2 + C\sigma_E^2} \times \rho_{\text{aux}}.$$

In sparse data $\rho_{\text{PRZ}} \approx \rho_{\text{aux}}$ and under our asymptotic setting $|\rho_{\text{aux}} - \rho_{\text{PRZ}}| \to 0$. Papaspiliopoulos et al. (2020) remark that $\rho_{\text{aux}}$ tends to decrease as the amount of data increases. When it does, then their algorithm takes $O(1)$ iterations and costs $O(N)$. They explain that $\rho_{\text{aux}}$ should decrease as the data set grows because the auxiliary process then gets greater connectivity. That connectivity increases for bounded $R$ and $C$ with increasing $N$ and from their notation, allowing multiple observations per $(i, j)$ pair it seems like they have this sort of infill asymptote in mind. For sparse data from electronic commerce we think that an asymptote like the one we study where $R$, $C$ and $N$ all grow is a better description. It would be interesting to see how $\rho_{\text{aux}}$

develops under such a model.

In Section 5.3 Papaspiliopoulos et al. (2020) state that the convergence rate of the collapsed Gibbs sampler is $O(1)$ regardless of the asymptotic regime. That section is about a more stringent 'balanced cells' condition where every $(i, j)$ combination is observed the same number of times, so it does not describe the 'balanced levels' setting where $N_{i\bullet} = N/R$ and $N_{\bullet j} = N/C$. Indeed they provide a counterexample in which there are two disjoint communities of users and two disjoint sets of items and each user in the first community has rated every item in the first item set (and no others) while each user in the second community has rated every item in the second item set (and no others). That configuration leads to an unbounded mixing time for collapsed Gibbs. It is also one where backfitting takes an increasing number of iterations as the sample size grows.

There are interesting parallels between methods to sample a high dimensional Gaussian distribution with covariance matrix $\Sigma$ and iterative solvers for the system $\Sigma\boldsymbol{x} = \boldsymbol{b}$. See Goodman and Sokal (1989) and Roberts and Sahu (1997) for more on how the convergence rates for these two problems coincide. We found that backfitting with one or both updates centered worked much better than uncentered backfitting. Papaspiliopoulos et al. (2020) used a collapsed sampler that analytically integrated out the global mean of their model in each update of a block of random effects.

Our approach treats $\sigma^2_A$, $\sigma^2_B$ and $\sigma^2_E$ as nuisance parameters. We plug in a consistent method of moments based estimator of them in order to focus on the backfitting iterations. In Bayesian computations, maximum a posteriori estimators of variance components under non-informative priors can be problematic for hierarchical models Gelman (2006), and so perhaps maximum likelihood estimation of these variance components would also have been challenging.

Whether one prefers a GLS estimate or a Bayesian one depends on context and goals. We believe that there is a strong computational advantage to GLS for large data sets. The cost of one backfitting iteration is comparable to the cost to generate one more sample in the MCMC. We may well find that only a dozen or so iterations are required for convergence of the GLS. A Bayesian analysis requires a much larger number of draws from the posterior distribution than that. For instance, Gelman and

Shirley (2011) recommend an effective sample size of about 100 posterior draws, with autocorrelations requiring a larger actual sample size. Vats et al. (2019) advocate even greater effective sample sizes.

It is usually reasonable to assume that there is a selection bias underlying which data points are observed. Accounting for any such selection bias must necessarily involve using information or assumptions from outside the data set at hand. We expect that any approach to take proper account of informative missingness must also make use of solutions to GLS perhaps after reweighting the observations. Before one develops any such methods, it is necessary to first be able to solve GLS without regard to missingness.

Many of the problems in electronic commerce involve categorical outcomes, especially binary ones, such as whether an item was purchased or not. Generalized linear mixed models are then appropriate ways to handle crossed random effects. In the next chapter we build scalable algorithm to do estimation and inference in such models.

# Chapter 3

# Generalized Linear Mixed Model with Crossed Random Effect

## 3.1 Introduction

Crossed random effects structures are ubiquitous in science, engineering and commerce. A biologist might study a response across many genotypes and environments (Bolker et al., 2009). A social scientist might study test scores by question and by student (Baayen et al., 2008). A web store might study responses of customers to various products. In all of these cases we have a pair of categorical variables with a potentially large number of levels. When those levels are a subset of a much larger set of potential levels then it is natural to model them as random effects. The problems we are interested in are large and sparse. There are $R$ (for rows) levels of one factor crossed with $C$ (for columns) of another factor. The total number of observations is $N$ and sparsity means that $N \ll RC$. The pattern of observations is also unstructured.

While large online commerce data sets have the crossed random effects structure, it is exceedingly difficult to analyze such data that way. Even for regression problems, the cost of computing the Gauss-Markov generalized least squares estimator grows as fast as $N^{3/2}$ or worse. The same holds for the cost of evaluating a Gaussian likelihood even once. The fundamental cause is that a linear equation of size $R + C$ must be solved. The usual algorithms solve it at cost proportional to $(R + C)^3$ and then because $RC \geqslant N$ we have $\max(R, C) \geqslant \sqrt{N}$ and now $(R + C)^3 = \Omega(N^{3/2})$. See Gao and Owen (2020) for details.

A cost of $N^{3/2}$ is infeasible in big data settings where $N$ is very large and growing. It is only recently that algorithms with an $O(N)$ cost have been produced. In the second chapter, we develop a backfitting algorithm based on the work of Buja et al. (1989). That algorithm computes the generalized least squares estimate and they give conditions where only $O(1)$ iterations are needed, with each iteration taking $O(N)$ work. A plain Gibbs sampler has the same $\Omega(N^{3/2})$ cost as naive linear modeling (Gao and Owen, 2017). A collapsed Gibbs sampler Papaspiliopoulos et al. (2020) can attain a cost of $O(N)$ under a strong balance condition that Ghosh and Zhong (2021) show can be weakened.

The response in an electronic commerce setting is usually categorical and very often binary, instead of the real valued responses that are common in other settings, such as agriculture. In a fixed effects setting, extending from least squares to generalized least squares is often done by a straightforward iteratively reweighted least squares approach. In a random effects setting, this path is not so simple because there is a very high dimensional integration problem that complicates it. Our main contribution is to find a way to adapt a penalized quasi-likelihood algorithm of Schall (1991) to mixed effects logistic regression models where there are crossed random effects. Schall's algorithm as written would cost $\Omega(N^{3/2})$ making it infeasible. We have adapted it by ignoring some off diagonal blocks in an $(R+C) \times (R+C)$ matrix and then showing that the effect of this simplification is asymptotically negligible. A second part of this adaptation comes from the fact that many of the feature variables in the commerce problem are functions of the product alone or of the customer alone. That brings a linear dependence that can slow convergence of iterative methods that alternate between updating random effects and fixed effects. We have devised a 'clubbing' algorithm that makes simultaneous updates to avoid this problem as described in more detail below.

We model a binary variable $Y_{ij} \in \{0, 1\}$ in terms of covariates $x_{ij} \in \mathbb{R}^p$ and row and column random effects $a_i$ and $b_j$. We assume that $p$ does not grow with $N$. We leave $p$ out of our cost estimates, giving the complexity in $N$. Conditionally on

$\boldsymbol{a} = (a_1, \dots, a_R)^\mathsf{T}$ and $\boldsymbol{b} = (b_1, \dots, b_C)^\mathsf{T}$ the $Y_{ij}$ are independent with

$$\Pr(Y_{ij} = 1 \,|\, \boldsymbol{a}, \boldsymbol{b}) = \Pr(Y_{ij} = 1 \,|\, a_i, b_j) = \frac{\exp(x_{ij}^\mathsf{T}\beta + a_i + b_j)}{1 + \exp(x_{ij}^\mathsf{T}\beta + a_i + b_j)}. \tag{3.1}$$

The random effects are $\boldsymbol{a} \sim \mathcal{N}(0, \sigma_A^2 I_R)$ independently of $\boldsymbol{b} \sim \mathcal{N}(0, \sigma_B^2 I_C)$. In this model the $x_{ij}$ are nonrandom, either because they were designed, or more usually because our analysis is conditional on their observed values.

We can also write (3.1) as

$$Y_{ij} = \begin{cases} 1, & x_{ij}^\mathsf{T}\beta + a_i + b_j + e_{ij} > 0 \\ 0, & \text{else} \end{cases} \tag{3.2}$$

for independent random variables $e_{ij}$ with the logistic CDF,

$$\Pr(e_{ij} \leqslant w) = \pi(w) \equiv \frac{e^w}{1 + e^w}. \tag{3.3}$$

The model (3.1) is a generalized linear mixed model (GLMM) owing to the appearance of both fixed effects $x_{ij}^\mathsf{T}\beta$ and random effects $a_i, b_j$.

While the linear model $Y_{ij} = x_{ij}^\mathsf{T}\beta + a_i + b_j + e_{ij}$ can be handled by the method of moments without assuming a distributional form for $a_i$, $b_j$ and $e_{ij}$, estimation for the model (3.2) depends on the shape of the distributions for $a_i$ and $b_j$ as well as $e_{ij}$. There is a near consensus in the literature that GLMM models are robust to mild departures from Gaussianity and correspondingly that such departures are hard to detect. See McCulloch and Neuhaus (2011) for references and also some contrary points of view. It is our view that the Gaussian model is of central importance and the problem is already hard enough under Gaussianity to motivate using that assumption.

The conditional likelihood of $\beta$ given $\boldsymbol{a}$ and $\boldsymbol{b}$ is

$$\begin{aligned} L(\beta \,|\, \boldsymbol{a}, \boldsymbol{b}) &= \prod_{(i,j)} \pi(x_{ij}^\mathsf{T}\beta + a_i + b_j)^{Y_{ij}} (1 - \pi(x_{ij}^\mathsf{T}\beta + a_i + b_j))^{1-Y_{ij}} \\ &= \prod_{(i,j)} \frac{e^{(x_{ij}^\mathsf{T}\beta + a_i + b_j)Y_{ij}}}{1 + e^{x_{ij}^\mathsf{T}\beta + a_i + b_j}} \end{aligned} \tag{3.4}$$

where the product is taken over pairs $(i,j)$ for which $(x_{ij},Y_{ij})$ is observed. The full likelihood incorporating random effect parameters is

$$L(\beta,\sigma_A^2,\sigma_B^2)=\int_{\mathbb{R}^{R+C}}L(\beta\,|\,\boldsymbol{a},\boldsymbol{b})\prod_{i=1}^{R}\frac{1}{\sigma_A}\varphi\Big(\frac{a_i}{\sigma_A}\Big)\prod_{j=1}^{C}\frac{1}{\sigma_B}\varphi\Big(\frac{b_j}{\sigma_B}\Big)\,\mathrm{d}\boldsymbol{a}\,\mathrm{d}\boldsymbol{b} \tag{3.5}$$

where $\boldsymbol{a}\in\mathbb{R}^R$ and $\boldsymbol{b}\in\mathbb{R}^C$ are vectors with components $a_i$ and $b_j$ respectively and $\varphi(\cdot)$ is the $\mathcal{N}(0,1)$ probability density function.

The high dimensional integral in (3.5) presents a major difficulty to finding estimates and confidence intervals for $\beta$, making the generalized linear mixed model (3.1) much harder to work with than Gaussian linear mixed models for regression studied in the second chapter. The integral is daunting not only because of its dimension but also because $L(\beta\,|\,\boldsymbol{a},\boldsymbol{b})$ is a product of $N$ probabilities and so it may easily underflow numerically.

The most popular way to handle the integral is via Laplace's method, which uses a single point in $\mathbb{R}^{R+C}$ to represent the integral, as Breslow and Clayton (1993) do in their penalized quasi-likelihood approach. Higher order Laplace integration methods cost $k^{R+C}$ for an integer $k\geqslant 1$, which is completely infeasible for any $k>1$ in our context. The lme4 code (Bates et al., 2015) uses sparse grid integration with a lower cost (Heiss and Winschel, 2008), but that also becomes too costly in our settings.

We have chosen to work with the quasi-likelihood formulation of Schall (1991). That algorithm is an iteratively reweighted penalized least squares solution that uses some weighted least squares fits to a working response vector in order to optimize the Laplace approximation of the integral. The estimand in Schall's setup is a posterior mode of $\beta$ under a diffuse prior that we describe below. It is thought to be close to the MLE except when some $\Pr(Y_{ij}=1\,|\,a_i,b_j)$ are very close to zero or one. Schall's algorithm is an iterative procedure with each iteration having two steps. Both of the steps have a cost that grows like $(R+C)^3=\Omega(N^{3/2})$. To get around this we use backfitting for the first step and an approximation to the trace of the inverse of an $(R+C)\times(R+C)$ matrix getting an algorithm that is $O(N)$ per iteration. We then study the impact of this approximation and find conditions under which is asymptotically negligible.

One of the main difficulties in random effects estimation arises from the fact that the intercept is hard to estimate. The easiest way to see this is to consider a balanced Gaussian model with $Y_{ij} = \mu + a_i + b_j + e_{ij}$. We would estimate $\mu$ by the average of all $RC$ observations getting variance $O(1/R + 1/C + 1/N)$ which is much larger than the $1/N$ rate from settings with $N$ IID observations.

The intercept remains difficult in the unbalanced setting with a binary response. The intercept contributes $\mu\mathbf{1}_N$ to the $N$-vector of logits of $\Pr(Y_{ij} = 1 \mid \boldsymbol{a}, \boldsymbol{b})$. The vector space spanned by indicators of the $a_i$ includes $\mu\mathbf{1}_N$. So does the one spanned by the indicators of $b_j$. This overlap among spaces slows the convergence of iterative algorithms to solve the quasi-likelihood equations. That motivated us to develop a generalized least squares algorithm in Ghosh et al. (2022) that alternates between updating $a_i$ along with the intercept and updating $b_j$ also with the intercept (see theorem 2.3.1).

A related issue arises when one of the predictor variables in $x_{ij}$, say $x_{ij\ell}$ is a binary variable whose value depends on only the row index $i$. That is $x_{ij\ell} = x_{i\bullet\ell}$ for all $j$. For instance one of the features in e-commerce might be a property of the customer or of the product while a feature in agriculture might be a property of the cultivar of wheat or a property of the environment in which it is grown. In such cases there is an overlap of vector spaces like the one described above for the intercept. The penalized likelihood estimate for the corresponding *subset* of $\boldsymbol{a}$ must sum to zero. This is true for every such partially aliased column $x_{ij\ell}$, so there can be many different such summation constraints. Similarly to the way the intercept is handled, the coefficient $\beta_\ell$ can then be efficiently updated *together* with $a_i$, a procedure we call *clubbing*. Clearly the same problem can happen with variables $x_{ij\ell}$ that are categorical with more than two levels or with variables whose level is defined by the index $j$. Our algorithm clubs together all $p$ variables in $\beta$ with $\boldsymbol{a}$ when updating $\boldsymbol{a}$, and also clubs $\beta$ with $\boldsymbol{b}$ when updating $\boldsymbol{b}$. The algorithm analyzed in the second chapter only clubs the intercept together with the random effects. The impetus for clubbing is that we found it made a much greater difference in the binary regression case than we saw in the generalized least squares setting.

One of the main open questions in generalized linear mixed models, like the ones we consider here, is at what rate do the estimates converge? Intuitively one might expect $O(N^{-1/2})$ or $O(\min(R, C)^{-1/2})$. In some settings different parameters converge at different rates. See Jiang (2013) for some discussion. It was only comparatively recently that consistency of $\hat{\beta}$ for the maximum likelihood estimate was proved in some generalized linear mixed models. See Jiang (2013).

An outline of this chapter is as follows. Section 3.2 introduces our notation and describes the missingness mechanism. Section 3.3 presents a penalized quasi-likelihood method based on a Laplace approximation. The optimization is done using Schall's method after incorporating backfitting and the matrix approximation referred to above. We then give conditions under which that approximation makes an asymptotically negligible difference. Section 3.4 describes in detail our clubbing strategy where we alternate between updating $(\hat{\beta}, \hat{\boldsymbol{a}})$ and $(\hat{\beta}, \hat{\boldsymbol{b}})$. Section 3.5 uses some simulated data to verify that the cost of a standard algorithm (glmer from Bates et al. (2015)) has a cost growing faster than $N^{3/2}$. Two versions of that algorithm show the expected superlinear cost and that superlinearity applies even to a single iteration. Our iteratively reweighted backfitting algorithm converges in $O(N)$ cost there. This holds in our example even though weighting will ordinarily violate the sufficient conditions that we give in theorem 2.4.1 for backfitting to cost $O(N)$. We also include a simulation that compares the accuracy of glmer, and our proposal and a naive logistic regression that ignores the random effects. Our algorithm's accuracy lies between that of two glmer algorithms, both of which have superlinear cost. A naive logistic regression is seen to be typically inconsistent. For glmer and our algorithm the error in the intercept coefficient decreases much more slowly than that of the other coefficients, in line with our earlier remarks. Section 3.6 illustrates our algorithm on a data set provided by Stitch Fix. Section 3.7 has a discussion about the Bayesian and frequentist perspectives. In Appendix B, we give details about the proofs and tabulate the results for the Stitch Fix data.

We conclude this section by describing some alternative approaches. Quasi-Monte Carlo (QMC) sampling (Niederreiter, 1992; Dick et al., 2013) is very well suited to many high dimensional integration problems but the integrand in (3.5) does not appear to be one of them. This integrand is very 'spiky' making it difficult for QMC. See Kuo et al. (2008) for some work in this area.

An option to alleviate the computational cost of likelihood computations is provided by composite likelihood methods (Lindsay, 1988). Composite likelihoods are inference functions constructed by multiplying component likelihoods. See Varin et al. (2011) for a review. Several articles discussed composite likelihood estimation of models for dependent binary data; see, for example, Heagerty and Lele (1998), Cox and Reid (2004) or Kuk (2007) . All these works consider a particular type of composite likelihood called pairwise likelihood that is built taking the product of bivariate marginal probabilities. Bellio and Varin (2005) investigated pairwise likelihood inference in logistic regression with crossed random effects. Their pairwise likelihood involves all the pairs of correlated observations, that is those pairs that share the row- or the column-random effect. The number of pairs involved in such pairwise likelihood does not satisfy the $\mathcal{O}(N)$ constraint. A possible remedy to the lack of scalability would be sampling the pairs of correlated observations so to achieve the $\mathcal{O}(N)$ computational cost. Although sampling might be a valid option, the estimates would depend on the pair selection mechanism. A comparison with composite likelihood would require consideration of precisely how to choose the $k$-tuples as well as the resulting efficiency. That would require an investigation of its own and is outside the scope of our investigation. We have opted to build upon the penalized quasi-likelihood method of Breslow and Clayton (1993). It is the most commonly used algorithm and there is well tested code in Bates et al. (2015) to use in comparisons.

It would be of great interest to handle mixed and generalized linear mixed models that incorporate SVD-like latent variable interactions. For the present we note that even without those interactions it is a significant challenge to get an efficient estimate of $\beta$ and also quantify the uncertainty in $\beta$ whether by Bayesian or frequentist methods while remaining within an $O(N)$ cost limit.

## 3.2 Notation and sampling models

We speak of rows and columns for our two factors. The rows are indexed by $i = 1, \dots, R$ and the columns by $j = 1, \dots, C$. Most of the $RC$ possible $(i, j)$ combinations are not observed in our motivating problems. We let $Z_{ij} \in \{0, 1\}$ take the value 1 if and only if $(x_{ij}, Y_{ij})$ has been observed. In our motivating problems we either never have the same $(i, j)$ combination observed twice, or we only keep the most recent such observation, or it happens so rarely that we can neglect it.

Sums over $i$ are from 1 to $R$ and sums over $j$ are from 1 to $C$. The number of times row $i$ has been observed is $N_{i\bullet} = \sum_j Z_{ij}$ and similarly, column $j$ has been observed $N_{\bullet j} = \sum_i Z_{ij}$ times. We suppose that the data are ordered so that all $N_{i\bullet} \geqslant 1$ and all $N_{\bullet j} \geqslant 1$. The total sample size is $N = \sum_{i=1}^{R} N_{i\bullet} = \sum_{j=1}^{C} N_{\bullet j}$. The sparsity condition implies that $N \ll RC$.

Sometimes we have to make a vector of length $N$ with an element for each observation or a matrix with $N$ rows, one per observation. We assume that some consistent ordering of the observations is used in all of these cases and we use a calligraphic font for most such quantities. The vector $\mathcal{Y} \in \{0, 1\}^N$ contains all the observed responses $Y_{ij}$. The matrix $\mathcal{Z}_A \in \{0, 1\}^{N\times R}$ has $R$ columns of which the $i$'th column contains ones for observations in row $i$ and zeros everywhere else. The matrix $\mathcal{Z}_B \in \{0, 1\}^{N\times C}$ is the corresponding incidence matrix for the columns of our observations. The product $\mathcal{Z}_A^\intercal \mathcal{Z}_B$ is our observation matrix $Z \in \{0, 1\}^{R\times C}$.

While the analyst considers $Z_{ij}$ to be fixed, we will study random $Z_{ij}$ in order to model the difficulties that the analyst will face and to prove that certain difficulties have vanishing probability in large samples. The model we use comes from section 2.4.2 of the second chapter. For a problem size parameter $S$, the number of rows and columns are $R = \lfloor S^\rho \rfloor$ and $C = \lfloor S^\kappa \rfloor$ respectively. Those values become very large in cases of interest and, to avoid uninteresting complications, we use $R = S^\rho$ and $C = S^\kappa$ in our formulas without taking integer parts. Then for some $\Upsilon > 1$, we suppose that

$$Z_{ij} \overset{\text{ind}}{\sim} \text{Bern}(p_{ij}) \quad \text{where} \quad \frac{S}{RC} \leqslant p_{ij} \leqslant \Upsilon \frac{S}{RC}. \tag{3.6}$$

The actual problem size $N$ satisfies $S \leqslant \mathbb{E}(N) \leqslant \Upsilon S$. The model (3.6) allows unequal observation probabilities but remains more restrictive than we would like. Numerical results in the second chapter show that their generalized least squares algorithm converges in $O(N)$ cost much more generally than their assumptions imply.

The relevant values of $\rho$ and $\kappa$ are positive so that $R$ and $C$ grow with $N$ but below 1 because neither $R$ nor $C$ can be larger than $N$. We also have $\rho + \kappa > 1$ to model sparsity, that is $N \ll RC$.

Under the model (3.6), we show in equation (2.32) that if $2\rho + \kappa < 2$ then all rows get an adequate sample size and no single row dominates the data:

$$\lim_{S\to\infty} \Pr\bigl((1-\epsilon)S^{1-\rho} \leqslant \min_i N_{i\bullet} \leqslant \max_i N_{i\bullet} \leqslant (\Upsilon + \epsilon)S^{1-\rho}\bigr) = 1 \tag{3.7}$$

for any $\epsilon > 0$. By symmetry, if $\rho + 2\kappa < 2$, then

$$\lim_{S\to\infty} \Pr\bigl((1-\epsilon)S^{1-\kappa} \leqslant \min_j N_{\bullet j} \leqslant \max_j N_{\bullet j} \leqslant (\Upsilon + \epsilon)S^{1-\kappa}\bigr) = 1 \tag{3.8}$$

holds for any $\epsilon > 0$.

Many settings can be expected to have informative missingness, where $Z_{ij}$ is related to $Y_{ij}$. We do not account for this because to do so would require problem specific assumptions from outside the data at hand. Also, the problems we face are already unsolved in the event that $Z_{ij}$ are not informative about $Y_{ij}$.

## 3.3 Likelihood approach

We will use an iterative approach of Schall (1991). It alternates between estimating $\boldsymbol{a}$, $\boldsymbol{b}$ and $\beta$ given $\sigma_A^2$, $\sigma_B^2$ and an overdispersion quantity $\phi$ and estimating the variances and overdispersion given $\boldsymbol{a}$, $\boldsymbol{b}$ and $\beta$. The resulting estimators maximize the Bayesian posterior probability density function for $(\beta, \boldsymbol{a}, \boldsymbol{b})$ under a diffuse prior for $\beta$ independent of a zero-mean Gaussian prior for $(\boldsymbol{a}, \boldsymbol{b})$ (see Schall (1991)). For the optimization over $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\beta$ we use weighted least squares for a working response, following the approach in Schall (1991). His approach requires the trace of the inverse of an $(R+C) \times (R+C)$ matrix and this is too expensive so we replace it by an approximation. We then give conditions where the approximation makes an asymptotically negligible difference.

Following Schall (1991), Breslow and Clayton (1993) derived the same set of estimating equations for $(\beta, \boldsymbol{a}, \boldsymbol{b})$ by maximizing the marginal likelihood using laplace approximations. We introduce some notation to discuss the relation between approach in Schall (1991) and penalized quasi-likelihood. We observe $Y_{ij}$ together with predictor variables $x_{ij}$. There are also unobserved random effects $(a_i, b_j)$ associated with this observation. We assumed that given $\boldsymbol{a}$ and $\boldsymbol{b}$, the $Y_{ij}$ are conditionally independent. Adapting some notation from Breslow and Clayton (1993) we let $\mu^{\boldsymbol{a},\boldsymbol{b}} \in \mathbb{R}^N$ have components $\mu_{ij}^{\boldsymbol{a},\boldsymbol{b}} \equiv \mathbb{E}(Y_{ij} \,|\, a_i, b_j) = \pi(x_{ij}^\mathsf{T}\beta + a_i + b_j)$. The model of Breslow and Clayton (1993) includes

$$\mathrm{var}(Y_{ij} \,|\, a_i, b_j) = \phi v(\mu_{ij})$$

for a function $v(\cdot)$ and an overdispersion parameter $\phi > 0$. In our case $v(\mu) = \mu(1-\mu)$. While Bernoulli random variables cannot be overdispersed or underdispersed we include $\phi$ because values $\phi \neq 1$ may signify a lack of fit.

We are going to use approximate quasi-likelihood $\mathrm{ql}(\beta, \sigma_A^2, \sigma_B^2, \phi)$ from Breslow and Clayton (1993). The gradient in equation (7) and (8) of Breslow and Clayton (1993) involves the quantity $v(\mu)\pi'(\mu)$ which equals 1 for $v$ and $\mu$ in our case.

Translating the score equation to our context, we obtain the following three equations:

$$\begin{aligned}
\mathcal{X}^\mathsf{T}(\mathcal{Y} - \mu^{\boldsymbol{a},\boldsymbol{b}}) &= \mathbf{0}, \\
\mathcal{Z}_A^\mathsf{T}(\mathcal{Y} - \mu^{\boldsymbol{a},\boldsymbol{b}}) &= \frac{\phi}{\sigma_A^2}\boldsymbol{a}, \quad \text{and} \\
\mathcal{Z}_B^\mathsf{T}(\mathcal{Y} - \mu^{\boldsymbol{a},\boldsymbol{b}}) &= \frac{\phi}{\sigma_B^2}\boldsymbol{b}.
\end{aligned} \tag{3.9}$$

Using the notion of working response $z$

$$z_{ij} = x_{ij}^\mathsf{T}\beta + a_i + b_j + \frac{Y_{ij} - \mu_{ij}}{\mu_{ij}(1 - \mu_{ij})} \tag{3.10}$$

with $\mu_{ij} = \pi(x_{ij}^\mathsf{T}\beta + a_i + b_j)$ for the logistic CDF $\pi(\cdot)$ from equation (3.3), we obtain the following estimating equations from (3.9):

$$\begin{aligned} \mathcal{X}^{\mathsf{T}}\mathcal{W}(z - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}) &= \mathbf{0}, \\ \mathcal{Z}_A^{\mathsf{T}}\mathcal{W}(z - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}) - \frac{1}{\sigma_A^2}\boldsymbol{a} &= \mathbf{0} \\ \mathcal{Z}_B^{\mathsf{T}}\mathcal{W}(z - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}) - \frac{1}{\sigma_B^2}\boldsymbol{b} &= \mathbf{0} \end{aligned} \tag{3.11}$$

where $\mathcal{W} = \mathrm{diag}(\mu^{\boldsymbol{a},\boldsymbol{b}}(1 - \mu^{\boldsymbol{a},\boldsymbol{b}}))/\phi \in [0, 1/(4\phi)]^{N\times N}$.

### 3.3.1 Schall's approach

Schall (1991) considers an iterative procedure to estimate the fixed effect, random effects and the variance component. At iteration $k$ we have estimates $\hat{\beta}^{(k)}$, $\boldsymbol{a}^{(k)}$, $\boldsymbol{b}^{(k)}$, $\hat{\sigma}_A^{2(k)}$, $\hat{\sigma}_B^{2(k)}$ and $\hat{\phi}^{(k)}$. For $k = 0$, we initialize them all to zeros except the overdispersion and variance parameters start at one. To obtain $\hat{\beta}^{(k+1)}$, $\hat{\boldsymbol{a}}^{(k+1)}$ and $\hat{\boldsymbol{b}}^{(k+1)}$ we solve the penalized weighted least squares problem

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \sum_{ij} Z_{ij}\, \widehat{W}_{ij}^{(k)}\big(z_{ij}^{(k)} - x_{ij}^{\mathsf{T}}\beta - a_i - b_j\big)^2 + \frac{\|\boldsymbol{a}\|^2}{\hat{\sigma}_A^{2(k)}} + \frac{\|\boldsymbol{b}\|^2}{\hat{\sigma}_B^{2(k)}}, \tag{3.12}$$

with weights $\widehat{W}_{ij}^{(k)} = \hat{\mu}_{ij}^{(k)}(1 - \hat{\mu}_{ij}^{(k)})/\hat{\phi}^{(k)}$. This optimization problem leads to exactly the same estimating equations (3.11) that we found for the approximate quasi-likelihood. Breslow and Clayton (1993) showed that their estimator based on Laplace approximation is the same as those of Stiratelli et al. (1984) and Schall (1991) (see Sect. 3 of Lee and Nelder (1996)).

It remains to update the overdispersion and variance parameters. Consider the following $(R + C) \times (R + C)$ matrix

$$\begin{aligned} T^{(k)} &= \begin{pmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ T_{21}^{(k)} & T_{22}^{(k)} \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{Z}_A^{\mathsf{T}} \widehat{\mathcal{W}}^{(k)} \mathcal{Z}_A + \hat{\sigma}_A^{-2(k)} I_R & \mathcal{Z}_A^{\mathsf{T}} \widehat{\mathcal{W}}^{(k)} \mathcal{Z}_B \\ \mathcal{Z}_B^{\mathsf{T}} \widehat{\mathcal{W}}^{(k)} \mathcal{Z}_A & \mathcal{Z}_B^{\mathsf{T}} \widehat{\mathcal{W}}^{(k)} \mathcal{Z}_B + \hat{\sigma}_B^{-2(k)} I_C \end{pmatrix} \end{aligned} \tag{3.13}$$

with $\widehat{\mathcal{W}}^{(k)} = \mathrm{diag}(\widehat{W}_{ij}^{(k)}) \in \mathbb{R}^{N \times N}$. Let

$$T^{*(k)} = \big(T^{(k)}\big)^{-1} = \begin{pmatrix} T_{11}^{*(k)} & T_{12}^{*(k)} \\ T_{21}^{*(k)} & T_{22}^{*(k)} \end{pmatrix}.$$

Schall's update evaluates

$$\nu_A^{(k+1)} = \mathrm{tr}(T_{11}^{*(k)}) / \hat{\sigma}_A^{2(k)} \quad \text{and} \quad \nu_B^{(k+1)} = \mathrm{tr}(T_{22}^{*(k)}) / \hat{\sigma}_B^{2(k)}$$

and then sets

$$\hat{\sigma}_A^{2(k+1)} = \frac{\|\hat{\boldsymbol{a}}^{(k+1)}\|^2}{R - \nu_A^{(k+1)}} \quad \text{and} \quad \hat{\sigma}_B^{2(k+1)} = \frac{\|\hat{\boldsymbol{b}}^{(k+1)}\|^2}{C - \nu_B^{(k+1)}}. \tag{3.14}$$

From these formulas it is clear that the $\nu$ parameters function as degrees of freedom estimates. Finally, we let

$$\hat{\phi}^{(k+1)} = \frac{\sum_{ij} Z_{ij} \hat{\mu}_{ij}^{(k+1)} (1 - \hat{\mu}_{ij}^{(k+1)}) \big( z_{ij}^{(k+1)} - x_{ij}^{\mathsf{T}} \hat{\beta}^{(k+1)} - \hat{a}_i^{(k+1)} - \hat{b}_j^{(k+1)} \big)^2}{N - p - (R - \nu_A^{(k+1)}) - (C - \nu_B^{(k+1)})} \tag{3.15}$$

where $p$ is the number of parameters in $\beta$ including the intercept.

The quantity Schall computes is almost but not quite the maximum likelihood estimate of $\beta$. Schall (1991) shows that it is a quantity studied by Stiratelli et al. (1984), namely the posterior mode of $\beta$ under a diffuse prior on $\beta$ independent of a zero-mean Gaussian prior for $(\boldsymbol{a}, \boldsymbol{b})$. Stiratelli et al. (1984) mentioned different approaches like empirical Bayes for the estimation of variance components of the prior distribution (See Leonard (1975), Laird (1978)). The update (3.12) for $\beta$, $\boldsymbol{a}$ and $\boldsymbol{b}$ is a Fisher scoring iteration to maximize their posterior density. The updates (3.14) and (3.15) for $\sigma_A^2$, $\sigma_B^2$ and $\phi$ are from an EM iteration to compute these dispersion components, after approximating the posterior distribution of $(\boldsymbol{a}, \boldsymbol{b})$ by a multivariate normal distribution with the same mode and curvature as the true posterior. In small data sets where we are able to compute the MLE we find the estimate from Schall's algorithm is very close to the MLE.

### 3.3.2 Modified Schall Approach

We want to solve the equation (3.12). It is instructive to begin with the case of just one factor. Then we obtain the following optimization problem:

$$\min_{\beta,\boldsymbol{a}} \sum_{i,j} Z_{ij} \widehat{W}_{ij}^{(k)} \bigl(z_{ij}^{(k)} - x_{ij}^\mathsf{T}\beta - a_i\bigr)^2 + \frac{\|\boldsymbol{a}\|^2}{\hat{\sigma}_A^{2(k)}}. \tag{3.16}$$

The normal equations from (3.16) yield

$$\mathbf{0} = \mathcal{X}^\mathsf{T} \widehat{\mathcal{W}}^{(k)} (z^{(k)} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A \hat{\boldsymbol{a}}), \quad \text{and} \tag{3.17}$$

$$\mathbf{0} = \mathcal{Z}_A^\mathsf{T} \widehat{\mathcal{W}}^{(k)} (z^{(k)} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A \hat{\boldsymbol{a}}) - \hat{\boldsymbol{a}}/\hat{\sigma}_A^{2(k)}. \tag{3.18}$$

Solving (3.18) for $\hat{\boldsymbol{a}}$ and multiplying the solution by $\mathcal{Z}_A$ yields

$$\begin{aligned} \mathcal{Z}_A \hat{\boldsymbol{a}} &= \mathcal{Z}_A \bigl(\mathcal{Z}_A^\mathsf{T} \widehat{\mathcal{W}}^{(k)} \mathcal{Z}_A + \hat{\sigma}_A^{-2(k)} I_R\bigr)^{-1} \mathcal{Z}_A^\mathsf{T} \widehat{\mathcal{W}}^{(k)} (z^{(k)} - \mathcal{X}\hat{\beta}) \\ &\equiv \mathcal{S}_A^{(k)} (z^{(k)} - \mathcal{X}\hat{\beta}) \end{aligned}$$

for an $N \times N$ ridge regression "smoother matrix" $\mathcal{S}_A^{(k)}$. (We have rederived one of the estimating equations in (3.11).) This smoother matrix implements weighted shrunken within-group means. Substituting $\mathcal{Z}_A \hat{\boldsymbol{a}}$ into equation (3.17) yields

$$\hat{\beta} = \bigl(\mathcal{X}^\mathsf{T} \widehat{\mathcal{W}}^{(k)} (I_N - \mathcal{S}_A^{(k)}) \mathcal{X}\bigr)^{-1} \mathcal{X}^\mathsf{T} \widehat{\mathcal{W}}^{(k)} (I_N - \mathcal{S}_A^{(k)}) z^{(k)}. \tag{3.19}$$

None of these steps takes superlinear time since weighted shrunken within-group means cost $O(N)$ time. Also observe that $\widehat{\mathcal{W}}^{(k)}\mathcal{S}_A^{(k)}$ is symmetric. We can use this fact to efficiently compute an estimated (asymptotic) covariance of $\hat{\beta}$. With $\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)} = \widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_A^{(k)})$, we have

$$\widehat{\mathrm{cov}}(\hat{\beta}) = (\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)} \cdot \widehat{\Sigma} \cdot \widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X}(\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X})^{-1}, \tag{3.20}$$

with $\widehat{\Sigma} = \hat{\sigma}_A^{-2(k)}\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T} + (\widehat{\mathcal{W}}^{(k)})^{-1}$ the covariance of the working response. Since $\widehat{\Sigma}$ has the form low-rank plus diagonal, we can compute this covariance with $O(N)$ computations.

With two factors we do not enjoy the same computational simplifications. The counterpart to equation (3.19) is

$$\hat{\beta} = (\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_{AB}^{(k)})\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_{AB}^{(k)})z^{(k)} \tag{3.21}$$

where

$$\mathcal{S}_{AB}^{(k)} = \mathcal{Z}(\mathcal{Z}^\mathsf{T}\widehat{\mathcal{W}}^{(k)}\mathcal{Z} + (D^{(k)})^{-1})^{-1}\mathcal{Z}$$

for $\mathcal{Z} = [\mathcal{Z}_A : \mathcal{Z}_B]$ and $D^{(k)} = \mathrm{diag}(\hat{\sigma}_A^{2(k)}I_R, \hat{\sigma}_B^{2(k)}I_C)$. Hence we would need to invert an $(R+C)\times(R+C)$ matrix $T^{(k)}$ in (3.13) to apply $\mathcal{S}_{AB}^{(k)}$ and thereby solve (3.21), incurring a cost far greater than $O(N)$.

However, in order to solve (3.21), all we need to do is apply the operator $\mathcal{S}_{AB}^{(k)}$ to each column of $\mathcal{X}$, and this can be done more efficiently. Consider a generic response vector $r$ (such as a column of $\mathcal{X}$) and the optimization problem

$$\min_{\boldsymbol{a},\boldsymbol{b}} \|r - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|^2_{\widehat{\mathcal{W}}^{(k)}} + \hat{\sigma}_A^{-2(k)}\|\boldsymbol{a}\|^2 + \hat{\sigma}_B^{-2(k)}\|\boldsymbol{b}\|^2. \tag{3.22}$$

It is clear that the fitted values are given by $\widehat{r} = \mathcal{S}_{AB}^{(k)}r$. Solving (3.22) leads to the following two blocks of estimating equations:

$$\begin{aligned}\mathcal{Z}_A\hat{\boldsymbol{a}} &= \mathcal{S}_A^{(k)}(r - \mathcal{Z}_B\hat{\boldsymbol{b}}),\\ \mathcal{Z}_B\hat{\boldsymbol{b}} &= \mathcal{S}_B^{(k)}(r - \mathcal{Z}_A\hat{\boldsymbol{a}}).\end{aligned} \tag{3.23}$$

We can solve these equations iteratively by block coordinate descent (backfitting).

This is done in parallel with $r$ being each column of $\mathcal{X}$ (separately) obtaining $\mathcal{S}_{AB}^{(k)}\mathcal{X}$ at convergence.

Similar to (3.20), we obtain the covariance estimate for the two factor case.

$$\widehat{\mathrm{cov}}(\hat{\beta}) = (\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)} \cdot \widehat{\Sigma} \cdot \widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X}(\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X})^{-1}, \tag{3.24}$$

with $\widehat{\Sigma} = \hat{\sigma}_A^{-2(k)}\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T} + \hat{\sigma}_B^{-2(k)}\mathcal{Z}_B\mathcal{Z}_B^\mathsf{T} + (\widehat{\mathcal{W}}^{(k)})^{-1}$ the covariance of the working response. Again, because of the low-rank-plus-diagonal nature of $\widehat{\Sigma}$, we can compute the covariance with $O(N)$ computations.

In practice, we will need to repeatedly minimize (3.12) for each step $k$ as the weights $\widehat{\mathcal{W}}^{(k)}$ and $\hat{\sigma}_A^{-2(k)}$ and $\hat{\sigma}_B^{-2(k)}$ vary, and only compute the covariance estimate after convergence. We develop a more efficient algorithm for this purpose which we describe in Section 3.4.

Now we consider the second step of Schall's method. Recall that $T^{(k)}$ is a $(R + C) \times (R + C)$ matrix without any special structure.

Although the diagonal blocks in $T^{(k)}$ are diagonal matrices, the off diagonal block have no special structure. Inverting $T^{(k)}$ would have a cost of $N^{3/2}$ or worse. Our approximation to Schall's algorithm uses

$$\nu_A^{(k+1)} = \mathrm{tr}((T_{11}^{(k)})^{-1})/\hat{\sigma}_A^{2(k)} \quad \text{and} \quad \nu_B^{(k+1)} = \mathrm{tr}((T_{22}^{(k)})^{-1})/\hat{\sigma}_B^{2(k)} \tag{3.25}$$

simply ignoring the off diagonal blocks of $T^{(k)}$. These can be computed in $O(N)$ time. Getting the trace of $(T_{11}^{(k)})^{-1}$ and $(T_{22}^{(k)})^{-1}$ costs $O(R + C)$ because those matrices are diagonal. The $i$'th diagonal element of $T_{11}^{(k)}$ is $\sum_j Z_{ij}\widehat{W}_{ij}^{(k)} + \hat{\sigma}_A^{-2(k)}$ and so all elements of $(T_{11}^{(k)})^{-1}$ and $(T_{22}^{(k)})^{-1}$ can be computed in $O(N)$ work. We show in the following subsection that this approximation ignoring the off diagonal blocks makes an asymptotically negligible difference.

Schall's approach requires the trace for each of two blocks of the inverse of the partitioned matrix

$$T = \begin{pmatrix} \mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_A + \sigma_A^{-2}I_R & \mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_B \\ \mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_A & \mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_B + \sigma_B^{-2}I_C \end{pmatrix} \tag{3.26}$$

with a diagonal weight matrix $\mathcal{W} \in (0, 1/(4\phi)]^{N\times N}$. Computing those traces directly costs $O((R+C)^3)$ because of the inversion step and this is infeasible in our applications. Instead ignoring the off-diagonal blocks of $T$ as mentioned above leads to our use of

$$\mathrm{tr}((\mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_A + \sigma_A^{-2}I_R)^{-1}) \quad \text{and} \quad \mathrm{tr}((\mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_B + \sigma_B^{-2}I_C)^{-1}) \tag{3.27}$$

for the two traces of blocks of $T^{-1}$.

We justify this approximation in two steps. First we give a representation for the error incurred in ignoring off diagonal blocks when taking the trace of the inverse. Then we show that error formula is asymptotically negligible under a sampling model for $Z_{ij}$. We work with the true weights $W_{ij}$ and comment later on the implications for estimated weights $\widehat{W}_{ij}$.

We define the weight sums

$$W_{i\bullet} = \sum_j Z_{ij}W_{ij} \quad \text{and} \quad W_{\bullet j} = \sum_i Z_{ij}W_{ij}.$$

Then writing

$$T = \begin{pmatrix} \mathrm{diag}(W_{i\bullet} + \sigma_A^{-2}) & W \\ W^\mathsf{T} & \mathrm{diag}(W_{\bullet j} + \sigma_B^{-2}) \end{pmatrix}$$

we see that the traces in (3.27) involve diagonal matrices and so they can be computed in $O(R+C)$ work after $O(N)$ work to compute their elements.

For $0 \leqslant \eta \leqslant 1$ let

$$T(\eta) = \begin{pmatrix} A & \eta B \\ \eta B^\mathsf{T} & C \end{pmatrix} \tag{3.28}$$

where matrices $A$, $B$ and $C$ are defined so that $T(1)$ is Schall's matrix $T$ from (3.26). There should be no confusion between the matrix $C$ here and our number of columns. Blocks $A$ and $C$ are diagonal in our problem and $T(\eta)$ is diagonally dominant for all $0 \leqslant \eta \leqslant 1$. Therefore $T = T(1)$ is invertible. In the proof of Lemma 3.3.1 we will see that $\mathrm{tr}(T(\eta)^{-1}) - \mathrm{tr}(T(0)^{-1}) = O(\eta^2)$ so that small perturbations $\eta B$ do not bring a large approximation error. However we need to study $\eta = 1$ so a small $\eta$ analysis is not quite enough, so the following Lemma 3.3.2 shows that $\eta = 1$ is within the radius of convergence.

*Lemma* 3.3.1. Let $T(\eta) \in \mathbb{R}^{(R+C)\times(R+C)}$ have the form (3.28) for positive definite diagonal matrices $A$ and $C$. Define $B_* = A^{-1/2}BC^{-1/2}$ and let $\rho$ be the spectral radius of

$$\begin{pmatrix} 0 & B_* \\ B_*^\mathsf{T} & 0 \end{pmatrix}.$$

Then for $0 \leqslant \eta < 1/\rho$,

$$\mathrm{tr}(T(\eta)^{-1}) = \mathrm{tr}(T(0)^{-1}) + \mathrm{tr}\Big(A^{-1}\sum_{k\geqslant 1}(\eta^2 B_* B_*^\mathsf{T})^k\Big) + \mathrm{tr}\Big(C^{-1}\sum_{k\geqslant 1}(\eta^2 B_*^\mathsf{T} B_*)^k\Big).$$

*Proof.* See Appendix B.1.1. □

The above Lemma gives the error in the trace. From the proof we see that the two specific error terms that we need to bound are

$$\mathrm{tr}\Big(A^{-1}\sum_{k\geqslant 1}(B_* B_*^\mathsf{T})^k\Big) \quad \text{and} \quad \mathrm{tr}\Big(C^{-1}\sum_{k\geqslant 1}(B_*^\mathsf{T} B_*)^k\Big) \tag{3.29}$$

after setting $\eta = 1$. To use Lemma 3.3.1 with $\eta = 1$ we need to verify that the given spectral radius $\rho$ is below 1 for our setting. We use Lemma 3.3.2 to control the spectral radius.

*Lemma* 3.3.2. In the notation of Lemma 3.3.1, let $B$ have nonnegative entries. Then the spectral radius $\rho$ satisfies

$$\rho \leqslant \Big(\max_i \sum_j \frac{B_{ij}}{C_{jj}}\Big)^{1/2}\Big(\max_j \sum_i \frac{B_{ij}}{A_{ii}}\Big)^{1/2}.$$

*Proof.* See Appendix B.1.2. □

In our present context $A = \mathrm{diag}(W_{i\bullet} + \sigma_A^{-2})$, $C = \mathrm{diag}(W_{\bullet j} + \sigma_B^{-2})$ and $B = W$.

Then

$$\max_i \sum_j \frac{B_{ij}}{C_{jj}} = \max_i \sum_j \frac{W_{ij}}{W_{i\bullet} + \sigma_A^{-2}} = \max_i \frac{W_{i\bullet}}{W_{i\bullet} + \sigma_A^{-2}} < 1$$

and we get

$$\rho < \left( \max_i \frac{W_{i\bullet}}{W_{i\bullet} + \sigma_A^{-2}} \times \max_j \frac{W_{\bullet j}}{W_{\bullet j} + \sigma_B^{-2}} \right)^{1/2} < 1$$

as required.

Now $W_{ij} = \mu_{ij}(1 - \mu_{ij})/\phi \in (0, 1/(4\phi)]$ for all $(i, j)$ in the data. We will need a strictly positive lower bound $W_{ij} \geqslant \underline{\omega} > 0$. In many applications, we can reasonably assume that $|x_{ij}^\mathsf{T}\beta|$ is bounded away from infinity for all $(i, j)$, even the unobserved $(i, j)$ pairs. However, our model uses $x_{ij}^\mathsf{T}\beta + a_i + b_j$ where both $a_i$ and $b_j$ are unbounded Gaussian random variables. Now $\max_{1 \leqslant i \leqslant R} a_i$ is asymptotically like $\sqrt{2\log(R)}\sigma_A$ and the $b_j$ satisfy a similar bound. Thus we will assume that

$$\lim_{S\to\infty} \Pr\Big( \max_{1\leqslant i\leqslant R} \max_{1\leqslant j\leqslant C} |x_{ij}^\mathsf{T}\beta + a_i + b_j| > \alpha \log(S) \Big) = 0$$

holds for any $\alpha > 0$. Then for any $\psi > 0$,

$$\lim_{S\to\infty} \Pr\Big( \min_{1\leqslant i\leqslant R} \min_{1\leqslant j\leqslant C} W_{ij} < S^{-\psi} \Big) = 0. \tag{3.30}$$

We need to bound the largest eigenvalue of $B_* B_*^\mathsf{T}$ from equation (3.29). In our context that matrix equals

$$\mathrm{diag}(W_{i\bullet} + \sigma_A^{-2})^{-1/2} W \mathrm{diag}(W_{\bullet j} + \sigma_B^{-2})^{-1} W^\mathsf{T} \mathrm{diag}(W_{i\bullet} + \sigma_A^{-2})^{-1/2}. \tag{3.31}$$

*Proposition* 3.3.1. Let $R$, $C$ and $Z_{ij}$ for $1 \leqslant i \leqslant R$ and $1 \leqslant j \leqslant C$ be sampled as in Section 3.2 with $\rho, \kappa \in (0, 1)$ and $\max(2\rho + \kappa, \rho + 2\kappa) < 2$. Let $\lambda_1$ be the largest eigenvalue of the matrix at (3.31). Then if $W_{ij} \leqslant \bar{\omega} < \infty$

$$\lim_{S\to\infty} \Pr\left( \lambda_1 \leqslant \frac{1}{1 + S^{\rho-1}/[(\Upsilon + \epsilon)\bar{\omega}\sigma_A^2]} \frac{1}{1 + S^{\kappa-1}/[(\Upsilon + \epsilon)\bar{\omega}\sigma_B^2]} \right) = 1$$

holds for any $\epsilon > 0$.

*Proof.* By equations (3.7) and (3.8) both

$$\max_i W_{i\bullet} \leqslant (\Upsilon + \epsilon)\bar{\omega} S^{1-\rho} \quad \text{and} \quad \max_j W_{\bullet j} \leqslant (\Upsilon + \epsilon)\bar{\omega} S^{1-\kappa}$$

hold with probability tending to 1 as $S \to \infty$ for any $\epsilon > 0$. Now

$$\lambda_1 \leqslant \max_i \frac{W_{i\bullet}}{W_{i\bullet} + \sigma_A^{-2}} \times \max_j \frac{W_{\bullet j}}{W_{\bullet j} + \sigma_B^{-2}}.$$

Hence, for any $\epsilon > 0$

$$\lim_{S\to\infty} \Pr\biggl(\lambda_1 \leqslant \frac{1}{1 + S^{\rho-1}/[(\Upsilon+\epsilon)\bar{\omega}\sigma_A^2]} \frac{1}{1 + S^{\kappa-1}/[(\Upsilon+\epsilon)\bar{\omega}\sigma_B^2]}\biggr) = 1.$$

□

*Theorem* 3.3.1. Let $R = S^\rho$, $C = S^\kappa$ and $Z_{ij}$ follow the sampling model from Section 3.2 for some $\Upsilon < \infty$. Assume that $\rho, \kappa \in (0, 1)$ with $\max(2\rho+\kappa, \rho+2\kappa) < 2$ and that $\Pr(\min_{ij} W_{ij} > S^{-\psi}) \to 1$ for all $\psi > 0$ as described prior to equation (3.30). Let the matrix $B_* B_*^\mathsf{T}$ in (3.31) have eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_R \geqslant 0$. Then there exists $\alpha < 1$ such that for any $\delta > 0$

$$\lim_{S\to\infty} \Pr\biggl(\sum_{i=1}^R 1\{\lambda_i > \delta\} > R^\alpha\biggr) = 0. \tag{3.32}$$

*Proof.* See Appendix B.1.3. □

*Theorem* 3.3.2. Let $R = S^\rho$, $C = S^\kappa$ and $Z_{ij}$ for $1 \leqslant i \leqslant R$ and $1 \leqslant j \leqslant C$ be sampled as in Section 3.2. Assume that $\rho, \kappa \in (0, 1)$ and that $\Pr(\min_{ij} W_{ij} > S^{-\psi}) \to 1$ for all $\psi > 0$ as described prior to equation (3.30). Then our approximation error from (3.29) is

$$\mathrm{Err} \equiv \mathrm{tr}\biggl(A^{-1} \sum_{k\geqslant 1} (B_* B_*^\mathsf{T})^k\biggr)$$

and for any $\gamma > 0$

$$\lim_{S\to\infty} \Pr(\mathrm{Err} > \gamma R) = 0.$$

*Proof.* See Appendix B.1.4. □

### 3.3.3 Error with $\widehat{W}$ versus $W$

Our algorithm makes a trace approximation to give Schall's algorithm a feasible cost. We have proved that this approximation brings a negligible error when the true $W_{ij} = \mu_{ij}(1-\mu_{ij})/\phi$ are used. In practice the algorithm uses estimates $\widehat{W}_{ij}, \hat{\sigma}_A^2, \hat{\sigma}_B^2$ and $\hat{\phi}$. The principle difference is that estimates $\widehat{W}_{ij}$ might be closer to zero than the true $W_{ij}$. We actually do not expect this to happen. The process of estimating $\beta$ and $\boldsymbol{a}$ and $\boldsymbol{b}$ biases them towards the origin and consequently away from very small weights. We could in principle choose a small value $\psi$ and impose a minimum $\widehat{W}_{ij} \geqslant N^{-\psi}$ in each stage of the algorithm but this has not been necessary.

We have also studied the algorithm assuming that the true $\sigma_A^2, \sigma_B^2$ and $\phi$ are used, while the algorithm runs with estimates $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\phi}$. The value of $\hat{\phi}$ is not consequential for Theorem 3.3.1 because it scales all the weights $\widehat{W}_{ij} = \hat{\mu}_{ij}(1-\hat{\mu}_{ij})/\hat{\phi}$ by the same factor leaving the weight ratios in the proof unchanged. Smaller $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ reduce the bound on $\lambda_1$ so they pose no difficulty. Larger values of $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ imply smaller ridge penalties on the weighted least squares problems and thereby increase the upper bound on the largest eigenvalue $\lambda_1$ in Proposition 3.3.1. Very large values are a priori implausible for our target setting. For instance, enormous $\sigma_A^2$ would tend to make $Y_{ij}$ equal 0 or 1 depending almost entirely on the row index $i$. In practice one could impose a constraint like $\max(\hat{\sigma}_A, \hat{\sigma}_B) \leqslant 10$ but we have not had to do that. For Theorem 3.3.2, we only need $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ to be $o(S^\delta)$ for any $\delta > 0$.

## 3.4 Clubbed backfitting

The iterative method in Section 3.3.2 to solve (3.12) might suffer from slow convergence because of confounding of a factor with one of the variables $x_{ij\ell}$. Let us consider the following optimization problem motivated by the single factor problem (3.16):

$$\min_{\beta,\boldsymbol{a}} \|z - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a}\|_{\boldsymbol{w}}^2 + \sigma_A^{-2}\|\boldsymbol{a}\|^2. \tag{3.33}$$

It may happen that a column of $\mathcal{X}$ is exactly equal to the sum of $k$ columns of $\mathcal{Z}_A$, e.g., columns $\ell_1, \ell_2, \dots, \ell_k$. For example if the rows in the data represent customers, this column of $\mathcal{X}$ could represent all male customers older than 50. This linear dependence has consequences for the solution $\boldsymbol{a}$ to (3.33).

*Lemma* 3.4.1. Suppose that column $q$ of $\mathcal{X}$ in problem (3.33) equals the sum of $k \geqslant 1$ distinct columns $\mathcal{I} = \{i_1, i_2, \dots, i_k\}$ of $\mathcal{Z}_A$. If $\sigma_A^{-2} > 0$ then the solution for $\boldsymbol{a}$ satisfies $\sum_{\ell=1}^k a_{i_\ell} = 0$.

*Proof.* For any $\boldsymbol{a} \in \mathbb{R}^R$, $\beta \in \mathbb{R}^p$ and $c \in \mathbb{R}$, define $\boldsymbol{a}^{(c)}$ and $\beta^{(c)}$ via

$$a_i^{(c)} = \begin{cases} a_i, & \text{if } i \in \mathcal{I} \\ a_i - c, & \text{else} \end{cases} \qquad \text{and} \qquad \beta_\ell^{(c)} = \begin{cases} \beta_\ell + c, & \text{if } \ell = q \\ \beta_\ell, & \text{else.} \end{cases}$$

If we evaluate the quadratic in (3.33) at $(\beta^{(c)}, \boldsymbol{a}^{(c)})$ then the first term does not depend on $c$ and the second term has a unique minimum at $c = (1/k)\sum_{i\in\mathcal{I}} a_i$. As a result $(\beta, \boldsymbol{a})$ can only be the solution if $c = (1/k)\sum_{\ell=1}^k a_{i_\ell} = 0$. □

Lemma 3.4.1 imposes a constraint on $\boldsymbol{a}$ for every column of $\mathcal{X}$ that equals a sum of columns of $\mathcal{Z}_A$. Backfitting will eventually converge to a solution satisfying those constraints, but it could take a long time to get there. We can speed things up by enforcing any known constraints. When as usual, $\mathcal{X}$ includes a column of ones for an intercept we get the constraint $\sum_{i=1}^R a_i = 0$, as a special case.

With two factors we can also have such aliasing with sums of columns of $\mathcal{Z}_B$ as well, with particular columns of $\mathcal{X}$. For example a column of $\mathcal{X}$ might represent a particular category of clothing, corresponding to a number of levels of the column factor. Furthermore there may be other non-trivial constraints that are implied by $\mathcal{X}$. As we saw in equation (3.12), we want to obtain $\arg\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \mathrm{pl}(\beta, \boldsymbol{a}, \boldsymbol{b})$ for the penalized least squares problem

$$\mathrm{pl}(\beta, \boldsymbol{a}, \boldsymbol{b}) = \sum_{i,j} Z_{ij}\widehat{W}_{ij}\left(z_{ij} - x_{ij}^\mathsf{T}\beta - a_i - b_j\right)^2 + \hat{\sigma}_A^{-2}\|\boldsymbol{a}\|^2 + \hat{\sigma}_B^{-2}\|\boldsymbol{b}\|^2. \tag{3.34}$$

We use the following iterative "clubbing" strategy to enforce all of these constraints automatically. Given $(\beta^{(k)}, \boldsymbol{a}^{(k)}, \boldsymbol{b}^{(k)})$ we first optimize $\mathrm{pl}(\cdot)$ over $\beta$ and $\boldsymbol{a}$ with $\boldsymbol{b} = \boldsymbol{b}^{(k)}$ fixed to get $\boldsymbol{a}^{(k+1)}$ and an intermediate quantity $\beta^{(k+\frac{1}{2})}$. Then we optimize $\mathrm{pl}(\cdot)$ over $\beta$ and $\boldsymbol{b}$ with $\boldsymbol{a} = \boldsymbol{a}^{(k+1)}$ fixed to get $\boldsymbol{b}^{(k+1)}$ and the next iterate $\beta^{(k+1)}$ of the regression vector.

It is convenient to describe the updates in terms of fitted quantities in $\mathbb{R}^N$. The needed parts $\beta^{(k)}$, $\boldsymbol{a}^{(k)}$ and $\boldsymbol{b}^{(k)}$ are easily obtained in the process. For the first part of the iteration we set

$$\begin{aligned} \mathcal{X}\beta^{(k+\frac{1}{2})} &= (\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}(I_N - \mathcal{S}_A)\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}(I_N - \mathcal{S}_A)(z - \mathcal{Z}_B\boldsymbol{b}^{(k)}), \quad \text{and} \\ \mathcal{Z}_A\boldsymbol{a}^{(k+1)} &= \mathcal{S}_A(z - \mathcal{X}\beta^{(k+\frac{1}{2})} - \mathcal{Z}_B\boldsymbol{b}^{(k)}) \end{aligned} \tag{3.35}$$

for a smoother matrix $\mathcal{S}_A = \mathcal{Z}_A(\mathcal{Z}_A^\mathsf{T}\widehat{\mathcal{W}}\mathcal{Z}_A + \hat{\sigma}_A^{-2}I_R)^{-1}\mathcal{Z}_A^\mathsf{T}\widehat{\mathcal{W}}$ that simply computes weighted group means. This gets applied to each column of $\mathcal{X}$ in the first equation, and to the residual in the second. The matrix inversion in the formula is handled by solving a $p \times p$ system of equations which adds a cost that is of constant order in $N$.

The equations (3.35) solve the following minimization problem,

$$\min_{\beta,\boldsymbol{a}} \sum_{i,j} Z_{ij}\widehat{W}_{ij}\big(z_{ij} - b_j^{(k)} - x_{ij}^\mathsf{T}\beta - a_i\big)^2 + \hat{\sigma}_A^{-2}\|\boldsymbol{a}\|^2. \tag{3.36}$$

If we absorb $b_j^{(k)}$ into $z_{ij}$ we see that (3.36) has the form (3.33) and so by Lemma 3.4.1 the solution $(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)})$ satisfies any constraint that is intrinsic to the design.

To complete the iteration we fix $\boldsymbol{a} = \boldsymbol{a}^{(k+1)}$ and optimize over $\beta$ and $\boldsymbol{b}$, via

$$\begin{aligned} \mathcal{X}\beta^{(k+1)} &= \mathcal{X}(\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}(I_N - \mathcal{S}_B)\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}(I_N - \mathcal{S}_B)(z - \mathcal{Z}_A\boldsymbol{a}^{(k+1)}), \quad \text{and} \\ \mathcal{Z}_B\boldsymbol{b}^{(k+1)} &= \mathcal{S}_B(z - \mathcal{X}\beta^{(k+1)} - \mathcal{Z}_A\boldsymbol{a}^{(k+1)}) \end{aligned} \tag{3.37}$$

for a smoother matrix $\mathcal{S}_B = \mathcal{Z}_B(\mathcal{Z}_B^\mathsf{T}\widehat{\mathcal{W}}\mathcal{Z}_B + \hat{\sigma}_B^{-2}I_C)^{-1}\mathcal{Z}_B^\mathsf{T}\widehat{\mathcal{W}}$. Once again Lemma 3.4.1 applies to the solution.

*Lemma* 3.4.2. If $\mathcal{X}$ is of full rank, and $\max(\hat{\sigma}_A^2, \hat{\sigma}_B^2) < \infty$ and the weights $\widehat{W}_{ij}$ are positive, then the iterative algorithm converges to a global minimum of the equation (3.34).

*Proof.* See Appendix B.1.5. □

We can compute these single factor operators in $O(N)$ computation and use them for each iteration of the inner (solving the weighted least squares problem for a fixed weight vector) backfitting loop rather than computing it every time we run this iteration. The iterative algorithm stops when the relative change in $\zeta \equiv \mathcal{X}\beta + \mathcal{Z}_A\boldsymbol{a} + \mathcal{Z}_B\boldsymbol{b}$ is below a certain threshold. At convergence of the clubbed backfitting, we obtain $\hat\beta, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}$ for a particular set of weights.

Next we update the weights using the new set of parameters and solve a new optimization problem until convergence. We stop when the relative change in the fitted values $\zeta$ obtained with a different set of weights is negligible. Not only does this clubbed variant of block-coordinate descent automatically satisfy the implicit constraints, but we expect (and indeed observe) faster convergence than other variants that do not enforce the constraints. Here is a concise description of the modified Schall algorithm with clubbed backfitting:

1) Stage $\ell$ of the algorithm provides $\hat\beta^{(\ell)}$, $\hat{\boldsymbol{a}}^{(\ell)}$, $\hat{\boldsymbol{b}}^{(\ell)}$ using the weights obtained in the $(\ell-1)$'st iteration and the clubbing method described in (3.35) and (3.37). Obtain $\hat\zeta^{(\ell)} = \mathcal{X}\hat\beta^{(\ell)} + \mathcal{Z}_A\hat{\boldsymbol{a}}^{(\ell)} + \mathcal{Z}_B\hat{\boldsymbol{b}}^{(\ell)}$.
2) Compute the new probability estimates: $\hat\mu_{ij}^{(\ell)} = \exp(\hat\zeta_{ij}^{(\ell)})/(1+\exp(\hat\zeta_{ij}^{(\ell)}))$.
3) Obtain $\hat\sigma_A^{2(\ell)}$, $\hat\sigma_B^{2(\ell)}$, and $\hat\phi^{(\ell)}$ using the trace approximation. The trace approximation for $\hat\sigma_A^{2(\ell)}$ uses $\nu$ from (3.25) and has the following form: $\|\hat{\boldsymbol{a}}^{(\ell)}\|^2/(R-\nu_A^{(\ell)})$. Similarly $\hat\sigma_B^{2(\ell)} = \|\hat{\boldsymbol{b}}^{(\ell)}\|^2/(C-\nu_B^{(\ell)})$ and

$$\hat\phi^{(\ell)} = \frac{\sum_{ij} Z_{ij}\hat\mu_{ij}^{(\ell)}(1-\hat\mu_{ij}^{(\ell)})\big(z_{ij}^{(\ell)} - x_{ij}^{\mathsf{T}}\hat\beta^{(\ell)} - \hat a_i^{(\ell)} - \hat b_j^{(\ell)}\big)^2}{N - p - (R-\nu_A^{(\ell)}) - (C-\nu_B^{(\ell)})}.$$

4) Recompute the weights with new estimates: $\widehat{W}_{ij}^{(\ell)} = \hat\mu_{ij}^{(\ell)}(1-\hat\mu_{ij}^{(\ell)})/\hat\phi^{(\ell)}$.

We iterate until

$$\frac{\|\hat\zeta^{(\ell)} - \hat\zeta^{(\ell-1)}\|^2}{\|\hat\zeta^{(\ell-1)}\|^2} < \epsilon$$

and then deliver $\hat\beta^{(\infty)} = \hat\beta^{(\ell)}$, $\hat{\boldsymbol{a}}^{(\infty)} = \hat{\boldsymbol{a}}^{(\ell)}$, $\hat{\boldsymbol{b}}^{(\infty)} = \hat{\boldsymbol{b}}^{(\ell)}$, $\hat\sigma_A^{2(\infty)} = \hat\sigma_A^{2(\ell)}$, $\hat\sigma_B^{2(\infty)} = \hat\sigma_B^{2(\ell)}$ and $\hat\phi^{(\infty)} = \hat\phi^{(\ell)}$. Our generalized linear mixed model coefficient estimate is then $\hat\beta_{\mathrm{GLMM}} =$

$\hat{\beta}^{(\infty)}$. Once we obtain the weights and parameter estimates at convergence we can compute $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLMM}})$ by backfitting on each column of $\mathcal{X}$ using $\hat{\sigma}_A^{2(\infty)}$, $\hat{\sigma}_B^{2(\infty)}$, $\widehat{\mathcal{W}}^{(\infty)}$ at (3.24). As a consequence of Lemma 3.4.1, if $\mathcal{X}$ contains the intercept the penalized least squares problem in (3.12) is equivalent to solving a constrained penalized least squares problem with constraint $\sum_{i=1}^R a_i = 0$ and $\sum_{j=1}^C b_j = 0$. We can make the algorithm more efficient if we use centered operators (see section 2.3.3) $\tilde{\mathcal{S}}_A^{(k)}$ and $\tilde{\mathcal{S}}_B^{(k)}$ in (3.23) instead of $\mathcal{S}_A^{(k)}$ and $\mathcal{S}_B^{(k)}$.

Our numerical results in the Section 3.5 use $\epsilon = 10^{-8}$. Open-source R code at `https://github.com/G28Sw/scalable-logistic-regression-with-crossed-random-effects` does these computations.

## 3.5 Timing and accuracy comparisons

In this section we simulate the data from a crossed random effects model with a binary response. A simulated setting with known $\beta$ lets us compare accuracy of the methods. We also can verify linear costs for backfitting and superlinear costs for a state of the art algorithm, glmer from Bates et al. (2015). We assume that the observation pattern follows the probability model from Section 3.2. We look at timings and we also consider accuracy. These experiments were carried out in R language (R Core Team (2015)) on a computer with the macOS operating system, 16 GB of memory and an Intel i7 processor.

### 3.5.1 Timings

To study how the computation time varies with $N$ we generated data over a range of sample sizes using $\rho = \kappa = 0.56$. Our predictors were $x_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ in seven dimensions, plus an intercept, making $p = 8$. The covariance matrix had $\Sigma_{k\ell} = \gamma^{|k-\ell|}$ with autoregression parameter $\gamma = 1/2$. We used two different set of $\beta$:

a) $\beta_\ell = 0$ for $\ell = 1, \ldots, 7$ and $\beta_0 = -2$, and

b) $\beta_\ell = -2 + 0.5 \times \ell$ for $\ell = 1, \ldots, 7$ and $\beta_0 = -2$.

The choice of negative $\beta_0$ was made to simulate a setting where $Y_{ij} = 1$ is a somewhat

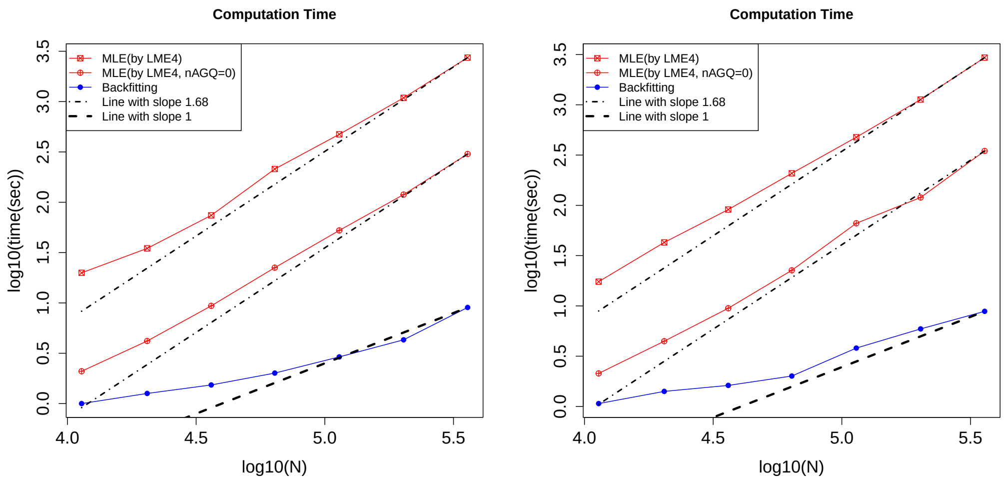



Figure 3.1: Time versus $N$ at $(\rho, \kappa) = (0.56, 0.56)$ for 0 (other than intercept) $\beta$ and trending $\beta$. The left panel plots the computational time for logistic regression when true $\beta_0 = -2$ and rest of the $\beta$'s are zero. The right panel has trending $\beta$. The cost for glmer is roughly $O(N^{\max(3\rho, 3\kappa)})$. Costs for backfitting is $O(N)$.

rare event. Some consequences for naive logistic regression of nonzero $\beta$ are described in Section 3.5.3. We took $\sigma_A = 0.8$ and $\sigma_B = 0.4$. The choice of the variance components was made based on estimates on real data. High value of $\max(\rho, \kappa)$ would lead to more computational time for glmer from Bates et al. (2015). Under the sparsity constraint, we chose to work with $(\rho, \kappa) = (0.56, 0.56)$.

We included our backfit iteratively reweighted least squares algorithm and two versions of the glmer R code with maximum number of function evaluations set at 1000. One version of glmer is the default using 1 Gaussian quadrature point. The other uses no such points and is obtained by calling glmer with the option `nAGQ = 0`. Not surprisingly, we will see that this version of glmer is faster than the default, but less accurate.

Theory suggests the computation time for glmer should be of the order $N^{\max(3\rho, 3\kappa)}$. So we expect the order of computation to be $N^{1.68}$ in this case, whereas Schall backfitting with the trace approximation should have a cost that is linear in $N$.

Figure 3.1 shows our timings for two choices of $\beta$. We see that for the largest values of $N$ in our simulation, the cost of glmer starts to follow a trend close to $N^{1.68}$.

As expected the default algorithm is slower. It appears to be slower by a constant factor. The cost of backfitting appears to grow more slowly than $O(N)$ over the given range but we believe this is because of startup costs and we see that by the end of the range of $N$ values that cost is growing nearly proportionally to $N$ as expected.

For both choices of $\beta$, backfitting took $10-12$ iterations to converge for the smaller values of $S$ and $8-9$ iterations for the larger ones. Our convergence criterion was a relative change of $10^{-8}$ as described in Section 3.4. At the largest values of $N$, the timings follow the same asymptotic rates, approximately $O(N^{1.68})$ and $O(N)$, as we saw for individual iterations. The costs tend to approach these asymptotes from above. That could be startup costs or it might arise because the number of iterations tends to decrease with $N$.

### 3.5.2 Accuracy

It is also important to compare the accuracy of these algorithms. We compute estimates of $\beta$ using logistic regression, clubbed backfitting, glmer from Bates et al. (2015) and compute the MSE.

We ran 100 replicates for each value of $S$ and then computed the MSE of the estimators $\hat{\beta}$, $\hat{\sigma}_A$ and $\hat{\sigma}_B$. For accuracy we chose $S$ to get $\log_{10}(N)$ in the range from 3 to 4.5 approximately for glmer algorithms. Because naive logistic regression and backfitting are $O(N)$ we studied them over a range up to $\log_{10}(N) \approx 5.5$.

Figure 3.2 handles case a) above with all $\beta_\ell = 0$ except for the intercept. The left panel shows the MSE for all $\beta_\ell$ except the intercept. The right panel shows the intercept. The intercept error in naive logistic regression does not decrease with $N$ although the error for the other coefficients does. The RMSE for non-intercept components is close to a reference line parallel to $O(1/N)$ and this holds for backfitting, naive logistic regression and both glmer algorithms. The RMSE for the intercept follows close to $O(1/\sqrt{N})$ for backfitting and glmer with `nAGQ=0`. The RMSE for default glmer appears to get a rate in between $O(1/\sqrt{N})$ and $O(1/N)$. We had expected the intercept coefficient to show slow convergence based on its partial confounding with the random effects described in the introduction. If the true MSE rate for the

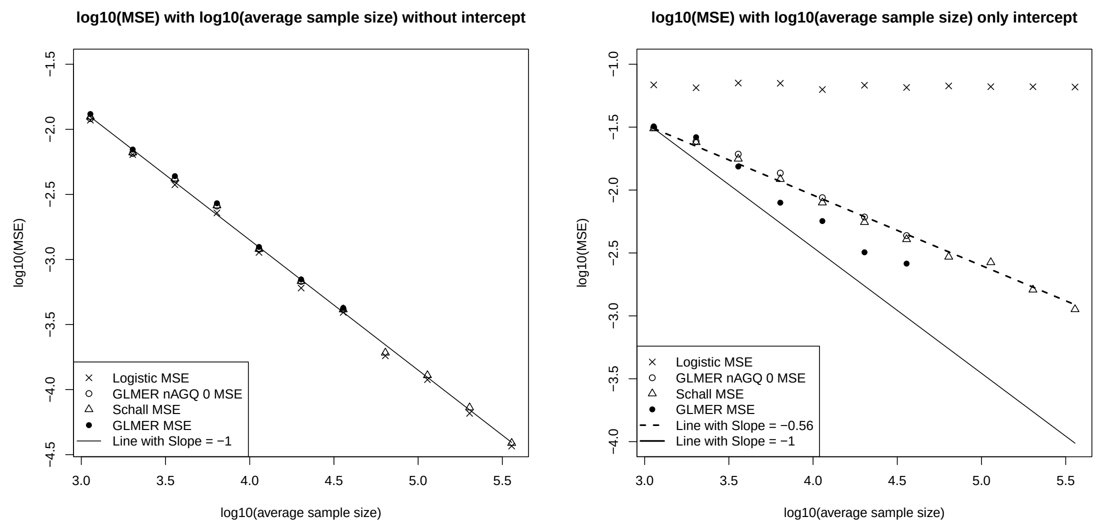



Figure 3.2: MSE of $\hat{\beta}$ vs average sample size at $(\rho, \kappa) = (0.56, 0.56)$ for 0 (other than intercept) $\beta$. The left panel plots the MSE of non intercept $\beta$ with true value being zero. The right panel plots the MSE for the intercept with true intercept being $-2$. MSE for glmer, backfitting and Schall seems to scale at $O(1/N)$ for non intercept regression coefficients. MSE for intercept seems to scale at a slower rate compared to $O(1/N)$, whereas logistic shows no improvement with increasing problem size.

intercept really is $o(N^{-1/2})$ for default glmer, then we are unable to explain that.

Figure 3.3 shows case b) with nonzero $\beta_\ell$. Plain logistic regression appears to be inconsistent for both intercept and non-intercept parameters. Once again backfitting is less accurate than the default glmer but more accurate than the other glmer choice. The default glmer has an RMSE close to $O(N^{-1})$ for non-intercept parameters. Backfitting and the faster glmer show evidence of having a worse than $O(N^{-1})$ convergence rate for non-intercept terms.

### 3.5.3 About naive logistic regression

This section provides an explanation for some of the biases we see for naive logistic regression. This analysis preceded the simulations and influenced our choice of simulation settings.

Suppose we have a mixed effects model with one random effect (no crossing). The

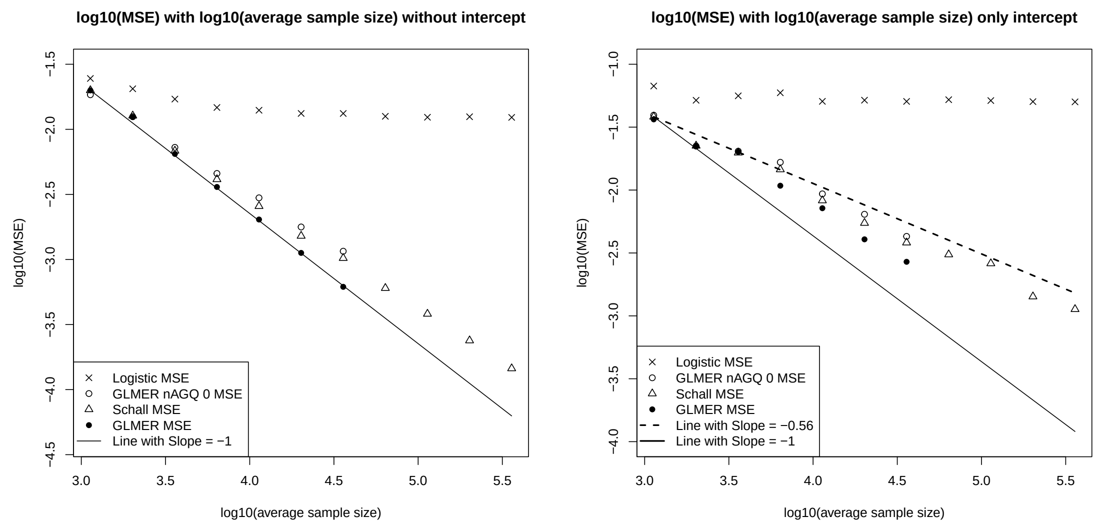



Figure 3.3: MSE of $\hat{\beta}$ vs average sample size at $(\rho, \kappa) = (0.56, 0.56)$ for trending $\beta$. The left panel plots the MSE of non intercept $\beta$. The right panel plots the MSE for the intercept with true intercept being $-2$. MSE for glmer and backfitting seems to scale at $\approx O(1/N)$ whereas logistic shows no improvement with increasing problem size. MSE for intercept seems to scale at a slower rate compared to $O(1/N)$.

responses $Y_{ij}$ satisfy $\Pr(Y_{ij} = 1 \mid x_{ij}, a_i) = \pi(x_{ij}^\mathsf{T}\beta + a_i)$ for $i = 1, \dots, R$. Since the columns are not linked we let $j = 1, \dots, N_{i\bullet}$ The random effects are $a_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The $x_{ij}$ are fixed either by design or by running a conditional analysis. We will suppose at first that $\sigma$ is 'small'.

If an analyst ignores the random effects and uses likelihood

$$\tilde{L} = \prod_{i=1}^{R} \prod_{j=1}^{N_{i\bullet}} \pi(x_{ij}^\mathsf{T}\beta)^{Y_{ij}}(1 - \pi(x_{ij}^\mathsf{T}\beta))^{1-Y_{ij}}$$

then their log likelihood function is

$$\tilde{\ell} = \sum_{i=1}^{R} \sum_{j=1}^{N_{i\bullet}} Y_{ij}x_{ij}^\mathsf{T}\beta - \log(1 + \exp(x_{ij}^\mathsf{T}\beta)).$$

Their naive score function is

$$\frac{\partial \tilde{\ell}}{\partial \beta} = \sum_{i=1}^{R} \sum_{j=1}^{N_{i\bullet}} \big(Y_{ij} - \pi(x_{ij}^\mathsf{T}\beta)\big)x_{ij}.$$

If the population mean of the score at the true parameter value is not zero, then eventually a statistical test based on the score would reject the true parameter value and we would expect an inconsistent MLE.

Randomness enters that expectation via $Y_{ij}$ which includes randomness from $a_i$. Now

$$\mathbb{E}\Big(\frac{\partial \tilde{\ell}}{\partial \beta}\Big) = \sum_{i=1}^{R} \sum_{j=1}^{N_{i\bullet}} \Big( \int_{-\infty}^{\infty} \frac{\exp(-a_i^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} \pi(x_{ij}^\mathsf{T}\beta + a_i)\,\mathrm{d}a_i - \pi(x_{ij}^\mathsf{T}\beta)\Big)x_{ij}.$$

For small $\sigma$, we anticipate that small $a$ will dominate, so we make a Taylor approximation

$$\pi(x_{ij}^\mathsf{T}\beta + a_i) \approx \pi(x_{ij}^\mathsf{T}\beta) + a_i\pi'(x_{ij}^\mathsf{T}\beta) + \frac{1}{2}a_i^2\pi''(x_{ij}^\mathsf{T}\beta)$$

and using this approximation

$$\mathbb{E}\Big(\frac{\partial \tilde{\ell}}{\partial \beta}\Big) \approx \frac{\sigma^2}{2} \sum_{i=1}^{R} \sum_{j=1}^{N_{i\bullet}} \pi''(x_{ij}^\mathsf{T}\beta)x_{ij}. \tag{3.38}$$

In light of equation (3.38), the analyst needs each variable $x_{ij\ell}$ to be nearly orthogonal to $\pi''(x_{ij}^\mathsf{T}\beta)$ with the true unknown $\beta$. This is virtually impossible to arrange.

In one special case we could have $\beta$ equal to zero apart from a negative intercept component. Then $\pi''(x_{ij}^\mathsf{T}\beta)$ is a positive constant and for any centered variables $x_{ij\ell}$ the $\ell$'th component of (3.38) will vanish. Of course the intercept variable is all ones and cannot be centered and it will therefore have a non vanishing score component and hence a bias.

## 3.6 Stitch Fix data example

Our algorithm is illustrated using data from Stitch Fix. In section 2.6 of the second chapter, a different data set from Stitch Fix was used in which the response was not binary. The binary response of interest in the current dataset was whether the customer thought an item was a top rated fit. There were $N = 5{,}000{,}000$ ratings from 744,482 clients on 3,547 items. We want to treat both clients and items as random effects. The data are not dominated by a single row or column because $\max_i N_{i\bullet}/N \doteq 1.24 \times 10^{-5}$ and $\max_j N_{\bullet j}/N \doteq 0.0276$. The data are sparse because $N/(RC) \doteq 0.0018$.

One of the predictors was the primary material of which an item was made with 20 levels such as 'linen' or 'wool'. The material is a property of the item or to put it another way, the item factor is nested within the levels of the material predictor. The other predictors for this response were 'client dress size', 'client chest size', 'client fit profile', 'item fit profile', 'client edgy' and 'client boho'. For instance the fit profiles had these levels: 'Fitted', 'Loose', 'Oversize', 'Straight' and 'Tight' as well as 'Missing'. Some clients indicated that they like 'edgy' styles and some that they like 'boho' styles.

In a business setting one would fit and compare a wide variety of different binary regression models in order to understand the data. Our purpose here is to understand large scale generalized linear mixed effects models and so we choose just one model for illustration. That model has

$$\begin{aligned}
\operatorname{logit}(\Pr(Y_{ij} = 1 \,|\, a_i, b_j)) = \beta_0 &+ \beta_1 \text{client fit profile}_i + \beta_2 \mathbb{I}\{\text{client edgy}\}_i \\
&+ \beta_3 \mathbb{I}\{\text{client boho}\}_i + \beta_4 \text{client chest size}_i \\
&+ \beta_5 \text{client dress size}_i + \beta_6 \text{material}_j \\
&+ \beta_7 \text{item fit profile}_j + a_i + b_j.
\end{aligned}$$

The categorical variables were incorporated using a one-hot encoding with a binary indicator variable for each level other than the most common level. The model has $p = 34$ parameters. Our backfitting algorithm took 14 iterations to convergence with a tolerance of $10^{-8}$.

Let $\hat{\beta}_{\mathrm{LR}}$ and $\hat{\beta}_{\mathrm{GLMM}}$ be the logistic regression and GLMM estimates of $\beta$ obtained by clubbed backfitting, respectively. We can compute their corresponding variance estimates $\widehat{\mathrm{cov}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}})$ and $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GLMM}})$. We can also find $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})$, the variance under our GLMM model of the $\hat{\beta}_{\mathrm{LR}}$. The estimated coefficients $\hat{\beta}_{\mathrm{LR}}$, $\hat{\beta}_{\mathrm{GLMM}}$ and their standard errors are presented in Appendix B.1.6.

We can quantify the naivete of logistic regression, coefficient by coefficient, via the ratio $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR},\ell})/\widehat{\mathrm{cov}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR},\ell})$. The left panel of Figure 3.4 plots these values. They range from 2.72 to 1467.05 and can be interpreted as factors by which logistic regression naively overestimates its sample size. The largest and second largest ratios are for material indicators corresponding to 'Modal' and 'Rayon', respectively. Not only is logistic regression estimating $\beta$ with significant bias, it greatly underestimates its sampling uncertainty. We can also identify the linear combination of $\hat{\beta}_{\mathrm{LR}}$ for which LR is most naive. We maximize the ratio

$$\boldsymbol{x}^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})\boldsymbol{x}/\boldsymbol{x}^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}})\boldsymbol{x}$$

over $\boldsymbol{x} \neq 0$. The resulting maximal ratio is the largest eigenvalue of

$$\widehat{\mathrm{cov}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}})^{-1}\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})$$

and it is about 1507 for the Stitch Fix data.

We can also quantify the inefficiency of logistic regression, coefficient by coefficient, via the ratio $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR},\ell})/\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GLMM},\ell})$. The right panel in Figure 3.4 plots these values. They range from just over 1 to 11.42 and can be interpreted as factors by which using logistic regression reduces the effective sample size. The two largest inefficiencies corresponds to item material 'Rayon' and 'Modal' respectively. The most inefficient linear combination of $\hat{\beta}$ reaches a variance ratio of 15.50.

Figure 3.5 plots inefficiency versus naivete for the 34 coefficients in our logistic regression model. The very worst coefficients by one measure tend to be worst by the other as well.

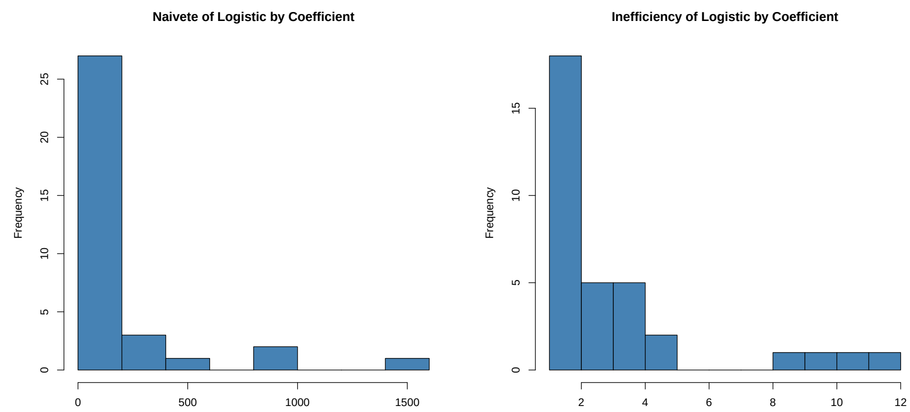



Figure 3.4: The left panel is a histogram of naivete of logistic regression quantified by $\widehat{\text{cov}}_{\text{GLMM}}(\hat{\beta}_{\text{LR},\ell})/\widehat{\text{cov}}_{\text{LR}}(\hat{\beta}_{\text{LR},\ell})$ for coefficients $\beta_\ell$ in the Stitch Fix data. The right panel is the inefficiency $\widehat{\text{cov}}_{\text{GLMM}}(\hat{\beta}_{\text{LR},\ell})/\widehat{\text{cov}}_{\text{GLMM}}(\hat{\beta}_{\text{GLMM},\ell})$.

## 3.7 Discussion

The most critical inference problems are at the margins where we can just barely resolve real effects from noise. Ignoring a crossed random effects correlation structure can be very naive, underestimating the true sampling variance of a parameter by several hundred-fold. Data analysts would then be overconfident about the extent to which their findings generalize to future data. This can be very troublesome, because even in large data sets, economically valuable effects can be surprisingly small compared to their sampling errors. Lewis and Rao (2015) find that the $R^2$ in a highly profitable project could be as small as 0.0000054. Kohavi et al. (2009) report that the Bing search engine had been so extensively optimized that only one third of executed experiments improved the quantity they were designed to improve. Even with very large data sets reproducibility of findings cannot be taken for granted. This may lead to very unreliable findings lacking reproducibility and even internal validity much less external validity.

The crossed random effects setting makes many of our standard methods going back to Searle et al. (2006) and Henderson (1953) computationally prohibitive. A cost that grows faster than $N^{3/2}$ is not possible for somebody with big data.

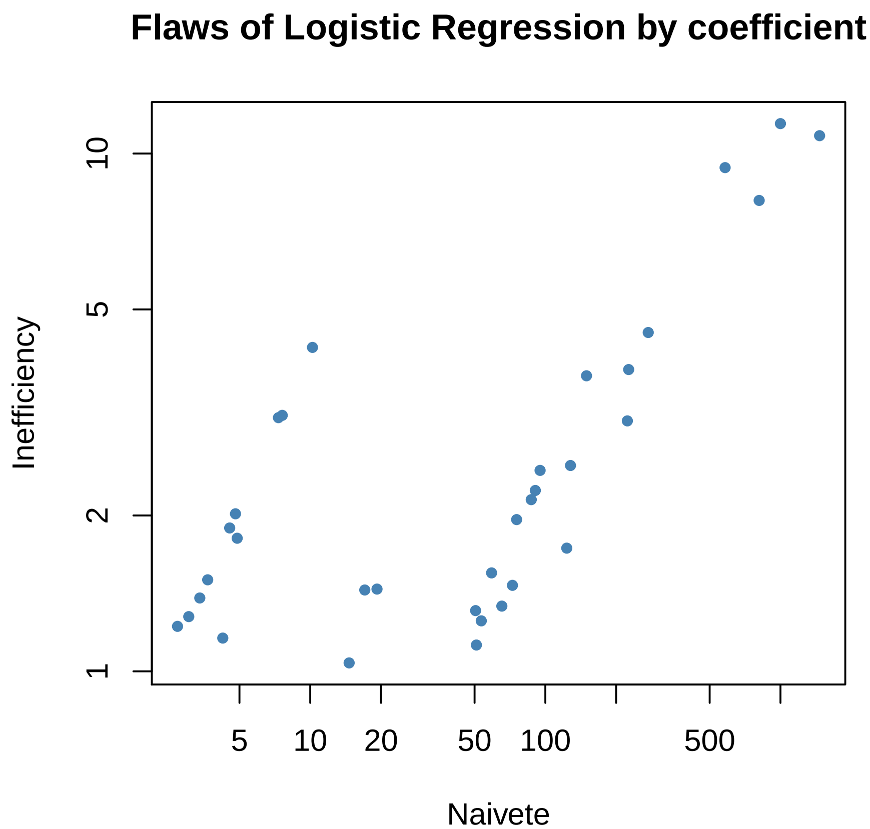



Figure 3.5: Inefficiency vs naivete for logistic regression coefficients in the Stitch Fix data.

There has been recent progress in speeding up these computations including frequentist approaches for least squares problems with crossed effects in Gao (2017), Gao and Owen (2020), and Ghosh et al. (2022) as well as Gibbs sampling in Papaspiliopoulos et al. (2020). Our contribution is to extend the backfitting approach to logistic regression. A critical step was our clubbing of $\hat{\beta}$ with one random effect at a time. This can be viewed as a frequentist counterpart to the collapsed Gibbs sampler in Papaspiliopoulos et al. (2020). There remains a theoretical gap between what the Bayesian and frequentist algorithms accomplish in practice and what can be proved about them.

Our goal remains partly achieved. There are several places in generalized linear mixed effects models where usual algorithms impose a cost that is $\gg N^{3/2}$ per iteration making them completely infeasible. We have removed all of those costs so that iterations cost $O(N)$ each. In our numerical examples we see a total cost of $O(N)$ so that the number of iterations has scaled as $O(1)$. What remains is to find more explanations of this phenomenon and sufficient conditions to bound the iteration count.

# Chapter 4

# Convergence of Collapsed Gibbs for Crossed Random Effect Models

## 4.1 Introduction

In this chapter, we study the convergence rate of a collapsed Gibbs sampler for crossed random effects models. In applications of crossed random effects models (such as recommender systems), the missing data phenomenon can be quite common. Our analysis applies to models with missingness mechanism (see Sect. 4.1.1 for a detailed description), which is a new feature in the Bayesian approach to this problem. Another feature of our analysis is that it applies to data with unbalanced levels, which substantially relaxes the balancedness condition required for the analysis in Papaspiliopoulos et al. (2020).

Crossed random effects models are regression models that relate a response variable to several categorical input variables. They are widely used in modeling enormous data sets coming from electronic commerce (such as recommender systems) and appear in literature under different names like cross-classified data, variance component models (Gelman (2005); Searle et al. (2006)).

In this chapter, we consider an intercept only crossed random effects model with two factors:

$$Y_{ij} = a^{(0)} + a_i^{(1)} + a_j^{(2)} + e_{ij}, \quad 1 \leq i \leq R, \quad 1 \leq j \leq C. \tag{4.1}$$

This facilitates the theoretical analysis. To distinguish this from model with continuous covariates, we use a slightly different notation than introduced before. Whereas the theory of this chapter refers to (4.1), adapted appropriately, our findings and their methodological implications are likely to apply to more general linear mixed models that include continuous covariates and probability distributions that are not Gaussian. This is attested in numerical experiments in Papaspiliopoulos et al. (2021). Here, $Y_{ij}$ is the response variable, $a^{(0)}$ is the global mean, $a_i^{(1)}$ and $a_j^{(2)}$ represent the two random effects, and $e_{ij}$ is the error. The first random effect has $R$ levels, and the second random effect has $C$ levels. We assume that $a_i^{(1)} \sim \mathcal{N}(0, \sigma_1^2)$ for each $1 \leq i \leq R$, $a_j^{(2)} \sim \mathcal{N}(0, \sigma_2^2)$ for each $1 \leq j \leq C$, and $e_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ for every $1 \leq i \leq R, 1 \leq j \leq C$, where $\sigma_1, \sigma_2, \sigma_E > 0$ (these random variables are assumed to be all independent). We denote the corresponding precision parameters by $\tau_1, \tau_2$ and $\tau_E$. We also assume a flat prior $p(a^{(0)}) \propto 1$ for the global mean.

In our analysis, we assume that the observation $Y_{ij}$ for some levels $(i, j)$ (with $1 \leq i \leq R$, $1 \leq j \leq C$) can be missing. The missingness pattern is described by a $\{0, 1\}$-valued matrix $Z$, which we describe in detail in Sect. 4.1.1.

For recommender system applications, we let $i$ denote the $i$th customer for $1 \leq i \leq R$, let $j$ denote the $j$th product for $1 \leq j \leq C$, and let $Y_{ij}$ be the rating of the $i$th customer for the $j$th product. Then $a_i^{(1)}$ is the random effect from the $i$th customer, and $a_j^{(2)}$ is the random effect from the $j$th product.

In this chapter, we are interested in scalable inference methods for crossed random effects models. Here, scalability means that the time complexity of the inference method should be at most linear in the number of observations and the number of parameters. This has been a great challenge for large data sets. For example, Gao and Owen (2020) showed that evaluating the likelihood of the model once already has complexity at least of order $N^{\frac{3}{2}}$ when there are $N$ observations. The lmer function in R package Bates et al. (2015) has a cost that grows like $N^{3/2}$ and Bates et al. (2015)

removed the MCMC option from the package because it was considered unreliable. As for MCMC algorithms, Gao and Owen (2017) showed that the Gibbs sampler takes order $N^{\frac{1}{2}}$ steps to mix, leading to order $N^{\frac{3}{2}}$ complexity (as the time cost for each iteration is of order $N$).

Liu (1994) introduced collapsing in Monte Carlo computations. Recently, Papaspiliopoulos et al. (2020) proposed a collapsed Gibbs sampler for crossed random effects models. They showed that under "balanced levels" condition the collapsed Gibbs sampler mixes in a constant number of steps in certain settings, and is therefore scalable. In the recommender system example, "balanced levels" condition reduces to each customer rating equal number of products and each item being rated by equal number of customers. We will review the collapsed Gibbs sampler and the balanced levels condition in Sect. 4.1.2. Further discussions on the application of collapsed Gibbs sampling to Bayesian hierarchical models are in Papaspiliopoulos et al. (2021). A different approach has been undertaken by Ghosh et al. (2022). We proposed an iterative algorithm based on backfitting and showed that in some regimes with unbalanced levels, the algorithm is scalable. Later we extended the algorithm to a generalized linear mixed model for logistic regression in Ghosh et al. (2021). We proposed *clubbed backfitting*, a two-step iterative procedure. It "clubs" the fixed effect and one of the random effects in each of the two steps.

The balanced levels condition is rarely satisfied in real applications such as recommender systems, unless we have a designed experiment for the data. Therefore, it is important to look for scalable inference methods that work beyond the balanced levels condition. In this chapter, we show that the collapsed Gibbs sampler is scalable in many natural scenarios beyond the balanced levels condition. We relax the balanced levels condition to balanced in expectation, then approximately balanced in expectation with some degree of inhomogeneity, then further relax it to arbitrary inhomogeneity assuming "almost balancedness". This motivates the study of this chapter. As our approach, we analyse the $L_2$ norm of a certain autoregression matrix, which widens the range of scenarios where our result applies compared to the backfitting-based algorithm in the second chapter in a frequentist paradigm. Moreover, when the row sums and column sums are "almost balanced" in expectation (see

Theorem 4.1.3 below for details), our proof strategy allows us to incorporate an *arbitrary* inhomogeneity level (measured by $\Upsilon$ in Theorem 4.1.3) for missingness pattern. The concrete results will be presented in Sect. 4.1.3.

### 4.1.1 Missingness mechanism

In this subsection, we describe the missingness mechanism of the model. The mechanism is similar to that of section 2.4.2 in the second chapter, but is new in the Bayesian setting.

We let $Z_{ij} \in \{0, 1\}$ for every $1 \leq i \leq R$ and $1 \leq j \leq C$, with $Z_{ij} = 1$ when $Y_{ij}$ is observed and $Z_{ij} = 0$ when $Y_{ij}$ is not observed. Now the $R \times C$ observation matrix $Z \in \{0, 1\}^{R\times C}$ formed by the elements $Z_{ij}$ has $N_{i\cdot} = \sum_{j=1}^{C} Z_{ij}$ observations in the $i$th row, and $N_{\cdot j} = \sum_{i=1}^{R} Z_{ij}$ observations in the $j$th column. The total number of observations is denoted by $N = \sum_{i=1}^{R}\sum_{j=1}^{C} Z_{ij}$.

In the following, we view $Z_{ij}$ as being non-random (i.e., we condition on the actual observations) when we perform collapsed Gibbs sampling. When studying convergence properties of the collapsed Gibbs sampler, we treat $Z_{ij}$ as independent random variables, with $Z_{ij}$ following the Bernoulli distribution with success probability $p_{ij}$ for every $1 \leq i \leq R, 1 \leq j \leq C$.

We mention by passing that our missingness mechanism does not incorporate informative missing, which can appear in real applications of crossed random effects models. However, as the analysis is already challenging under the current missingness mechanism, we leave the handling of informative missing to future works.

### 4.1.2 Collapsed Gibbs sampling

Papaspiliopoulos et al. (2020) proposed a collapsed Gibbs sampler for sampling from the posterior of the crossed random effects model. They showed that under the balanced levels condition, the sampler is scalable. Here, a data set has balanced levels if for each factor, the number of observations for each level of the factor is the same (the number can vary with the factor). Via the notations from Sect. 4.1.1, this condition can be expressed as $N_{i\cdot} = \frac{N}{R}$ for every $1 \leq i \leq R$, and $N_{\cdot j} = \frac{N}{C}$ for every

$1 \leq j \leq C$.

Below we introduce the details of the collapsed Gibbs sampler, following Papaspiliopoulos et al. (2020). We denote by $\boldsymbol{a}^{(1)} = (a_1^{(1)}, \cdots, a_R^{(1)})$, $\boldsymbol{a}^{(2)} = (a_1^{(2)}, \cdots, a_C^{(2)})$. We also denote by $y$ the observations. The posterior distribution of $(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)})$ is denoted by $\mathcal{L}(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)} \mid y)$.

Each step of the collapsed Gibbs sampler involves integrating out the global mean $a^{(0)}$ and then sampling in blocks the levels of $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$. In practice, for the first block, we sample first from $\mathcal{L}(a^{(0)}|y, \boldsymbol{a}^{(2)})$ and then from $\mathcal{L}(a_i^{(1)}|\cdot)$ for $i = 1, \cdots, R$; for the second block, we sample first from $\mathcal{L}(a^{(0)}|y, \boldsymbol{a}^{(1)})$ and then from $\mathcal{L}(a_j^{(2)}|\cdot)$ for $j = 1, \cdots, C$. Here, $\mathcal{L}(X|\cdot)$ denotes the full conditional distribution of $X$ conditionally on the rest variables.

We list the involved conditional distributions below, following Papaspiliopoulos et al. (2020).

*Proposition* 4.1.1 (Papaspiliopoulos et al. (2020)). Let $\tilde{y}_j^{(k)}$ be the weighted average of all observations whose level on factor $k$ is $j$ for $k \in \{1, 2\}$. For any $i = 1, \cdots, R$,

$$\mathcal{L}(a_i^{(1)}|\cdot) = \mathcal{N}\{\frac{N_{i\cdot}\sigma_1^2}{N_{i\cdot}\sigma_1^2 + \sigma_E^2}(\tilde{y}_i^{(1)} - a^{(0)} - \frac{\sum_{j=1}^{C} a_j^{(2)} Z_{i,j}}{N_{i\cdot}}), \frac{\sigma_E^2\sigma_1^2}{N_{i\cdot}\sigma_1^2 + \sigma_E^2}\}; \tag{4.2}$$

for any $j = 1, \cdots, C$,

$$\mathcal{L}(a_j^{(2)}|\cdot) = \mathcal{N}\{\frac{N_{\cdot j}\sigma_2^2}{N_{\cdot j}\sigma_2^2 + \sigma_E^2}(\tilde{y}_j^{(2)} - a^{(0)} - \frac{\sum_{i=1}^{R} a_i^{(1)} Z_{i,j}}{N_{\cdot j}}), \frac{\sigma_E^2\sigma_2^2}{N_{\cdot j}\sigma_2^2 + \sigma_E^2}\}. \tag{4.3}$$

For any $i = 1, \cdots, R$, we let $s_i^{(1)} = \frac{N_{i\cdot}\sigma_1^2}{N_{i\cdot}\sigma_1^2 + \sigma_E^2}$; for any $j = 1, \cdots, C$, we let $s_j^{(2)} = \frac{N_{\cdot j}\sigma_2^2}{N_{\cdot j}\sigma_2^2 + \sigma_E^2}$. Then we have

$$\mathcal{L}(a^{(0)}|y, \boldsymbol{a}^{(2)}) = \mathcal{N}(\frac{1}{\sum_{i=1}^{R} s_i^{(1)}} \sum_{i=1}^{R} s_i^{(1)}(\tilde{y}_i^{(1)} - \frac{\sum_{j=1}^{C} a_j^{(2)} Z_{i,j}}{N_{i\cdot}}), \frac{\sigma_1^2}{\sum_{i=1}^{R} s_i^{(1)}}), \tag{4.4}$$

$$\mathcal{L}(a^{(0)}|y, \boldsymbol{a}^{(1)}) = \mathcal{N}(\frac{1}{\sum_{j=1}^{C} s_j^{(2)}} \sum_{j=1}^{C} s_j^{(2)}(\tilde{y}_j^{(2)} - \frac{\sum_{i=1}^{R} a_i^{(1)} Z_{i,j}}{N_{\cdot j}}), \frac{\sigma_2^2}{\sum_{j=1}^{C} s_j^{(2)}}). \tag{4.5}$$

### 4.1.3 Main results

In this subsection, we describe the main results of this chapter. We measure the convergence rate of the collapsed Gibbs sampler by the relaxation time, denoted by $t_{rel}$. We recall that the relaxation time of a Markov chain is defined as the reciprocal of its spectral gap, where the spectral gap is the difference between 1 and the geometric rate of convergence of the Markov chain. The readers are referred to Papaspiliopoulos et al. (2020); Roberts and Sahu (1997) for further details.

We show that in three regimes, the collapsed Gibbs sampler mixes in a *constant* number of steps (independent of the problem size) in terms of relaxation time with high probability (with respect to the randomness in the observation matrix $Z$). Therefore, the collapsed Gibbs sampler is scalable for these regimes.

In our analysis, we assume that $Z_{ij}$ follows independent Bernoulli distribution with success probability $p_{ij}$ for each $1 \leq i \leq R, 1 \leq j \leq C$. Our problem size will be indexed by $S$, and we let $S \to \infty$ in the asymptotic analysis. The total sum $N$ of the matrix $Z$ satisfies $\mathbb{E}[N] \asymp S$, where $A \asymp B$ means $k_1 A \leq B \leq k_2 A$ for some absolute constants $k_1, k_2 > 0$. We also assume the scaling $R = \lceil S^\rho \rceil, C = \lceil S^\kappa \rceil$ for $\rho, \kappa \in (0,1)$. We are interested in regimes where the observations are sparse, that is, $\rho + \kappa > 1$ (so that $S \ll RC$). Our analysis holds for $\rho, \kappa \in (0,1)$ such that

$$\rho + \frac{1}{2}\kappa < 1, \quad \kappa + \frac{1}{2}\rho < 1. \tag{4.6}$$

This is wider than the regime covered by the analysis of Ghosh et al. (2022). The range of $\rho, \kappa$ covered by our analysis is indicated in Figure 4.1. For the theoretical analysis we assume the precision parameters to be known. Empirically we observe a similar phenomenon when they are unknown and we sample precision parameters $\tau$ from the conditional distribution $\mathcal{L}(\tau \mid \boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)})$ and update $(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)})$ with the Gibbs sampler and its collapsed version, respectively.

The first result assumes that $Z_{ij}$ are i.i.d. This corresponds to the situation where the missingness pattern is homogeneous across different cells $(i,j)$ for $1 \leq i \leq R, 1 \leq j \leq C$. This missingness mechanism is known as missing completely at random (MCAR) (see Little and Rubin (2019)).

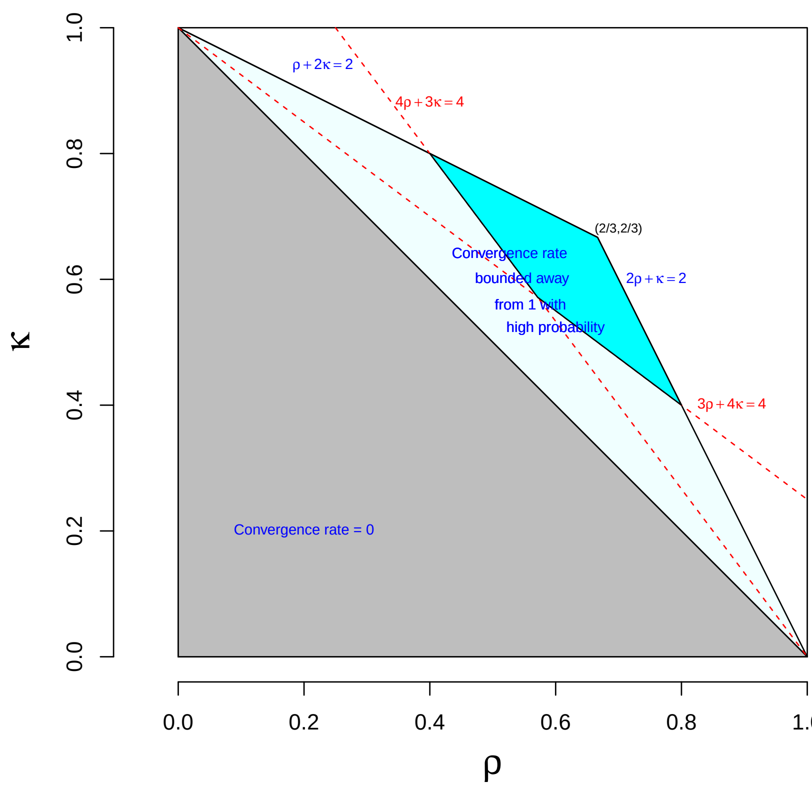



Figure 4.1: The area shaded with light blue and cyan is the domain of interest for Theorems 4.1.1-4.1.3. Ghosh et al. (2022) proved convergence for the backfitting procedure for $(\rho, \kappa)$ in the light blue region.

*Theorem* 4.1.1. Suppose that $Z_{ij} \sim \text{Bern}(\frac{S}{RC})$ for every $1 \le i \le R$ and $1 \le j \le C$. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), and that $\sigma_1^2, \sigma_2^2, \sigma_E^2$ are independent of $S$.

Then there exists a positive constant $\xi$ (independent of $S$), such that

$$\lim_{S \to \infty} \mathbb{P}(t_{rel} \le \xi) = 1.$$

*Proof.* It follows from Proposition 4.2.1 and Theorem 4.3.1 below. □

The second result assumes that $\frac{S}{RC} \le p_{ij} \le \frac{\Upsilon S}{RC}$ with $\Upsilon \in [1, 1.52]$. In this regime, we allow inhomogeneous missingness pattern across different cells $(i, j)$ for $1 \le i \le R, 1 \le j \le C$. This missingness mechanism corresponds to missing at random (MAR) (see Little and Rubin (2019)).

*Theorem* 4.1.2. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), and that $\sigma_1^2, \sigma_2^2, \sigma_E^2$ are independent of $S$. Also assume that there exists a constant $\Upsilon \in [1, 1.52]$, such that for every $1 \le i \le R$ and $1 \le j \le C$,

$$\frac{S}{RC} \le p_{ij} \le \frac{\Upsilon S}{RC}.$$

Then there exists a positive constant $\xi$ (independent of $S$), such that

$$\lim_{S \to \infty} \mathbb{P}(t_{rel} \le \xi) = 1.$$

*Proof.* It follows from Proposition 4.2.1 and Theorem 4.3.2 below. □

*Remark* 4.1.1. Theorem 4.1.1 is a special case of Theorem 4.1.2 with $\Upsilon = 1$.

The third result assumes that the row sums and column sums of the matrix $Z$ are "almost balanced" in expectation (given by the conditions (4.7)-(4.8) below). In the recommender system application, this corresponds to the case when every customer/ product has similar behavior in expectation. With this almost balancedness condition, we can allow for an arbitrary inhomogeneity parameter $\Upsilon > 1$ (with $\frac{S}{\Upsilon RC} \le p_{ij} \le \frac{\Upsilon S}{RC}$). This missingness mechanism also corresponds to missing at random (MAR) (see Little and Rubin (2019)).

*Theorem* 4.1.3. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), and that $\sigma_1^2, \sigma_2^2, \sigma_E^2$ are independent of $S$. Also assume that there exist a series of numbers $\epsilon(S) > 0$ with $\lim_{S\to\infty} \epsilon(S) = 0$, such that for every $1 \leq i \leq R$ and $1 \leq j \leq C$,

$$\frac{S}{C}(1-\epsilon(S)) \leq \sum_{i=1}^{R} p_{ij} \leq \frac{S}{C}(1+\epsilon(S)), \tag{4.7}$$

$$\frac{S}{R}(1-\epsilon(S)) \leq \sum_{j=1}^{C} p_{ij} \leq \frac{S}{R}(1+\epsilon(S))). \tag{4.8}$$

Further assume that there exists a constant $\Upsilon \geq 1$, independent of $S$, such that for every $1 \leq i \leq R$ and $1 \leq j \leq C$,

$$\frac{S}{\Upsilon RC} \leq p_{ij} \leq \frac{\Upsilon S}{RC}.$$

Then there exists a positive constant $\xi$ (independent of $S$), such that

$$\lim_{S\to\infty} \mathbb{P}(t_{rel} \leq \xi) = 1.$$

*Proof.* It follows from Proposition 4.2.1 and Theorem 4.3.3 below. □

The proof strategy of our main results is substantially different from that used in Papaspiliopoulos et al. (2020). Specifically, the theoretical analysis of Papaspiliopoulos et al. (2020) is based on multigrid decomposition, which crucially depends on the balanced levels condition. In our proof, instead, we utilize a connection between the relaxation time of the collapsed Gibbs sampler and the autoregression matrix of the chain without performing a multigrid decomposition (see Sect. 4.2 for details). The analysis of the autoregression matrix, as given in Sect. 4.3, involves many further tools, including concentration inequalities and results from random matrix theory. Equipped with these tools, our analysis provides sufficient conditions for convergence within a finite number of steps in the three settings covered by Theorems 4.1.1–4.1.3. In comparison, the convergence results of Papaspiliopoulos et al. (2020) depend on the rate of convergence of an auxiliary Markov chain (denoted by $\rho_{aux}$ in the chapter), for which not much is known in many practical settings.

Now we describe the main contents of the following sections. Sect. 4.2 introduces our main strategy for studying the relaxation time of the collapsed Gibbs sampler with unbalanced levels. Applying this strategy to three concrete scenarios, Sect. 4.3 provides relaxation time upper bounds in such settings, which show that the collapsed Gibbs sampler is scalable in these settings. Then through numerical simulations, we show in Sect. 4.4 the scalability of the collapsed Gibbs sampler in various settings, even beyond those covered by our theoretical analysis. In Sect. 4.5, we further illustrate the scalability of the collapsed Gibbs sampler on real data from Stitch Fix. Finally, in Sect. 4.6, we present some conclusions and discussions.

## 4.2 Relaxation time and autoregression matrix

Suppose the chain $\big(\boldsymbol{a}^{(1)}(t), \boldsymbol{a}^{(2)}(t)\big)$ is a two-component Gibbs sampler alternating updates from $\mathcal{L}(\boldsymbol{a}^{(1)}|y, \boldsymbol{a}^{(2)})$ and $\mathcal{L}(\boldsymbol{a}^{(2)}|y, \boldsymbol{a}^{(1)})$. Thus, $\boldsymbol{a}^{(2)}(t)$ is marginally a Markov chain and its rate of convergence equals that of $\big(\boldsymbol{a}^{(1)}(t), \boldsymbol{a}^{(2)}(t)\big)$ (see, for instance, Roberts and Rosenthal (2001)). Let $B_1$ and $B_2$ be defined by $\mathbb{E}(\boldsymbol{a}^{(1)}|\boldsymbol{a}^{(2)}, y) = B_1\boldsymbol{a}^{(2)} + b_1$ and $\mathbb{E}(\boldsymbol{a}^{(2)}|\boldsymbol{a}^{(1)}, y) = B_2\boldsymbol{a}^{(1)} + b_2$. Then $\boldsymbol{a}^{(2)}(t)$ is a Gaussian autoregressive process with autoregression matrix $B_2B_1$. By (Roberts and Sahu, 1997, Theorem 1), to study the relaxation time, it suffices to look at the autoregression matrix.

We consider a crossed random effects model with two factors (that is, $K = 2$) as in the Introduction. Let $\lambda_A = \frac{\sigma_E^2}{\sigma_1^2}$ and $\lambda_B = \frac{\sigma_E^2}{\sigma_2^2}$. Then $s_i^{(1)} = \frac{N_{i\bullet}}{N_{i\bullet}+\lambda_A}$ and $s_j^{(2)} = \frac{N_{\bullet j}}{N_{\bullet j}+\lambda_B}$ following the definitions as in Proposition 4.1.1. Along with these, let us define two sets of weights $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ with components $w_i^{(1)} = \frac{s_i^{(1)}}{\sum_{i'} s_{i'}^{(1)}}$ and $w_j^{(2)} = \frac{s_j^{(2)}}{\sum_{j'} s_{j'}^{(2)}}$. Also we define the two diagonal matrices $D_1 = \mathrm{diag}(N_{i\bullet} + \lambda_A)$ and $D_2 = \mathrm{diag}(N_{\bullet j} + \lambda_B)$. Throughout the chapter, for any $n \in \{1, 2, \cdots\}$, we denote by $I_n$ the identity matrix of size $n$ and $\mathbf{1}_n$ the $n$-dimensional vector whose entries are all 1.

*Proposition* 4.2.1. The relaxation time $t_{rel}$ of the collapsed Gibbs sampler is characterised by the spectral radius $\rho(M)$ of the autoregression matrix $M := (I_C -$

$\boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T})M_0$ (with $M_0 := D_2^{-1}Z^\mathsf{T}(I_R - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T})D_1^{-1}Z)$ by

$$t_{rel} = \frac{1}{1-\rho(M)}. \tag{4.9}$$

*Proof.* See Appendix C.1.1. □

## 4.3 Analysis of the autoregression matrix

For the analysis below we would assume the following model. We let $Z_{ij} \sim \text{Bern}(p_{ij})$ for every $1 \le i \le R$ and $1 \le j \le C$ independently with

$$\frac{1}{\Upsilon'}\frac{S}{RC} \leqslant p_{ij} \leqslant \Upsilon\frac{S}{RC} \quad \text{for} \quad 1 \leqslant \Upsilon, \Upsilon' < \infty, R = S^\rho, C = S^\kappa. \tag{4.10}$$

That is $1/\Upsilon' \leqslant p_{ij}S^{\rho+\kappa-1} \leqslant \Upsilon$. Letting $p_{ij}$ depend on $i$ and $j$ allows the probability model to capture stylistic preferences affecting the missingness pattern in the ratings data. Note that we have ignored the issue of taking integer parts for $R$ and $C$, as it will not influence our conclusion.

For any matrix $A$, we denote by $\|A\|_2$ its spectral norm, and denote by $\rho(A)$ its spectral radius below. The meanings of variables like $\delta_1, \delta_2, \epsilon$ defined in proofs can vary across different proofs, but will remain the same within one proof.

For the three regimes discussed in Sect. 4.1.3, the corresponding results for the autoregression matrix $M$ (as given in Proposition 4.2.1) are given by Theorems 4.3.1-4.3.3 as follows. Theorems 4.1.1-4.1.3 follow from combining Theorems 4.3.1-4.3.3 and Proposition 4.2.1.

*Theorem* 4.3.1. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), and that $\sigma_1^2, \sigma_2^2, \sigma_E^2$ are independent of $S$. Also assume that $\Upsilon' = \Upsilon = 1$.

Then for any fixed $\delta > 0$, we have

$$\lim_{S\to\infty} \mathbb{P}(\|M\|_2 \le \delta) = 1.$$

*Proof.* See Sect. 4.3.2. □

*Theorem* 4.3.2. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), and that $\sigma_1^2, \sigma_2^2, \sigma_E^2$ are independent of $S$. Also assume that $\Upsilon' = 1$ and $\Upsilon \in [1, 1.52]$. Take

$$\phi(\Upsilon) = \frac{1}{\Upsilon^3} - (\Upsilon - 1)^2.$$

Then for any fixed $\delta > 0$, we have

$$\lim_{S \to \infty} \mathbb{P}\big(\|M\|_2 \leq 1 - \phi(\Upsilon) + \delta\big) = 1.$$

*Proof.* See Sect. 4.3.3. □

*Remark* 4.3.1. We note that for $\Upsilon \in [1, 1.52]$, $\phi(\Upsilon) \geq 0.0143 > 0$.

*Theorem* 4.3.3. Assume that $\rho$ and $\kappa$ satisfy condition (4.6), $\Upsilon' = \Upsilon$, and $\sigma_1^2, \sigma_2^2, \sigma_E^2, \Upsilon$ are independent of $S$. Also assume that there exist a series of numbers $\epsilon(S) > 0$ with $\lim\limits_{S \to \infty} \epsilon(S) = 0$, such that for every $1 \leq i \leq R$ and $1 \leq j \leq C$,

$$\frac{S}{C}(1 - \epsilon(S)) \leq \sum_{i=1}^{R} p_{ij} \leq \frac{S}{C}(1 + \epsilon(S)),$$

$$\frac{S}{R}(1 - \epsilon(S)) \leq \sum_{j=1}^{C} p_{ij} \leq \frac{S}{R}(1 + \epsilon(S)).$$

Then there exists $\delta > 0$ (independent of $S$), such that

$$\lim_{S \to \infty} \mathbb{P}(\|M\|_2 \leq 1 - \delta) = 1.$$

*Proof.* See Sect. 4.3.4. □

The rest of this section is devoted to the proofs of Theorems 4.3.1-4.3.3, in Sect. 4.3.2-4.3.4. Some preparatory results are presented in Sect. 4.3.1. For any $n \in \{1, 2, \cdots\}$, we denote by $[n] = \{1, 2, \cdots, n\}$.

### 4.3.1 Preparatory results

In this section, we present several preparatory results, which will be used in the proofs of Theorems 4.3.1–4.3.3.

*Lemma* 4.3.1. If $X \sim \mathrm{Bin}(n,p)$, then for any $t \geqslant 0$,

$$\mathbb{P}(X \geqslant np + t) \leqslant \exp(-2t^2/n), \quad \text{and}$$
$$\mathbb{P}(X \leqslant np - t) \leqslant \exp(-2t^2/n).$$

*Proof.* This follows from Hoeffding's theorem. □

*Lemma* 4.3.2. Under the model in (4.10), for any $\psi > 0$,

$$\mathbb{P}\Big((\frac{1}{\Upsilon'} - \psi)S^{1-\rho} \leqslant \min_i \sum_j Z_{ij} \leqslant \max_i \sum_j Z_{ij} \leqslant (\Upsilon + \psi)S^{1-\rho}\Big)$$
$$\geq 1 - 2R\exp(-2S^{2-\kappa-2\rho}\psi^2). \tag{4.11}$$

Hence if $\rho + \frac{1}{2}\kappa < 1$, then for any fixed $\psi > 0$,

$$\lim_{S\to\infty} \mathbb{P}\Big((\frac{1}{\Upsilon'} - \psi)S^{1-\rho} \leqslant \min_i \sum_j Z_{ij} \leqslant \max_i \sum_j Z_{ij} \leqslant (\Upsilon + \psi)S^{1-\rho}\Big) = 1. \tag{4.12}$$

Likewise, for any $\psi > 0$,

$$\mathbb{P}\Big((\frac{1}{\Upsilon'} - \psi)S^{1-\kappa} \leqslant \min_j \sum_i Z_{ij} \leqslant \max_j \sum_i Z_{ij} \leqslant (\Upsilon + \psi)S^{1-\kappa}\Big)$$
$$\geq 1 - 2C\exp(-2S^{2-\rho-2\kappa}\psi^2). \tag{4.13}$$

Hence if $\frac{1}{2}\rho + \kappa < 1$, then for any fixed $\psi > 0$,

$$\lim_{S\to\infty} \mathbb{P}\Big((\frac{1}{\Upsilon'} - \psi)S^{1-\kappa} \leqslant \min_j \sum_i Z_{ij} \leqslant \max_j \sum_i Z_{ij} \leqslant (\Upsilon + \psi)S^{1-\kappa}\Big) = 1. \tag{4.14}$$

*Proof.* See Appendix C.1.2. □

*Lemma* 4.3.3. $\|Z\|_2 \leq \sqrt{\max_i N_{i\bullet} \max_j N_{\bullet j}}$. Under the model in (4.10) and the condition in (4.6), with high probability $\|Z\|_2 \leq \frac{(\Upsilon+1)S}{\sqrt{RC}}$.

*Proof.* See Appendix C.1.3 □

The following bound on spectral norms of random matrices was proved by Latała (2005).

*Proposition* 4.3.1. [Latała (2005), Theorem 2] Suppose $X$ is a random matrix with independent mean zero entries $X_{ij}$. Then

$$\mathbb{E}\|X\|_2 \leq K\left(\max_i \sqrt{\sum_j \mathbb{E}X_{ij}^2} + \max_j \sqrt{\sum_i \mathbb{E}X_{ij}^2} + \sqrt[4]{\sum_{ij} \mathbb{E}X_{ij}^4}\right), \tag{4.15}$$

where $K > 0$ is an absolute constant.

*Lemma* 4.3.4. Under the model in (4.10),

$$\mathbb{E}\|Z - \mathbb{E}(Z)\|_2 \leq K\left(\sqrt{\frac{\Upsilon S}{R}} + \sqrt{\frac{\Upsilon S}{C}} + \sqrt[4]{\Upsilon S}\right), \tag{4.16}$$

where $K > 0$ is an absolute constant.

*Proof.* For any $1 \leq i \leq R$ and $1 \leq j \leq C$,

$$\mathbb{E}[(Z_{ij} - p_{ij})^2] = p_{ij}(1 - p_{ij}) \leq p_{ij} \leq \frac{\Upsilon S}{RC}, \tag{4.17}$$

$$\mathbb{E}[(Z_{ij} - p_{ij})^4] = p_{ij}(1 - p_{ij})((1 - p_{ij})^3 + p_{ij}^3) \leq p_{ij} \leq \frac{\Upsilon S}{RC}. \tag{4.18}$$

As $\{Z_{ij} - p_{ij}\}_{1\leq i\leq R, 1\leq j\leq C}$ are independent with zero mean, we can apply Proposition 4.3.1 and get

$$\mathbb{E}\left\|Z - \mathbb{E}(Z)\right\|_2 \leq K\Bigg(\max_i \sqrt{\sum_j \mathbb{E}(Z_{ij} - p_{ij})^2} + \max_j \sqrt{\sum_i \mathbb{E}(Z_{ij} - p_{ij})^2} + \sqrt[4]{\sum_{ij} \mathbb{E}(Z_{ij} - p_{ij})^4}\Bigg)$$

$$\leq K\left(\sqrt{\frac{\Upsilon S}{R}}+\sqrt{\frac{\Upsilon S}{C}}+\sqrt[4]{\Upsilon S}\right), \tag{4.19}$$

where $K > 0$ is an absolute constant. □

### 4.3.2 First regime

In this section, we complete the proof of Theorem 4.3.1.

*Proof of Theorem 4.3.1.* The proof proceeds by approximating factors of the matrix $M$ by easier-to-handle quantities. The main technical tools involved are concentration inequalities for $N_{i\cdot}, N_{\cdot j}$ and Rafał Latała's bound on spectral norms of random matrices (see Proposition 4.3.1 and Lemma 4.3.4).

To start with we bring in some notations : for any $i \in [R], j \in [C]$, we let $d_i = (D_1^{-1})_{ii} = \frac{1}{N_{i\cdot}+\lambda_A}$, $d'_j = (D_2^{-1})_{jj} = \frac{1}{N_{\cdot j}+\lambda_B}$. Let

$$L = (I_R - \boldsymbol{w}^{(1)}\mathbf{1}_R^\top)D_1^{-1}, \quad L_0 = \frac{R}{S}(I_R - \frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\top).$$

We note that $M = (I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\top)M_0$ with $M_0 = D_2^{-1}Z^\top LZ$. In the following, we upper bound $\|M\|_2$ through four steps:

- Step 1: provide an upper bound of $\|M_0\|_2$ in terms of $\|Z^\top LZ\|_2$;
- Step 2: approximate $Z^\top LZ$ by $Z^\top L_0 Z$, and upper bound the spectral norm of their difference;
- Step 3: upper bound $\|Z^\top L_0 Z\|_2$ by Lemma 4.3.4, which combined with Steps 1-2 provides an upper bound of $\|M_0\|_2$ by the triangle inequality

$$\|Z^\top LZ\|_2 \leq \|Z^\top L_0 Z\|_2 + \|Z^\top LZ - Z^\top L_0 Z\|;$$

- Step 4: upper bound $\|M\|_2$ using the upper bound of $\|M_0\|_2$ and the inequality

$$\|M\|_2 \leq \|I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\top\|_2\|M_0\|_2.$$

**Step 1**: Firstly we bound the quantities $N_{i\cdot}, N_{\cdot j}$ by Hoeffding's inequality. We take $\psi(S) = \min\{\frac{1}{20}, \frac{1}{\log(S+1)}\}$, and let $\mathcal{A}_S$ be the event that for any $i \in [R], j \in [C]$, $|N_{i\cdot} - \frac{S}{R}| \leq \frac{S}{R}\psi(S)$, $|N_{\cdot j} - \frac{S}{C}| \leq \frac{S}{C}\psi(S)$. Using Lemma 4.3.2 with $\Upsilon = \Upsilon' = 1$, under the condition on $(\rho, \kappa)$ mentioned in the theorem,

$$\lim_{S\to\infty} \mathbb{P}(\mathcal{A}_S^c) = 0.$$

In the following, we assume that the event $\mathcal{A}_S$ holds. Note that for any $j \in [C]$,

$$\begin{aligned} |d'_j - \frac{C}{S}| &\leq \frac{|\sum_{i=1}^R Z_{ij} - \frac{S}{C}| + \lambda_B}{(\sum_{i=1}^R Z_{ij})\frac{S}{C}} \\ &\leq \frac{\frac{S}{C}\psi(S) + \lambda_B}{\frac{S^2}{2C^2}} = \delta_2 \frac{C}{S}, \end{aligned}$$

with $\delta_2 = 2(\psi(S) + \frac{C}{S}\lambda_B)$. Hence

$$\|D_2^{-1} - \frac{C}{S} I_C\|_2 = \max_{j\in[C]} |d'_j - \frac{C}{S}| \leq \delta_2 \frac{C}{S}. \tag{4.20}$$

This leads to

$$\begin{aligned} &\|M_0 - \frac{C}{S} Z^\mathsf{T} L Z\|_2 = \|D_2^{-1} Z^\mathsf{T} L Z - \frac{C}{S} Z^\mathsf{T} L Z\|_2 \\ &\leq \|D_2^{-1} - \frac{C}{S} I_C\|_2 \|Z^\mathsf{T} L Z\|_2 \leq \delta_2 \frac{C}{S} \|Z^\mathsf{T} L Z\|_2. \end{aligned}$$

Using the triangle inequality we obtain

$$\|M_0\|_2 \leq (1 + \delta_2)\frac{C}{S} \|Z^\mathsf{T} L Z\|_2. \tag{4.21}$$

**Step 2**: We start by bounding $\|L - L_0\|_2$. We let $D = \boldsymbol{w}^{(1)} \mathbf{1}_R^\mathsf{T} D_1^{-1}$ with $D_{ij} = w_i^{(1)} d_j$. This gives

$$\|L - L_0\|_2 \leq \|D_1^{-1} - \frac{R}{S} I_R\|_2 + \|D - \frac{1}{S}\mathbf{1}_R \mathbf{1}_R^\mathsf{T}\|_2. \tag{4.22}$$

Using a similar result as equation (4.20),

$$\|D_1^{-1} - \frac{R}{S} I_R\|_2 = \max_{i\in[R]} |d_i - \frac{R}{S}| \le \delta_1 \frac{R}{S}. \tag{4.23}$$

for $\delta_1 = 2(\psi(S) + \frac{R}{S}\lambda_A)$. For any $i, j \in [R]$, we have

$$\begin{aligned}
|w_i^{(1)} d_j - \frac{1}{S}| &\le |\frac{s_i^{(1)} d_j}{\sum_{i'=1}^{R} s_{i'}^{(1)}} - \frac{s_i^{(1)} d_j}{R}| + |\frac{s_i^{(1)} d_j}{R} - \frac{d_j}{R}| + |\frac{d_j}{R} - \frac{1}{S}| \\
&\le |s_i^{(1)} d_j| |\frac{1}{\sum_{i'=1}^{R} s_{i'}^{(1)}} - \frac{1}{R}| + \frac{d_j}{R} |s_i^{(1)} - 1| + |\frac{d_j}{R} - \frac{1}{S}| \\
&\le \frac{R}{S}(1+\delta_1)\frac{1}{R}(1 + \frac{2\lambda_A R}{S} - 1) + \frac{d_j}{R} \frac{\frac{2\lambda_A R}{S}}{1 + \frac{2\lambda_A R}{S}} + \frac{\delta_1}{S} \\
&\le \frac{4(1+\delta_1)\lambda_A R}{S^2} + \frac{\delta_1}{S} \le \delta_1 \frac{C_0}{S},
\end{aligned} \tag{4.24}$$

where $C_0$ is a constant. In the above derivation, we have used $|s_i^{(1)}| \le 1$ and $|d_j - \frac{R}{S}| \le \delta_1 \frac{R}{S}$ for the first and last terms while going from the inequality in the second line to the third line. We have also used the following string of inequalities :

$$1 \ge s_i^{(1)} = \frac{1}{1 + \frac{\lambda_A}{N_{i\cdot}}} \ge \frac{1}{1 + \frac{\lambda_A}{(1-\psi(S))\frac{S}{R}}} \ge \frac{1}{1 + \frac{2\lambda_A R}{S}}. \tag{4.25}$$

By (4.24), we have

$$\|D - \frac{1}{S}\mathbf{1}_R \mathbf{1}_R^{\mathsf{T}}\|_2 \le \|D - \frac{1}{S}\mathbf{1}_R \mathbf{1}_R^{\mathsf{T}}\|_F = \sqrt{\sum_{i,j\in[R]} (D_{ij} - \frac{1}{S})^2} \le \delta_1 C_0 \frac{R}{S}. \tag{4.26}$$

Plugging in the results from equations (4.23) and (4.26) to (4.22), we obtain

$$\|L - L_0\|_2 \le \delta_1 (1 + C_0) \frac{R}{S}.$$

Let $\mathcal{B}_S$ be the event that $\|Z\|, \|Z^{\mathsf{T}}\| \le \frac{2S}{\sqrt{RC}}$. By Lemma 4.3.3,

$$\lim_{S\to\infty} \mathbb{P}(\mathcal{B}_S^c) = 0.$$

Below we assume that $\mathcal{B}_S$ holds. We have

$$\begin{aligned}
&\|Z^\mathsf{T} L Z - Z^\mathsf{T} L_0 Z\|_2 \leq \|Z^\mathsf{T}\|_2 \|L - L_0\|_2 \|Z\|_2 \\
&\leq (1 + C_0)\delta_1 \frac{R}{S}\frac{4S^2}{RC} \leq C_0' \delta_1 \frac{S}{C},
\end{aligned} \tag{4.27}$$

where $C_0'$ is a constant.

**Step 3**: Using the fact that $\mathbb{E}(Z) = k\mathbf{1}_R\mathbf{1}_C^\mathsf{T}$ for some constant $k$ and $L_0\mathbf{1}_R = \mathbf{0}$, we deduce that

$$Z^\mathsf{T} L_0 Z = (Z - \mathbb{E}(Z))^\mathsf{T} L_0 (Z - \mathbb{E}(Z)).$$

We also note

$$\begin{aligned}
\|(Z - \mathbb{E}(Z))^\mathsf{T} L_0 (Z - \mathbb{E}(Z))\|_2 &= \frac{S}{R}\|L_0(Z - \mathbb{E}(Z))\|_2^2 \quad \text{using } L_0^2 = \tfrac{R}{S} L_0 \\
\|L_0(Z - \mathbb{E}(Z))\|_2 &\leq \frac{R}{S}\|Z - \mathbb{E}(Z)\|_2 \quad \text{because } \|L_0\|_2 = \tfrac{R}{S}
\end{aligned} \tag{4.28}$$

By Lemma 4.3.4 (with $\Upsilon = 1$), we have

$$\mathbb{E}\|Z - \mathbb{E}(Z)\|_2 \leq K\left(\sqrt{\frac{S}{R}} + \sqrt{\frac{S}{C}} + \sqrt[4]{S}\right), \tag{4.29}$$

where $K > 0$ is an absolute constant. Plugging the inequality of (4.28) into equation (4.29) we obtain,

$$\mathbb{E}\left\|L_0\left(Z - \mathbb{E}(Z)\right)\right\|_2 \leq K\left(\sqrt{\frac{R}{S}} + \frac{R}{S^{0.5}C^{0.5}} + \frac{R}{S^{0.75}}\right).$$

We take $\epsilon = \frac{1}{\log(S+1)}$. By Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\left\|L_0\left(Z - \mathbb{E}(Z)\right)\right\|_2 > \frac{1}{\epsilon}\mathbb{E}\left\|L_0\left(Z - \mathbb{E}(Z)\right)\right\|_2\right) < \epsilon \\
\Longrightarrow &\mathbb{P}\left(\left\|L_0\left(Z - \mathbb{E}(Z)\right)\right\|_2 > \frac{K}{\epsilon}\left(\sqrt{\frac{R}{S}} + \frac{R}{S^{0.5}C^{0.5}} + \frac{R}{S^{0.75}}\right)\right) < \epsilon
\end{aligned}$$

$$\implies \mathbb{P}\left(\frac{S}{R}\|L_0(Z-\mathbb{E}(Z))\|_2^2 > 3(\frac{K}{\epsilon})^2 \max\{1, \frac{R}{C}, \frac{R}{S^{0.5}}\}\right) < \epsilon$$
$$\implies \mathbb{P}\left(\|Z^\mathsf{T} L_0 Z\|_2 > 3(\frac{K}{\epsilon})^2 \max\{1, \frac{R}{C}, \frac{R}{S^{0.5}}\}\right) < \epsilon.$$

Let $\mathcal{D}_\epsilon$ be the event that $\|Z^\mathsf{T} L_0 Z\|_2 \leq 3(\frac{K}{\epsilon})^2 \max\{1, \frac{R}{C}, \frac{R}{S^{0.5}}\}$. In view of the inequality (4.27), we obtain on $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$ ,

$$\|Z^\mathsf{T} L Z\|_2 \leq C_0'\delta_1 \frac{S}{C} + 3(\frac{K}{\epsilon})^2 \max\{1, \frac{R}{C}, \frac{R}{S^{0.5}}\}.$$

From equation (4.21), on $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$ we get

$$\|M_0\|_2 < (1+\delta_2)\big(C_0'\delta_1 + 3(\frac{K}{\epsilon})^2 \max\{\frac{C}{S}, \frac{R}{S}, \frac{RC}{S^{3/2}}\}\big). \tag{4.30}$$

The quantity on the right hand side goes to 0 when $\rho + \kappa < \frac{3}{2}$, which is implied by the condition $\max\{\rho + \frac{1}{2}\kappa, \kappa + \frac{1}{2}\rho\} < 1$.

**Step 4**: For the last step, we note that

$$\|M\|_2 \leq \|I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \|M_0\|_2. \tag{4.31}$$

Similar to the inequality in (4.25),

$$1 \geq s_j^{(2)} = \frac{1}{1 + \frac{\lambda_B}{N_{\cdot j}}} \geq \frac{1}{1 + \frac{\lambda_B}{(1-\psi(S))\frac{S}{C}}} \geq \frac{1}{1 + \frac{2\lambda_B C}{S}}. \tag{4.32}$$

Noting that $\boldsymbol{w}^{(2)} = \frac{s_j^{(2)}}{\sum_{j'=1}^{C} s_{j'}^{(2)}}$, we obtain for any $j \in [C]$,

$$\frac{1}{C}\frac{1}{1 + \frac{2\lambda_B C}{S}} \leq w_j^{(2)} \leq \frac{1}{C}(1 + \frac{2\lambda_B C}{S}). \tag{4.33}$$

Therefore, there exists $\delta(S) \to 0$, such that

$$|w_j^{(2)} - \frac{1}{C}| \leq \delta(S)\frac{1}{C}. \tag{4.34}$$

Observe that

$$\begin{aligned}\|I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\top\|_2 &\le \|I_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\|_2 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\top - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\|_2 \\ &\le 1 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\top - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\|_F.\end{aligned}$$

Using (4.34), the second term can be made arbitrarily small. Plugging this into the inequality (4.31), we obtain $\|M\|_2 \to 0$ on $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$. Noting that $\lim_{S\to\infty} \mathbb{P}(\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon) = 1$, we conclude the proof.

□

### 4.3.3 Second regime

In this section, we give the proof of Theorem 4.3.2. We note that Theorem 4.3.1 is a special case of Theorem 4.3.2 with $\Upsilon = 1$. However, the simpler case illustrated in Theorem 4.3.1 sets up the field for the extensions in Theorems 4.3.2 and 4.3.3.

*Proof of Theorem 4.3.2.* The proof of Theorem 4.3.2 involves upper bounding the spectral norm of a conjugated version $\mathcal{M}$ of the matrix $M$. We proceed by approximating $\mathcal{M}$ by a deterministic matrix $\mathcal{M}'$, which is obtained from replacing each factor of $\mathcal{M}$ by easier-to-handle quantities (see below for the precise definitions of $\mathcal{M}, \mathcal{M}'$). By the triangle inequality

$$\|\mathcal{M}\|_2 \le \|\mathcal{M} - \mathcal{M}'\|_2 + \|\mathcal{M}'\|_2,$$

we accomplish the proof in two steps. For the first step, we upper bound $\|\mathcal{M} - \mathcal{M}'\|_2$. The main technical tools involved in this step are concentration inequalities for the row and column sums $N_{i\cdot}, N_{\cdot j}$ and Rafał Latała's bound on spectral norms of random matrices (see Proposition 4.3.1 and Lemma 4.3.4). The second step involves upper bounding $\|\mathcal{M}'\|_2$. The analysis in this step involves studying certain quadratic forms (similar to the Dirichlet form in the setting of reversible Markov chains) related to the matrix $\mathcal{M}'$.

First we set up some notations. We let $\bar{N}_{i\cdot} = \sum_{j=1}^{C} p_{ij}$ and $\bar{N}_{\cdot j} = \sum_{i=1}^{R} p_{ij}$ for any $i \in [R], j \in [C]$. We also let $\bar{D}_1 = diag(\bar{N}_{i\cdot}), \bar{D}_2 = diag(\bar{N}_{\cdot j})$, and $\bar{Z} = (p_{ij})$. We take $\psi(S) = \min\{\frac{1}{20}, \frac{1}{\log(S+1)}\}$, and assume that $S$ is sufficiently large so that $\frac{R}{S}\lambda_A, \frac{C}{S}\lambda_B \leq \frac{1}{20}$ (note that $\rho, \kappa < 1$ by the condition (4.6)).

We define

$$\mathcal{M} = D_2^{\frac{1}{2}} M D_2^{-\frac{1}{2}},$$

$$\mathcal{M}' = (I_C - \frac{1}{C}\bar{D}_2^{\frac{1}{2}}\mathbf{1}_C\mathbf{1}_C^\mathsf{T}\bar{D}_2^{-\frac{1}{2}})(\bar{D}_2^{-\frac{1}{2}}\bar{Z}^\mathsf{T}\bar{D}_1^{-\frac{1}{2}})(I_R - \frac{1}{R}\bar{D}_1^{\frac{1}{2}}\mathbf{1}_R\mathbf{1}_R^\mathsf{T}\bar{D}_1^{-\frac{1}{2}})(\bar{D}_1^{-\frac{1}{2}}\bar{Z}\bar{D}_2^{-\frac{1}{2}}).$$

Note that $\rho(M) = \rho(\mathcal{M}) \leq \|\mathcal{M}\|_2 \leq \|\mathcal{M} - \mathcal{M}'\|_2 + \|\mathcal{M}'\|_2$. In the following, we bound $\|\mathcal{M} - \mathcal{M}'\|_2$ and $\|\mathcal{M}'\|_2$.

**Step 1**: For this step, we provide an upper bound on $\|\mathcal{M} - \mathcal{M}'\|_2$.

We first bound the quantities $N_{i\cdot}$ and $N_{\cdot j}$ for $i \in [R], j \in [C]$. By Hoeffding's inequality, for any $i \in [R], j \in [C]$,

$$\mathbb{P}\Big(|N_{i\cdot} - \bar{N}_{i\cdot}| \geq \frac{S}{R}\psi(S)\Big) \leq 2e^{-\frac{2S^2\psi(S)^2}{CR^2}}, \tag{4.35}$$

$$\mathbb{P}\Big(|N_{\cdot j} - \bar{N}_{\cdot j}| \geq \frac{S}{C}\psi(S)\Big) \leq 2e^{-\frac{2S^2\psi(S)^2}{RC^2}}. \tag{4.36}$$

We denote by $A_S$ be the event that for any $i \in [R], j \in [C]$, $|N_{i\cdot} - \bar{N}_{i\cdot}| \leq \frac{S}{R}\psi(S)$, $|N_{\cdot j} - \bar{N}_{j\cdot}| \leq \frac{S}{C}\psi(S)$. By the union bound we obtain

$$\mathbb{P}(\mathcal{A}_S^c) \leq 2Re^{-\frac{2S^2\psi(S)^2}{CR^2}} + 2Ce^{-\frac{2S^2\psi(S)^2}{RC^2}}. \tag{4.37}$$

Hence

$$\lim_{S\to\infty} \mathbb{P}(\mathcal{A}_S^c) = 0. \tag{4.38}$$

In the following we assume that $\mathcal{A}_S$ holds. We let $\delta_1(S) = 3(\psi(S) + \frac{R}{S}\lambda_A)$ and $\delta_2(S) = 3(\psi(S) + \frac{C}{S}\lambda_B)$. Note that by our assumption, $\delta_1(S), \delta_2(S) \leq \frac{1}{2}$. For any $i \in [R]$,

$$|\frac{1}{N_{i\cdot} + \lambda_A} - \frac{1}{\bar{N}_{i\cdot}}| \leq \frac{|N_{i\cdot} - \bar{N}_{i\cdot}| + \lambda_A}{(N_{i\cdot} + \lambda_A)\bar{N}_{i\cdot}} \leq \frac{\frac{S}{R}\psi(S) + \lambda_A}{(\frac{S}{R})^2(1 - \psi(S))} \leq \delta_1(S)\frac{R}{S}.$$

Similarly, for any $j \in [C]$,

$$|\frac{1}{N_{\cdot j} + \lambda_B} - \frac{1}{\bar{N}_{\cdot j}}| \leq \delta_2(S)\frac{C}{S}.$$

Hence we have that

$$\|D_1^{-1} - \bar{D}_1^{-1}\|_2 \leq \delta_1(S)\frac{R}{S}, \quad \|D_2^{-1} - \bar{D}_2^{-1}\|_2 \leq \delta_2(S)\frac{C}{S}. \tag{4.39}$$

Now note that

$$|\sqrt{N_{i\cdot} + \lambda_A} - \bar{N}_{i\cdot}| \leq \frac{|N_{i\cdot} - \bar{N}_{i\cdot}| + \lambda_A}{\sqrt{N_{i\cdot} + \lambda_A} + \sqrt{\bar{N}_{i\cdot}}} \leq \frac{\frac{S}{R}\psi(S) + \lambda_A}{\sqrt{\frac{S}{R}}} = \frac{\delta_1(S)}{3}\sqrt{\frac{S}{R}}, \tag{4.40}$$

$$\begin{aligned} |\frac{1}{\sqrt{N_{i\cdot} + \lambda_A}} - \frac{1}{\sqrt{\bar{N}_{i\cdot}}}| &= \frac{|\sqrt{N_{i\cdot} + \lambda_A} - \sqrt{\bar{N}_{i\cdot}}|}{\sqrt{N_{i\cdot} + \lambda_A}\sqrt{\bar{N}_{i\cdot}}} \\ \leq \quad & \frac{\frac{\delta_1(S)}{3}\sqrt{\frac{S}{R}}}{\sqrt{(\frac{S}{R})^2(1 - \psi(S))}} \leq \delta_1(S)\sqrt{\frac{R}{S}}. \end{aligned} \tag{4.41}$$

Hence

$$\|D_1^{\frac{1}{2}} - \bar{D}_1^{\frac{1}{2}}\|_2 \leq \delta_1(S)\sqrt{\frac{S}{R}}, \quad \|D_1^{-\frac{1}{2}} - \bar{D}_1^{-\frac{1}{2}}\|_2 \leq \delta_1(S)\sqrt{\frac{R}{S}}. \tag{4.42}$$

Similarly,

$$\|D_2^{\frac{1}{2}} - \bar{D}_2^{\frac{1}{2}}\|_2 \leq \delta_2(S)\sqrt{\frac{S}{C}}, \quad \|D_2^{-\frac{1}{2}} - \bar{D}_2^{-\frac{1}{2}}\|_2 \leq \delta_2(S)\sqrt{\frac{C}{S}}. \tag{4.43}$$

Now note that $\|\bar{D}_1^{-1}\|_2 \leq \frac{R}{S}$, $\|\bar{D}_2^{-1}\|_2 \leq \frac{C}{S}$, $\|\bar{D}_1^{-\frac{1}{2}}\|_2 \leq \sqrt{\frac{R}{S}}$, and $\|\bar{D}_2^{-\frac{1}{2}}\|_2 \leq \sqrt{\frac{C}{S}}$. Hence by the above estimates, we have $\|D_1^{-1}\|_2 \leq \frac{2R}{S}$, $\|D_2^{-1}\|_2 \leq \frac{2C}{S}$, $\|D_1^{-\frac{1}{2}}\|_2 \leq 2\sqrt{\frac{R}{S}}$, and $\|D_2^{-\frac{1}{2}}\|_2 \leq 2\sqrt{\frac{C}{S}}$.

Using an argument similar to (4.25), for any $i \in [R]$, we have $\frac{1}{1+\frac{2\lambda_A R}{S}} \leq s_i^{(1)} \leq 1$ and for any $j \in [C]$, $\frac{1}{1+\frac{2\lambda_B C}{S}} \leq s_j^{(2)} \leq 1$. Following (4.34), we conclude there exists

$\delta_3(S) \to 0$, such that

$$|w_i^{(1)} - \frac{1}{R}| \le \delta_3(S)\frac{1}{R}, \quad |w_j^{(2)} - \frac{1}{C}| \le \delta_3(S)\frac{1}{C}.$$

As a result of the above two inequalities,

$$\begin{aligned} &\|\frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\mathsf{T} - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 \le \|\frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\mathsf{T} - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_F \le \delta_3(S), \\ &\|\frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\mathsf{T} - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \le \|\frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\mathsf{T} - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_F \le \delta_3(S). \end{aligned} \tag{4.44}$$

We assume $\delta_3(S) \le \frac{1}{4}$ (taking large $S$). Noting that $0 \le w_i^{(1)} \le \frac{1}{R}(1 + \delta_3(S))$ and $0 \le w_j^{(2)} \le \frac{1}{C}(1 + \delta_3(S))$, we obtain

$$\begin{aligned} &\|I_R - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 \le \|I_R\|_2 + \|\boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 \le 1 + \|\boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_F \le 4, \\ &\|I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \le \|I_C\|_2 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \le 1 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_F \le 4. \end{aligned} \tag{4.45}$$

To simplify notations, in the rest of this proof we denote by

$$\dot{I}_R = I_R - \frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\mathsf{T}, \quad \dot{I}_C = I_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\mathsf{T}; \tag{4.46}$$

$$\tilde{I}_R = I_R - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}, \quad \tilde{I}_C = I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}. \tag{4.47}$$

By Lemma 4.3.3 and the fact that $\Upsilon < 2$ (under the condition in the theorem), we have $\|\bar{Z}\|_2, \|\bar{Z}^\mathsf{T}\|_2 \le \frac{3S}{\sqrt{RC}}$. Let $\mathcal{B}_S$ be the event that $\|Z\|_2, \|Z^\mathsf{T}\|_2 \le \frac{3S}{\sqrt{RC}}$. Using Lemma 4.3.3 again, we have $\lim_{S\to\infty} \mathbb{P}(\mathcal{B}_S^c) = 0$. In the following, we assume that the event $\mathcal{B}_S$ holds.

By Lemma 4.3.4, we have

$$\mathbb{E}[\|Z - \bar{Z}\|_2] \le K_1 \left( \sqrt{\frac{\Upsilon S}{R}} + \sqrt{\frac{\Upsilon S}{C}} + \sqrt[4]{\Upsilon S} \right),$$

where $K_1 > 0$ is an absolute constant. For any $\epsilon > 0$, let $\mathcal{D}_\epsilon$ be the event that

$$\|Z - \bar{Z}\|_2 \le \frac{K_1}{\epsilon}\left(\sqrt{\frac{\Upsilon S}{R}} + \sqrt{\frac{\Upsilon S}{C}} + \sqrt[4]{\Upsilon S}\right). \tag{4.48}$$

By Markov's inequality, we have $\mathbb{P}(\mathcal{D}_\epsilon^c) \le \epsilon$.

For the following we assume that $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$ holds. Note that we have (recall the notations (4.46)-(4.47))

$$\mathcal{M}' = \bar{D}_2^{\frac{1}{2}} \dot{I}_C \bar{D}_2^{-1} \bar{Z}^\mathsf{T} \dot{I}_R \bar{D}_1^{-1} \bar{Z} \bar{D}_2^{-\frac{1}{2}}.$$

Hence combining the above estimates (replacing each factor of $\mathcal{M}$ by the corresponding factor of $\mathcal{M}'$ one by one), we have that

$$\begin{aligned}
&\|\mathcal{M} - \mathcal{M}'\|_2 \\
\le\ & \|D_2^{\frac{1}{2}} - \bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\tilde{I}_C\|_2 \cdot \|D_2^{-1}\|_2 \cdot \|Z^\mathsf{T}\|_2 \cdot \|\tilde{I}_R\|_2 \cdot \|D_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\tilde{I}_C - \dot{I}_C\|_2 \cdot \|D_2^{-1}\|_2 \cdot \|Z^\mathsf{T}\|_2 \cdot \|\tilde{I}_R\|_2 \cdot \|D_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|D_2^{-1} - \bar{D}_2^{-1}\|_2 \cdot \|Z^\mathsf{T}\|_2 \cdot \|\tilde{I}_R\|_2 \cdot \|D_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|\bar{D}_2^{-1}\|_2 \cdot \|Z^\mathsf{T} - \bar{Z}^\mathsf{T}\|_2 \cdot \|\tilde{I}_R\|_2 \cdot \|D_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|\bar{D}_2^{-1}\|_2 \cdot \|\bar{Z}^\mathsf{T}\|_2 \cdot \|\tilde{I}_R - \dot{I}_R\|_2 \cdot \|D_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|\bar{D}_2^{-1}\|_2 \cdot \|\bar{Z}^\mathsf{T}\|_2 \cdot \|\dot{I}_R\|_2 \cdot \|D_1^{-1} - \bar{D}_1^{-1}\|_2 \cdot \|Z\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|\bar{D}_2^{-1}\|_2 \cdot \|\bar{Z}^\mathsf{T}\|_2 \cdot \|\dot{I}_R\|_2 \cdot \|\bar{D}_1^{-1}\|_2 \cdot \|Z - \bar{Z}\|_2 \cdot \|D_2^{-\frac{1}{2}}\|_2 \\
+\ & \|\bar{D}_2^{\frac{1}{2}}\|_2 \cdot \|\dot{I}_C\|_2 \cdot \|\bar{D}_2^{-1}\|_2 \cdot \|\bar{Z}^\mathsf{T}\|_2 \cdot \|\dot{I}_R\|_2 \cdot \|\bar{D}_1^{-1}\|_2 \cdot \|\bar{Z}\|_2 \cdot \|D_2^{-\frac{1}{2}} - \bar{D}_2^{-\frac{1}{2}}\|_2 \\
\le\ & K\left(\delta_1(S) + \delta_2(S) + \delta_3(S) + \epsilon^{-1}(S^{-\frac{1}{2}(1-\rho)} + S^{-\frac{1}{2}(1-\kappa)} + S^{-(\frac{3}{4}-\frac{1}{2}\rho-\frac{1}{2}\kappa)})\right),
\end{aligned}$$

where $K > 0$ is an absolute constant. Note that by the condition (4.6) we have $\phi(\rho, \kappa) := \min\{\frac{1}{2}(1-\rho), \frac{1}{2}(1-\kappa), \frac{3}{4} - \frac{1}{2}(\rho+\kappa)\} > 0$. We take $\epsilon = S^{-\frac{1}{2}\phi(\rho,\kappa)}$. Then

$$\lim_{S\to\infty} \mathbb{P}((\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon)^c) = 0. \tag{4.49}$$

We take $\delta(S) = K(\delta_1(S) + \delta_2(S) + \delta_3(S) + 3S^{-\frac{1}{2}\phi(\rho,\kappa)})$, and note that $\lim_{S\to\infty} \delta(S) = 0$.

Moreover, when the event $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$ holds,

$$\|\mathcal{M} - \mathcal{M}'\|_2 \leq \delta(S). \tag{4.50}$$

**Step 2**: For this step, we provide an upper bound on $\|\mathcal{M}'\|_2$.

We let

$$\mathcal{M}_1' = (I_R - \frac{1}{R}\bar{D}_1^{\frac{1}{2}}\mathbf{1}_R\mathbf{1}_R^\mathsf{T}\bar{D}_1^{-\frac{1}{2}})(\bar{D}_1^{-\frac{1}{2}}\bar{Z}\bar{D}_2^{-\frac{1}{2}}),$$

$$\mathcal{M}_2' = (I_C - \frac{1}{C}\bar{D}_2^{\frac{1}{2}}\mathbf{1}_C\mathbf{1}_C^\mathsf{T}\bar{D}_2^{-\frac{1}{2}})(\bar{D}_2^{-\frac{1}{2}}\bar{Z}^\mathsf{T}\bar{D}_1^{-\frac{1}{2}}).$$

Note that

$$\|\mathcal{M}'\|_2 \leq \|\mathcal{M}_2'\|_2\|\mathcal{M}_1'\|_2. \tag{4.51}$$

We bound $\|\mathcal{M}_1'\|_2$ in the following. Note that $(\mathcal{M}_1')^\mathsf{T}\mathcal{M}_1'\bar{D}_2^{\frac{1}{2}}1_C = 0$. In the following, we take any $u = (u_s)_{s\in[C]}$ such that $\sum_{s=1}^C u_s\bar{N}_{\cdot s}^{\frac{1}{2}} = 0$. By computation, we obtain that

$$\begin{aligned} u^\mathsf{T}(\mathcal{M}_1')^\mathsf{T}\mathcal{M}_1'u &= \sum_{s=1}^C\sum_{k=1}^C \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}\frac{S_0}{R^2}(\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l\cdot}})(\sum_{l=1}^R \frac{p_{lk}}{\bar{N}_{l\cdot}}) \\ &+ \sum_{s=1}^C\sum_{k=1}^C \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}(\sum_{l=1}^R \frac{p_{ls}p_{lk}}{\bar{N}_{l\cdot}}), \end{aligned} \tag{4.52}$$

where $S_0 := \sum_{i=1}^R\sum_{j=1}^C p_{ij} \in [S, \Upsilon S]$.

Using the fact that $\sum_{s=1}^C u_s\bar{N}_{\cdot s}^{\frac{1}{2}} = 0$, we have

$$\begin{aligned} &\sum_{s=1}^C\sum_{k=1}^C \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}\frac{S_0}{R^2}(\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l\cdot}})(\sum_{l=1}^R \frac{p_{lk}}{\bar{N}_{l\cdot}}) = \frac{S_0}{R^2}\left(\sum_{s=1}^C \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}(\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l\cdot}})\right)^2 \\ = \ &\frac{S_0}{R^2}\left(\sum_{s=1}^C \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}(\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l\cdot}} - \frac{R}{S}\bar{N}_{\cdot s})\right)^2. \end{aligned} \tag{4.53}$$

By the fact that $\frac{S}{R} \leq \bar{N}_{l\cdot} \leq \frac{\Upsilon S}{R}$, we have for every $s \in [C]$,

$$0 \geq \sum_{l=1}^{R} \frac{p_{ls}}{\bar{N}_{l\cdot}} - \frac{R}{S}\bar{N}_{\cdot s} \geq -(1-\frac{1}{\Upsilon})\frac{R}{S}\bar{N}_{\cdot s}.$$

Hence

$$\begin{aligned} & \left(\sum_{s=1}^{C} \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}(\sum_{l=1}^{R} \frac{p_{ls}}{\bar{N}_{l\cdot}} - \frac{R}{S}\bar{N}_{\cdot s})\right)^2 \leq (1-\frac{1}{\Upsilon})^2 \frac{R^2}{S^2}(\sum_{s=1}^{C} |u_s|\sqrt{\bar{N}_{\cdot s}})^2 \\ \leq \;& (1-\frac{1}{\Upsilon})^2 \frac{R^2}{S^2}(\sum_{s=1}^{C}\bar{N}_{\cdot s})(\sum_{s=1}^{C} u_s^2) = (1-\frac{1}{\Upsilon})^2 \frac{R^2 S_0}{S^2}(\sum_{s=1}^{C} u_s^2). \end{aligned} \tag{4.54}$$

By combining (4.53), (4.54), and the fact that $S_0 \leq \Upsilon S$ , we obtain

$$\begin{aligned} & \sum_{s=1}^{C}\sum_{k=1}^{C} \frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}\frac{S_0}{R^2}(\sum_{l=1}^{R}\frac{p_{ls}}{\bar{N}_{l\cdot}})(\sum_{l=1}^{R}\frac{p_{lk}}{\bar{N}_{l\cdot}}) \\ \leq \;& (1-\frac{1}{\Upsilon})^2\frac{S_0^2}{S^2}(\sum_{s=1}^{C} u_s^2) \leq (\Upsilon-1)^2(\sum_{s=1}^{C} u_s^2). \end{aligned} \tag{4.55}$$

Let $P(s,k) = \sum_{l=1}^{R} \frac{p_{ls}p_{lk}}{\bar{N}_{l\cdot}}$ for any $s,k \in [C]$. We note that $P(s,k) = P(k,s)$ and $\sum_{k=1}^{C} P(s,k) = \bar{N}_{\cdot s}$. This gives

$$\sum_{s=1}^{C} u_s^2 = \sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s^2}{\bar{N}_{\cdot s}}P(s,k) = \frac{1}{2}\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s^2}{\bar{N}_{\cdot s}}P(s,k) + \frac{1}{2}\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_k^2}{\bar{N}_{\cdot k}}P(s,k).$$

Hence

$$\begin{aligned} & -\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}P(s,k) + \sum_{s=1}^{C} u_s^2 \\ = \;& \frac{1}{2}\sum_{s=1}^{C}\sum_{k=1}^{C} P(s,k)(\frac{u_s}{\sqrt{\bar{N}_{\cdot s}}} - \frac{u_k}{\sqrt{\bar{N}_{\cdot k}}})^2. \end{aligned} \tag{4.56}$$

Now note that for any $s, k \in [C]$,

$$P(s,k) \geq R\frac{(\frac{S}{RC})^2}{\Upsilon\frac{S}{R}} = \Upsilon^{-1}\frac{S}{C^2} \geq \Upsilon^{-1}SC^{-2}\frac{\bar{N}_{\cdot s}\bar{N}_{\cdot k}}{(\frac{\Upsilon S}{C})^2} = \frac{\bar{N}_{\cdot s}\bar{N}_{\cdot k}}{\Upsilon^3 S}.$$

Hence by (4.56) we have

$$\begin{aligned}
& -\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}P(s,k) + \sum_{s=1}^{C}u_s^2 \\
\geq \;& \frac{1}{2\Upsilon^3 S}\sum_{s=1}^{C}\sum_{k=1}^{C}\bar{N}_{\cdot s}\bar{N}_{\cdot k}\left(\frac{u_s}{\sqrt{\bar{N}_{\cdot s}}} - \frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}\right)^2 \\
\geq \;& \frac{1}{\Upsilon^3}\sum_{s=1}^{C}u_s^2 - \frac{1}{\Upsilon^3 S}\left(\sum_{s=1}^{C}\sqrt{\bar{N}_{\cdot s}}u_s\right)^2 = \frac{1}{\Upsilon^3}\sum_{s=1}^{C}u_s^2, \qquad (4.57)
\end{aligned}$$

where in the last line we have used that $S_0 \geq S$ and $\sum_{s=1}^{C}\sqrt{\bar{N}_{\cdot s}}u_s = 0$. This leads to

$$\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s}{\sqrt{\bar{N}_{\cdot s}}}\frac{u_k}{\sqrt{\bar{N}_{\cdot k}}}P(s,k) \leq \left(1 - \frac{1}{\Upsilon^3}\right)\sum_{s=1}^{C}u_s^2. \qquad (4.58)$$

By combining (4.52), (4.55), and (4.58), we conclude that

$$u^{\mathsf{T}}(\mathcal{M}_1')^{\mathsf{T}}\mathcal{M}_1' u \leq \left(1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2\right)\left(\sum_{s=1}^{C}u_s^2\right). \qquad (4.59)$$

This leads to

$$\|\mathcal{M}_1'\|_2 = \sqrt{\|(\mathcal{M}_1')^{\mathsf{T}}\mathcal{M}_1'\|_2} \leq \sqrt{1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2}. \qquad (4.60)$$

The following bound for $\|\mathcal{M}_2'\|$ can be similarly obtained:

$$\|\mathcal{M}_2'\|_2 \leq \sqrt{1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2}. \qquad (4.61)$$

Hence when $1 \leq \Upsilon \leq 1.52$,

$$\|\mathcal{M}'\|_2 \leq 1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2 \leq 0.9857. \tag{4.62}$$

Recall that on the event $\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon$, $\|\mathcal{M} - \mathcal{M}'\|_2 \leq \delta(S)$, hence $\|\mathcal{M}\|_2 \leq 1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2 + \delta(S)$. Noting that $\lim_{S\to\infty} \mathbb{P}(\mathcal{A}_S \cap \mathcal{B}_S \cap \mathcal{D}_\epsilon) = 1$, we obtain the conclusion of the theorem.

□

### 4.3.4 Third regime

In this section, we give the proof of Theorem 4.3.3. The proof of Theorem 4.3.3 follows a similar route as that of Theorem 4.3.2.

*Proof of Theorem 4.3.3.* We reuse the notations $\bar{D}_1, \bar{D}_2, \mathcal{M}, \mathcal{M}', \cdots$ from Sect. 4.3.3. Similar to what we did for the second regime, the proof splits into two major steps. For the first step, we upper bound $\|\mathcal{M} - \mathcal{M}'\|_2$; for the second step, we upper bound $\|\mathcal{M}'\|_2$.

**Step 1**: We start by bounding $\|\mathcal{M} - \mathcal{M}'\|_2$.

We take $S$ sufficiently large so that $\epsilon(S) < \frac{1}{4}$. As a consequence, $\bar{N}_{i.} \geq \frac{3}{4}\frac{S}{R}$ and $\bar{N}_{.j} \geq \frac{3}{4}\frac{S}{C}$ for any $i \in [R], j \in [C]$.

We take $\psi(S) = \min\{\frac{1}{20}, \frac{1}{\log(S+1)}\}$, and let $\mathcal{A}_S$ be the event that

$$\begin{aligned} \mid N_{i\bullet} - \bar{N}_{i.} \mid &\leq \frac{S}{R}\psi(S) \text{ for any } i \in [R], \\ \mid N_{\bullet j} - \bar{N}_{.j} \mid &\leq \frac{S}{C}\psi(S) \text{ for any } j \in [C]. \end{aligned} \tag{4.63}$$

By an argument similar to that in Sect. 4.3.3, we have

$$\lim_{S\to\infty} \mathbb{P}(\mathcal{A}_S^c) = 0. \tag{4.64}$$

For the following, we would restrict our attention to $\mathcal{A}_S$. Recalling $\bar{N}_{i.} \geq \frac{3}{4}\frac{S}{R}$ and $\bar{N}_{.j} \geq \frac{3}{4}\frac{S}{C}$, we obtain $N_{i\bullet} \geq \frac{3}{4}\frac{S}{R} - \frac{S}{R}\psi(S)$ and $N_{\bullet j} \geq \frac{3}{4}\frac{S}{C} - \frac{S}{C}\psi(S)$. We let $\delta_1(S) =$

$3(\psi(S) + \frac{R}{S}\lambda_A)$, $\delta_2(S) = 3(\psi(S) + \frac{C}{S}\lambda_B)$. For any $i \in [R]$,

$$|\frac{1}{N_{i\cdot} + \lambda_A} - \frac{1}{\bar{N}_{i\cdot}}| \le \frac{|N_{i\cdot} - \bar{N}_{i\cdot}| + \lambda_A}{(N_{i\cdot} + \lambda_A)\bar{N}_{i\cdot}} \le \frac{\frac{S}{R}\psi(S) + \lambda_A}{\frac{S}{R}(\frac{3}{4} - \frac{1}{20})\frac{3}{4}\frac{S}{R}} \le \delta_1(S)\frac{R}{S}.$$

Similarly, for any $j \in [C]$,

$$|\frac{1}{N_{\cdot j} + \lambda_B} - \frac{1}{\bar{N}_{\cdot j}}| \le \delta_2(S)\frac{C}{S}.$$

Hence we have that

$$\|D_1^{-1} - \bar{D}_1^{-1}\|_2 \le \delta_1(S)\frac{R}{S}, \quad \|D_2^{-1} - \bar{D}_2^{-1}\|_2 \le \delta_2(S)\frac{C}{S},$$

Now note that

$$|\sqrt{N_{i\cdot} + \lambda_A} - \bar{N}_{i\cdot}| \le \frac{|N_{i\cdot} - \bar{N}_{i\cdot}| + \lambda_A}{\sqrt{N_{i\cdot} + \lambda_A} + \sqrt{\bar{N}_{i\cdot}}} \le \frac{\frac{S}{R}\psi(S) + \lambda_A}{\sqrt{\frac{S}{R}(\frac{3}{4} - \frac{1}{20})}} \le \delta_1(S)\sqrt{\frac{S}{R}},$$

$$|\frac{1}{\sqrt{N_{i\cdot} + \lambda_A}} - \frac{1}{\sqrt{\bar{N}_{i\cdot}}}| = \frac{|\sqrt{N_{i\cdot} + \lambda_A} - \sqrt{\bar{N}_{i\cdot}}|}{\sqrt{N_{i\cdot} + \lambda_A}\sqrt{\bar{N}_{i\cdot}}} \le 2\delta_1(S)\sqrt{\frac{R}{S}}.$$

Hence

$$\|D_1^{\frac{1}{2}} - \bar{D}_1^{\frac{1}{2}}\| \le \delta_1(S)\sqrt{\frac{S}{R}}, \quad \|D_1^{-\frac{1}{2}} - \bar{D}_1^{-\frac{1}{2}}\| \le 2\delta_1(S)\sqrt{\frac{R}{S}}.$$

For any $i \in [R]$, we have

$$1 \ge s_i^{(1)} = \frac{1}{1 + \frac{\lambda_A}{N_{i\cdot}}} \ge \frac{1}{1 + \frac{\lambda_A}{(\frac{3}{4} - \frac{1}{20})\frac{S}{R}}} \ge \frac{1}{1 + \frac{2\lambda_A R}{S}}.$$

Using a similar argument as in equation (4.33), for any $i \in [R]$,

$$\frac{1}{R}\frac{1}{1 + \frac{2\lambda_A R}{S}} \le w_i^{(1)} \le \frac{1}{R}(1 + \frac{2\lambda_A R}{S}).$$

Similarly, for any $j \in [C]$,

$$\frac{1}{C}\frac{1}{1+\frac{2\lambda_B C}{S}} \le w_j^{(2)} \le \frac{1}{C}(1+\frac{2\lambda_B C}{S}).$$

Therefore, there exists $\delta_3(S) \to 0$, such that for any $i \in [R], j \in [C]$,

$$|w_i^{(1)} - \frac{1}{R}| \le \delta_3(S)\frac{1}{R}, \quad |w_j^{(2)} - \frac{1}{C}| \le \delta_3(S)\frac{1}{C}.$$

We assume $\delta_3(S) \le \frac{1}{4}$. From equation (4.63) we obtain $N_{i\bullet} \le (\frac{5}{4}+\frac{1}{20})\frac{S}{R}$ and $N_{\bullet j} \le (\frac{5}{4}+\frac{1}{20})\frac{S}{C}$. Lemma 4.3.3 along with the bounds on $N_{i\bullet}$ and $N_{\bullet j}$ leads to $\|Z\|, \|Z^\mathsf{T}\| \le \frac{2S}{\sqrt{RC}}$.

By Lemma 4.3.4, following the argument in Sect. 4.3.3, there is an absolute constant $K_1 > 0$, such that if we let $\mathcal{D}_\epsilon$ be the event that

$$\|Z - \bar{Z}\|_2 \le K_1\frac{1}{\epsilon}(\sqrt{\Upsilon}\sqrt{\frac{S}{R}} + \sqrt{\Upsilon}\sqrt{\frac{S}{C}} + \Upsilon^{\frac{1}{4}}S^{\frac{1}{4}}),$$

then $\mathbb{P}(\mathcal{D}_\epsilon^c) \le \epsilon$.

Noting that $0 \le w_i^{(1)} \le \frac{1}{R}(1+\delta_3(S))$ and $0 \le w_j^{(2)} \le \frac{1}{C}(1+\delta_3(S))$, we have

$$\begin{aligned}
\|I_R - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 &\le \|I_R\|_2 + \|\boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 \le 1 + \|\boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_F \le 4,\\
\|I_C - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 &\le \|I_C\|_2 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \le 1 + \|\boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_F \le 4.
\end{aligned}$$

We also have

$$\|\frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\mathsf{T} - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_2 \le \|\frac{1}{R}\mathbf{1}_R\mathbf{1}_R^\mathsf{T} - \boldsymbol{w}^{(1)}\mathbf{1}_R^\mathsf{T}\|_F \le \delta_3(S),$$

$$\|\frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\mathsf{T} - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_2 \le \|\frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\mathsf{T} - \boldsymbol{w}^{(2)}\mathbf{1}_C^\mathsf{T}\|_F \le \delta_3(S).$$

Using all the above results and the decomposition from Sect. 4.3.3, we deduce that for some $\delta(S) \to 0$ and some $\epsilon \to 0$, on $\mathcal{A}_S \cap \mathcal{D}_\epsilon$ we have $\|\mathcal{M} - \mathcal{M}'\|_2 \le \delta(S)$ and $\mathbb{P}((\mathcal{A}_S \cap \mathcal{D}_\epsilon)^c) \to 0$.

**Step 2**: In the following, we upper bound $\|\mathcal{M}'\|_2$. Letting $\mathcal{M}_1', \mathcal{M}_2'$ be defined similarly as in Sect. 4.3.3, we have

$$\|\mathcal{M}'\|_2 \leq \|\mathcal{M}'_1\|_2 \|\mathcal{M}'_2\|_2.$$

We bound $\|\mathcal{M}'_1\|_2$ in the following. Note that $(\mathcal{M}'_1)^\mathsf{T}\mathcal{M}'_1 \bar{D}_2^{\frac{1}{2}} 1_C = 0$. In the following, we take $u = (u_j)$ such that $\sum_{j=1}^C u_j \bar{N}_{.j}^{\frac{1}{2}} = 0$. Note that the equations (4.52) and (4.53) still hold for the third regime, with $S_0 = \sum_{l=1}^R \sum_{k=1}^C p_{lk} = \sum_{l=1}^R \bar{N}_{l.} \in [S(1-\epsilon(S)), S(1+\epsilon(S))]$.

Noting that $\frac{S}{R}(1-\epsilon(S)) \leq \bar{N}_{l.} \leq \frac{S}{R}(1+\epsilon(S))$, we obtain

$$|\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l.}} - \frac{R}{S}\bar{N}_{.s}| \leq 2\frac{R}{S}\epsilon(S)\bar{N}_{.s}. \tag{4.65}$$

Hence by (4.53) and (4.65),

$$\begin{aligned}
& \sum_{s=1}^C \sum_{k=1}^C \frac{u_s}{\sqrt{\bar{N}_{.s}}} \frac{u_k}{\sqrt{\bar{N}_{.k}}} \frac{S_0}{R^2} (\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l.}})(\sum_{l=1}^R \frac{p_{lk}}{\bar{N}_{l.}}) \\
= \; & \frac{S_0}{R^2}(\sum_{s=1}^C \frac{u_s}{\sqrt{\bar{N}_{.s}}}(\sum_{l=1}^R \frac{p_{ls}}{\bar{N}_{l.}} - \frac{R}{S}\bar{N}_{.s}))^2 \\
\leq \; & \frac{4\epsilon(S)^2 S_0}{S^2}(\sum_{s=1}^C |u_s|\sqrt{\bar{N}_{.s}})^2 \leq \frac{4\epsilon(S)^2 S_0}{S^2}(\sum_{s=1}^C \bar{N}_{.s})(\sum_{s=1}^C u_s^2) \\
\leq \; & \frac{4\epsilon(S)^2 S_0^2}{S^2}(\sum_{s=1}^C u_s^2) \leq 16\epsilon(S)^2(\sum_{s=1}^C u_s^2).
\end{aligned} \tag{4.66}$$

Let $P(s,k) = \sum_{l=1}^R \frac{p_{ls}p_{lk}}{\bar{N}_{l.}}$. Then

$$P(s,k) = \sum_{l=1}^R \frac{p_{ls}p_{lk}}{\bar{N}_{l.}} \geq R(\frac{S}{\Upsilon RC})^2 \frac{1}{\frac{5}{4}\frac{S}{R}} \geq \frac{S}{2\Upsilon^2 C^2} \geq \frac{1}{8\Upsilon^2 S}\bar{N}_{.s}\bar{N}_{.k}.$$

Note that the equation (4.56) still holds for the third regime. Hence

$$\begin{aligned}
& -\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s}{\sqrt{\bar{N}_{.s}}}\frac{u_k}{\sqrt{\bar{N}_{.k}}}P(s,k)+\sum_{s=1}^{C}u_s^2 \\
\geq\ & \frac{1}{16\Upsilon^2 S}\sum_{s=1}^{C}\sum_{k=1}^{C}\bar{N}_{.s}\bar{N}_{.k}\left(\frac{u_s}{\sqrt{\bar{N}_{.s}}}-\frac{u_k}{\sqrt{\bar{N}_{.k}}}\right)^2 \\
=\ & \frac{S_0}{8\Upsilon^2 S}\sum_{s=1}^{C}u_s^2-\frac{1}{8\Upsilon^2 S}\left(\sum_{s=1}^{C}\sqrt{\bar{N}_{.s}}u_s\right)^2\geq\frac{1}{16\Upsilon^2}\sum_{s=1}^{C}u_s^2.
\end{aligned}$$

where in the last line we have used that $S_0 \geq \frac{S}{2}$ and $\sum_{s=1}^{C}\sqrt{\bar{N}_{.s}}u_s = 0$. This leads to

$$\sum_{s=1}^{C}\sum_{k=1}^{C}\frac{u_s}{\sqrt{\bar{N}_{.s}}}\frac{u_k}{\sqrt{\bar{N}_{.k}}}P(s,k)\leq\left(1-\frac{1}{16\Upsilon^2}\right)\sum_{s=1}^{C}u_s^2. \tag{4.67}$$

Plugging inequalities (4.66) and (4.67) into (4.52), we obtain

$$\|\mathcal{M}_1'\|_2=\sqrt{\rho((\mathcal{M}_1')^\intercal\mathcal{M}_1')}\leq\sqrt{1-\frac{1}{16\Upsilon^2}+16\epsilon(S)^2}.$$

A similar bound can be obtained for $\|\mathcal{M}_2'\|_2$:

$$\|\mathcal{M}_2'\|_2\leq\sqrt{1-\frac{1}{16\Upsilon^2}+16\epsilon(S)^2}.$$

Hence for any $1 \leq \Upsilon < \infty$ independent of $S$, there exists $\delta_0 > 0$, such that for $S$ sufficiently large,

$$\|\mathcal{M}'\|_2 \leq 1-\delta_0.$$

Note that $\|\mathcal{M}-\mathcal{M}'\|_2 \leq \delta(S)$ on $\mathcal{A}_S \cap \mathcal{D}_\epsilon$. Noting that $\lim_{S\to\infty}\mathbb{P}(\mathcal{A}_S\cap\mathcal{D}_\epsilon) = 1$ and $\delta(S) \to 0$ as $S \to \infty$, we obtain the conclusion of the theorem.

□

## 4.4 Simulation studies

We start off by showing some trace plots of the Gibbs sampler and the collapsed Gibbs sampler in Figure 4.2 for 10000 iterations in simulated data for two different problem sizes.

Given $p_{ij}$, we generate our observation matrix via $Z_{ij} \overset{\text{ind}}{\sim} \text{Bern}(p_{ij})$. These probabilities are first generated via $p_{ij} = U_{ij} S^{1-\rho-\kappa}$ where $U_{ij} \overset{\text{iid}}{\sim} \mathbb{U}[1, 1.52]$. Note that 1.52 is the largest value of $\Upsilon$ up to second decimal digit for which $\phi(\Upsilon) = 1 - \frac{1}{\Upsilon^3} + (\Upsilon - 1)^2 < 1$. If $Z_{ij} = 1$, we generate $y_{ij}$ from the following model :

$$y_{ij} = 2 + a_i^{(1)} + a_j^{(2)} + e_{ij} \tag{4.68}$$

where $a_i^{(1)} \sim \mathcal{N}(0, \sigma_1^2)$ for each $1 \le i \le R$, $a_j^{(2)} \sim \mathcal{N}(0, \sigma_2^2)$ for each $1 \le j \le C$, and $e_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ for every $1 \le i \le R, 1 \le j \le C$. We take $(\rho, \kappa) = (0.52, 0.52)$ and $\sigma_1 = \sigma_2 = \sigma_E = 1$.

We denote $\mu_1 = \frac{1}{R} \sum_{i=1}^{R} a_i^{(1)}$ and $\mu_2 = \frac{1}{C} \sum_{j=1}^{C} a_j^{(2)}$. We assume Gaussian prior on mean parameters and known precision parameters. Figure 4.2 indicates that the collapsed Gibbs sampler for the mean parameter in the model in (4.10) with $\Upsilon' = 1$ and $\Upsilon = 1.52$ mixes faster than the Gibbs sampler. This is consistent with the observation made by Papaspiliopoulos et al. (2020).

The trace plots of Gibbs sampler gives a sense of serial correlation of the draws. We plot the autocorrelations in Figure 4.3.

We give a comparison of effective sample size (see Geyer (2011)) for the parameters for problem size $S = 10^\ell$ for $\ell$ varying in $\{3, 3.25, 3.5, \cdots, 5\}$.

Figure 4.4 shows the effective sample size for Gibbs sampler and collapsed Gibbs sampler for 10000 iterations after discarding first 1000 as burn-in.

Figure 4.4 suggests that the effective sample size for Gibbs sampler goes down with problem size, while the effective sample size of its collapsed version remains constant. As noted by Papaspiliopoulos et al. (2020), the mixing should depend on $(\rho, \kappa)$. We give a similar set of plots in Appendix C.2 for a different value of $(\rho, \kappa)$.

In the following, we study the convergence rate of the collapsed Gibbs sampler more quantitatively through the matrix norm of the autoregression matrix (according

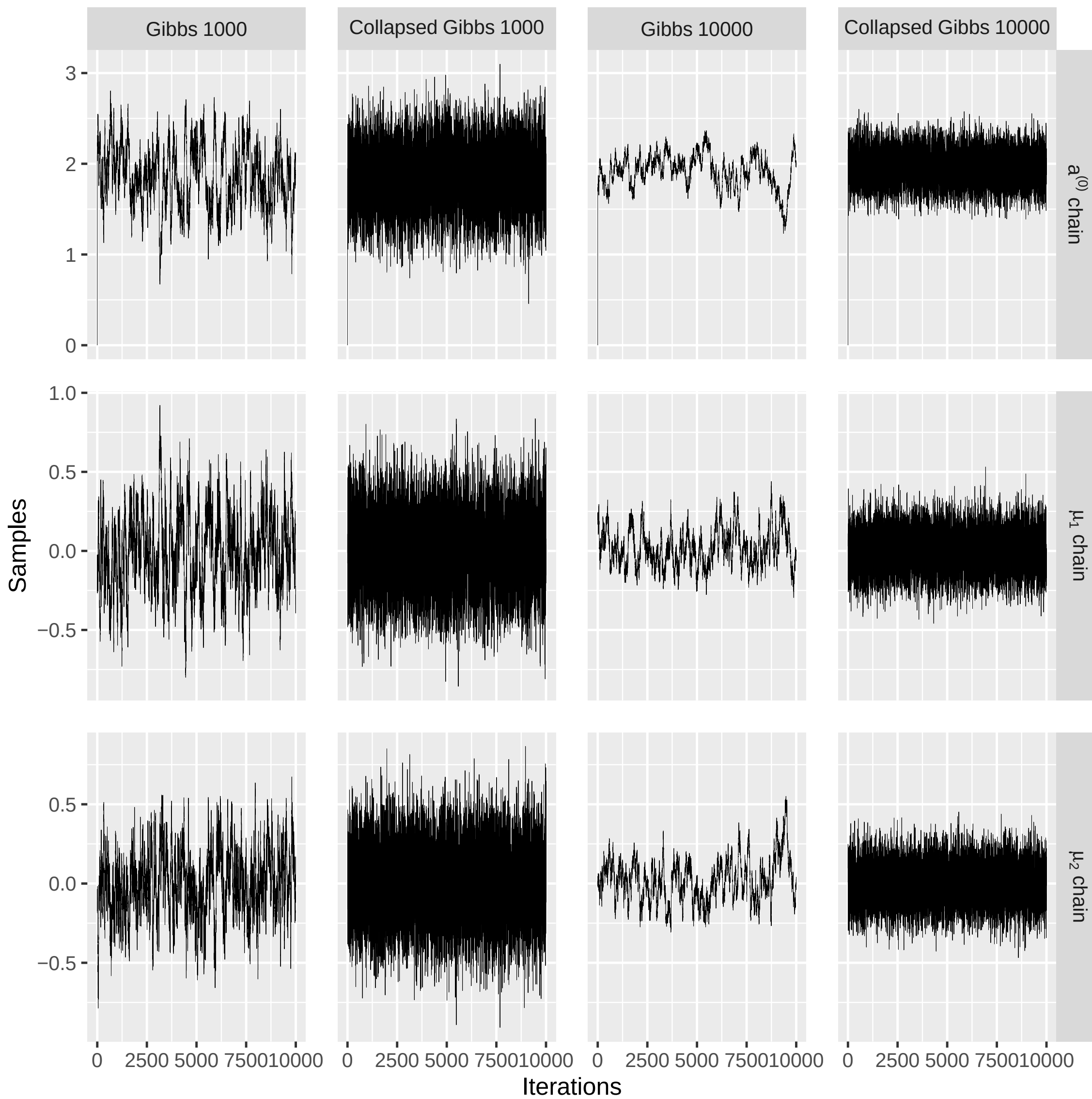



Figure 4.2: Trace plots of Gibbs sampler and collapsed Gibbs sampler for mean parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.52} \rceil, C = \lceil S^{0.52} \rceil$. This indicates poor mixing of Gibbs sampler and increasing complexity with problem size.

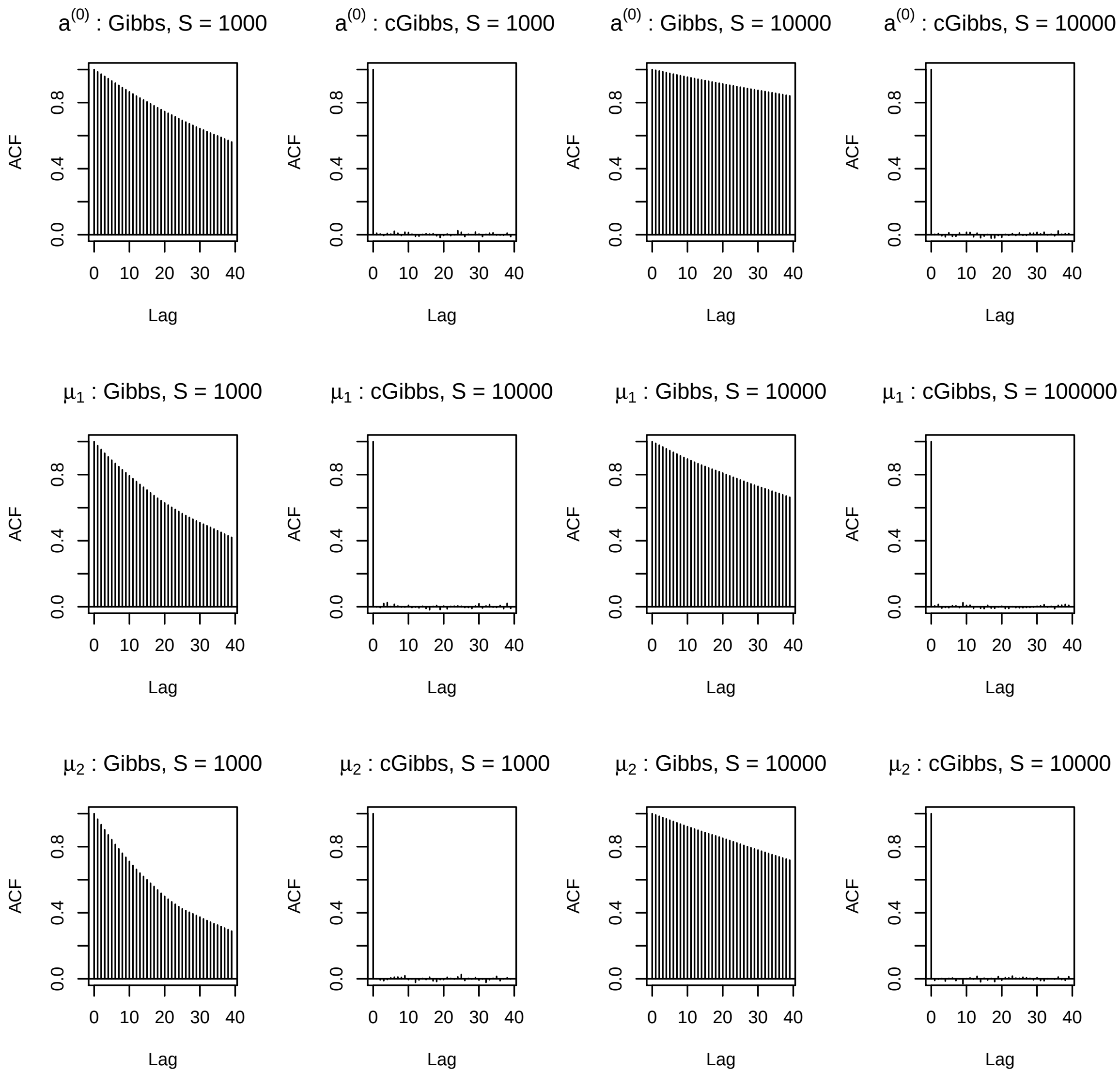



Figure 4.3: Autocorrelation plot of Gibbs sampler and collapsed Gibbs (cGibbs) sampler for mean parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.52} \rceil, C = \lceil S^{0.52} \rceil$. This indicates that the autocorrelation of the Gibbs sampler is significantly higher than that of the collapsed Gibbs sampler.

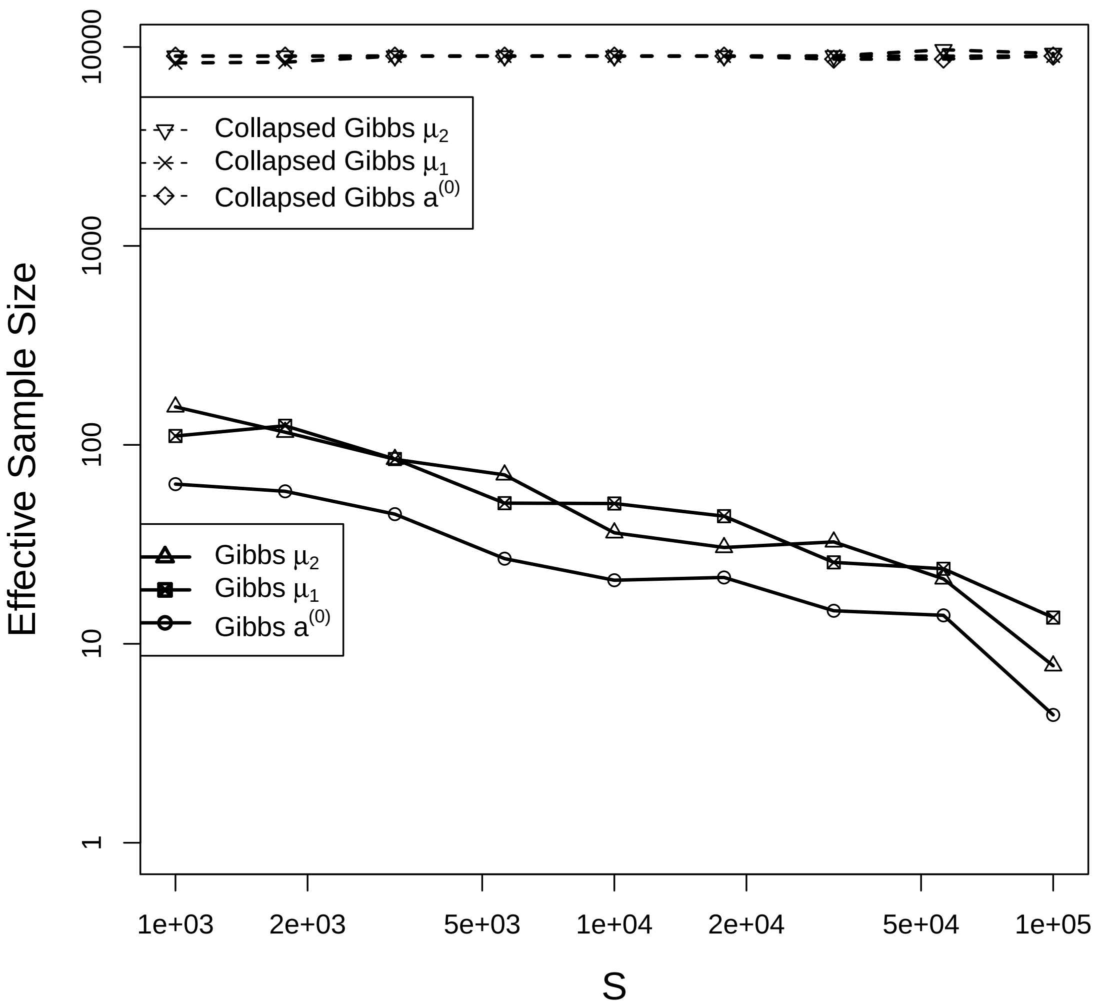



Figure 4.4: Effective sample size for mean parameters of crossed random effects of Gibbs and collapsed Gibbs sampler versus problem sizes ($S$) under the model (4.68) with $R = \lceil S^{0.52} \rceil, C = \lceil S^{0.52} \rceil$. The plot has a logarithmic horizontal and vertical scale.

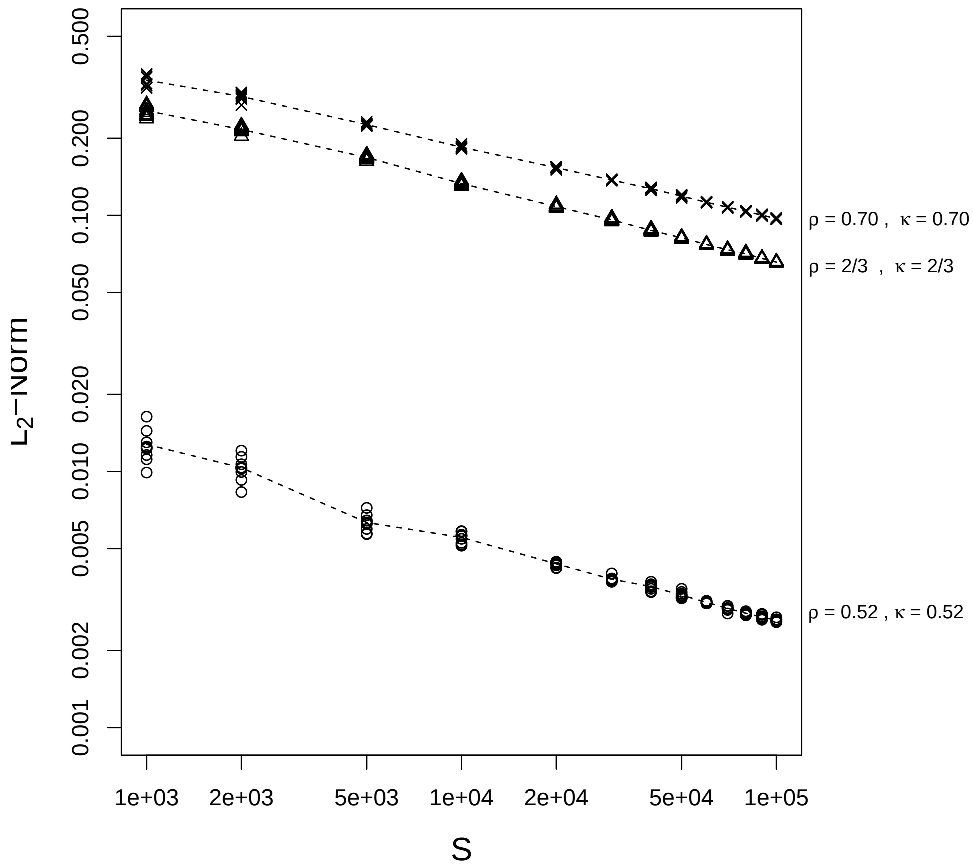



Figure 4.5: $L_2$ norm versus $S$ with a logarithmic vertical scale.

to the results in Sect. 4.2). We sample from the model multiple times at various values of $S$ and plot $\|M\|_2$ versus $S$ on a logarithmic scale. Figure 4.5 shows the results. We observe that $\|M\|_2$ is below 1 and decreasing with $S$ for all the examples $(\rho, \kappa)$ that satisfies the condition in (4.6). We also look at $(\rho, \kappa) = (0.7, 0.7)$ on the 45 degree line that is outside the triangle in (4.6). The behavior of $\|M\|_2$ shows that the condition in (4.6) is a sufficient condition, but might not be necessary.

The values of $\lambda_A$ and $\lambda_B$ appear in expressions $N_{i\bullet} + \lambda_A$ and $N_{\bullet j} + \lambda_B$ where their contribution is asymptotically negligible, so conservatively setting them to zero will nonetheless be realistic for large data sets.

## 4.5 Application on Real Data

We illustrate the two samplers on some data from Stitch Fix. The same data was used in section 2.6 of the second chapter.

Using our previous notations for crossed random effects models, we describe the details of the data set as follows. Recall there are $N$ = 5,000,000 ratings in total, produced by $R$ = 762,752 customers on $C$ = 6,318 items. These correspond to $(\rho, \kappa) = (0.88, 0.57)$, which does not satisfy the condition (4.6). The response variable $y_{ij}$ is the rating of satisfaction (on a ten point scale) of customer $i$ on item $j$. Features related to the customers and the items are also included in the data set. However, as our purpose here is to study and compare large scale Gibbs sampling and collapsed Gibbs sampling, we use the following basic model. The goal of this analysis is to test the algorithms, and the following model in (4.69) is not a very unreasonable model.

$$y_{ij} = a^{(0)} + a_i^{(1)} + a_j^{(2)} + e_{ij} \tag{4.69}$$

for customer $i$ and item $j$.

We used a standard flat prior for all of the precision parameters, $\tau_E, \tau_1$ and $\tau_2$. That is, we assumed $p(\tau^{-\frac{1}{2}}) \propto 1$ for $\tau \in \{\tau_E, \tau_1, \tau_2\}$ following Papaspiliopoulos et al. (2020). We alternately sample precision parameters $\tau$ from the conditional distribution $\mathcal{L}(\tau \mid \boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)})$ and update $(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)})$ with the Gibbs Sampler and its collapsed version, respectively. In Figures 4.6 and 4.7, we give the trace plots for the global mean, the means of the two random effects and the precision parameters for the Gibbs sampler and the collapsed Gibbs sampler for 10000 iterations on Stitch Fix data. We discard the first 1000 samples as burn-in. We also tabulate the effective sample size in Table 4.1 and note that collapsed Gibbs sampler has a much larger amount of effective sample size. The similar effective sample size for $a^{(0)}$ and $\mu_2$ for Gibbs sampler might be resulting from the strong correlation of the two chains as observed in the Figure C.6 in Appendix C.2 .

Gelman and Shirley (2011) recommends an effective sample size of about 100 posterior draws. Vats et al. (2019) advocates even greater effective sample sizes. Clearly Gibbs sampler fails to achieve that within 10000 iterations. These experiments

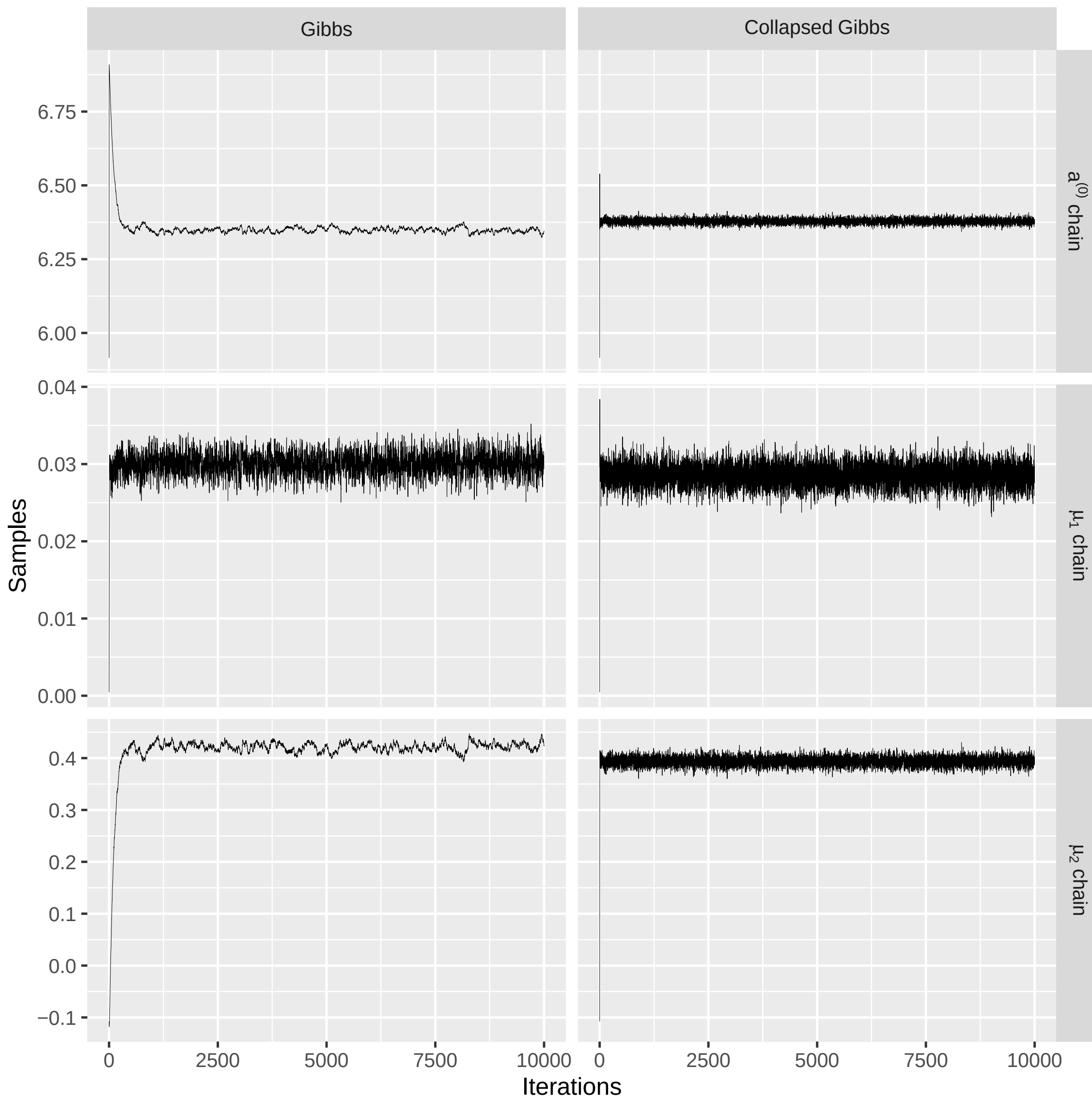



Figure 4.6: Trace plots for mean parameters of Gibbs and collapsed Gibbs sampler for crossed random effects model on Stitch Fix data.

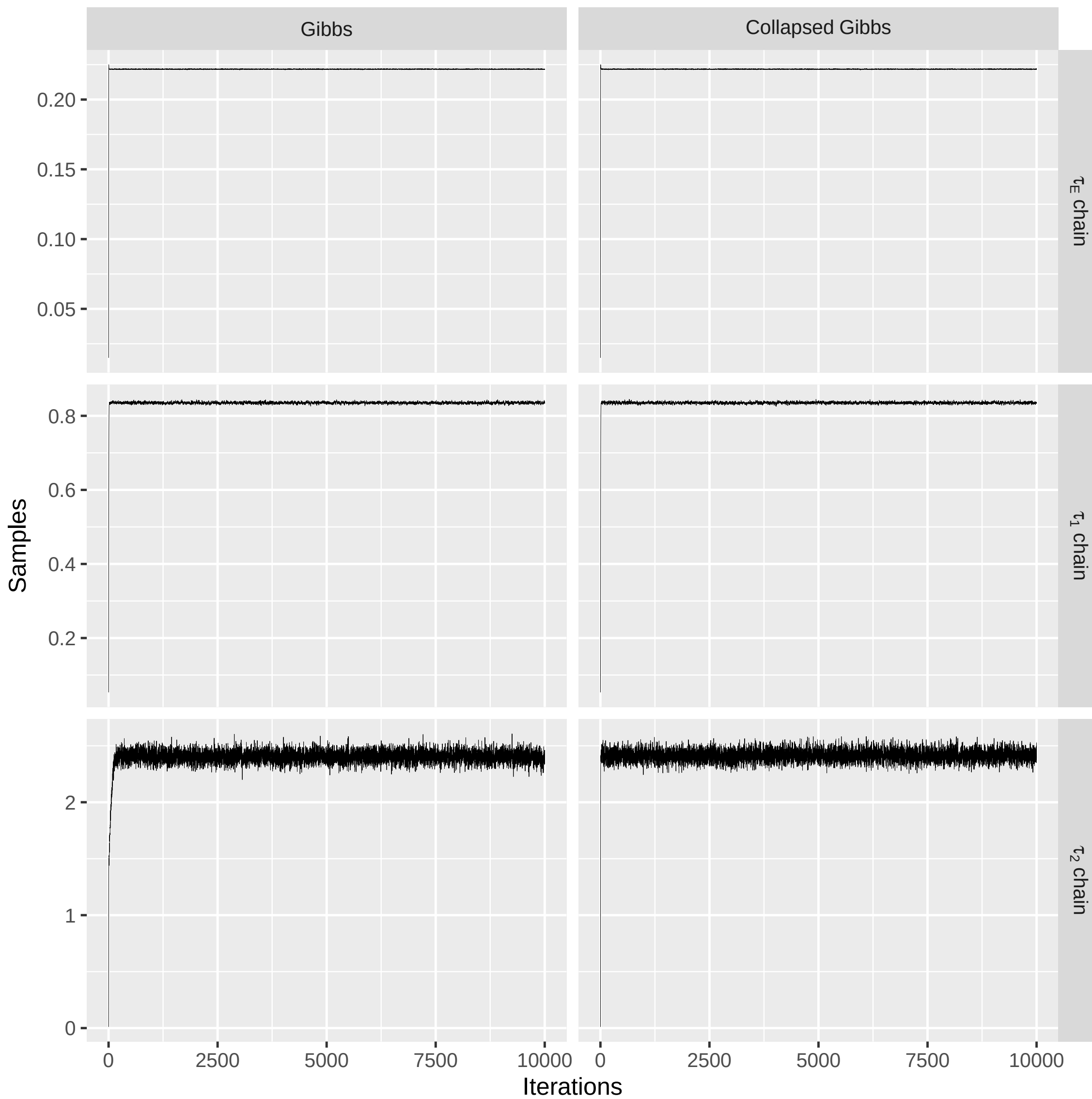



Figure 4.7: Trace plots for precision parameters of Gibbs and collapsed Gibbs sampler for crossed random effects model on Stitch Fix data.

| Parameter | Gibbs | Collapsed Gibbs |
|---|---|---|
| $a^{(0)}$ | 67 | 9000 |
| $\mu_1$ | 1745 | 9000 |
| $\mu_2$ | 62 | 8566 |
| $\tau_E$ | 6025 | 6033 |
| $\tau_1$ | 1746 | 1821 |
| $\tau_2$ | 4749 | 5302 |

Table 4.1: Comparison of effective sample size for different sampling schemes on the Stitch Fix data. Numbers are from 10000 iterations for each scheme, discarding the first 1000 samples as burn-in.

were carried out in R on a computer with the macOS operating system, 16 GB of memory and an Intel i7 processor.

## 4.6 Discussion

In this work, we have given sufficient conditions for finite relaxation time of a collapsed Gibbs sampler applied to crossed random effects models with unbalanced levels and a missingness mechanism. Our focus is on the Bayesian paradigm, and is therefore substantially different from the approach undertaken in the second chapter, in which we obtained the GLS estimate in a scalable fashion under less stringent conditions compared to the "balanced levels" condition. We studied the $L_1$ norm of the backfitting matrix in the second chapter and found in practice $L_2$ norm behaves desirably for scalability compared to the $L_1$ norm. We also saw that the $L_2$ norm and spectral radius are in a close range. Studying the $L_2$ norm using the tools from random matrix theory lets us weaken the condition on the missingness mechanism that has been considered in theorem 2.4.1. Moreover, our proof strategy allows us to incorporate an arbitrary inhomogeneity level (measured by $\Upsilon$) for missingness pattern when the row sums and column sums are "almost balanced" in expectation. Such kind of results are new to the best of our knowledge. The convergence rate of the collapsed Gibbs sampler when the data satisfy the "balanced levels" condition was obtained by (Papaspiliopoulos et al., 2020, Theorem 4). This requires that $N_{i\cdot} = \frac{N}{R}$ for all $i = 1, \cdots, R$ and $N_{\cdot j} = \frac{N}{C}$ for all $j = 1, \cdots, C$. In their result, the convergence rate (denoted by $\rho_{PRZ}$) of the collapsed Gibbs sampler is expressed through the convergence rate $\rho_{aux}$ of an auxiliary process.

Compared to their results, our results work for several regimes with unbalanced levels, and give unconditional (i.e., without assuming finite relaxation time of any auxiliary process) convergence results. We also believe that the asymptote like the one that we study where $R, C, N$ simultaneously grow is a better description for sparse data coming from electronic commerce.

There can be multiple ways to go forward. One immediate direction this work can move forward is to obtain a characterisation of the convergence rate of the collapsed Gibbs sampler when the number of crossed random effects is bigger than 2. The progress made here turns out to be useful for this problem which we discuss in the next chapter.

# Chapter 5

# Coordinatewise Inference for Crossed Random Effect Model

## 5.1 Introduction

In this chapter we develop complexity theory for algorithms used to make crossed random effect models.

A general version of these models is as follows:

$$\begin{aligned} \eta_{i_1\cdots i_K} &= x_{i_1\cdots i_K}^T a^{(0)} + \sum_{k=1}^{K} z_{i_1\cdots i_K}^T a_{i_k}^{(k)} \\ y_{i_1\cdots i_K} \mid \eta_{i_1\cdots i_K} &\sim p(y_{i_1\cdots i_K} \mid \eta_{i_1\cdots i_K}). \end{aligned} \tag{5.1}$$

There are $K$ categorical covariates, which we will call factors, each with $p_k$ levels, the crossing of which defines a high-dimensional contingency table. We are interested in situations where at least one of the factors has a large number of levels. Each cell of the table is indexed by the level of each factor, $i_1 \cdots i_K$. For each available observation the factors determine the cell it belongs to, and there is a response variable $y_{i_1 \ldots i_K}$. For each observation there might be additional numerical covariates available, denoted in (5.1) by $x_{i_1\cdots i_K}$ and $z_{i_1\cdots i_K}$. The coefficients $a^{(0)}$ are known as the fixed effects, in the contexts we are interested in these are low-dimensional and in the formulation we adopt are given a flat prior distribution, $p(a^{(0)}) \propto 1$. The coefficients $a_i^{(k)}$ are known as the random effects and they are modeled via Gaussian distributions, $a_i^{(k)} \overset{i.i.d}{\sim} \mathcal{N}(0, \tau_k^{-1})$. We will denote by $a$ the set of all regression coefficients, $y$ the set of all available responses, and $p(a \mid y)$ the posterior density resulting from (5.1) and the prior on the coefficients.

The theory in this chapter is particularly motivated by problems where both $n$ and $p$ are large and the contingency matrix is sparsely observed, despite the fact that none of our arguments is asymptotic in nature. A couple of examples can provide some intuition for the problems we are interested in with more details provided in the aforementioned references. In the context of survey data as typically arise in political science one factor might be the interaction term between geographical administrative unit and race, in which case there will be hundreds of levels. In the context of recommendation one factor might be client and another product, in which cases there will be hundreds of thousands of levels.

We consider the three canonical inference problems and corresponding algorithms associated to (5.1), the details of which are given in Section 5.2.2 together with references in the literature. The first is generalized least squares inference for the regression coefficients, which amounts to maximizing $p(a \mid y)$, using backfitting. The second is Bayesian inference by Gibbs sampling from $p(a \mid y)$, and the third is approximate Bayesian inference by approximating $p(a \mid y)$ by a product of densities using coordinate ascent variational inference (CAVI). All three algorithms are iterative schemes that involve coordinate-wise updates, either in terms of optimization or sampling. From the theory point of view that we adopt in this chapter, there are two fundamental complexity parameters. The first is the total number of observations that we will denote by $n$. The second complexity parameter is the number of regression coefficients $a$ that we will denote by $p$ and is proportional to $\sum_k p_k$.

The goal of this chapter is to study the geometric rate of convergence of the aforementioned coordinate-wise updating algorithms for $K \geq 2$. Previous work, in particular Papaspiliopoulos et al. (2021); Ghosh et al. (2022) has established that for crossed random effect models when the rate of convergence is bounded away from 1 as $n$ and $p$ vary, then the overall computational complexity of these algorithms is

$\mathcal{O}(n+p)$, a property that we will refer to as scalability. A main result states that under certain assumptions, and for typical data design the convergence rate of certain implementations of the algorithms is bounded away from 1, hence the algorithms are scalable. In fact, we obtain a stronger result that in high-dimensional regimes where both $n$ and $p$ are large, the rate goes to 0, therefore there is in effect a blessing of dimensionality. The result is quite subtle for two main reasons we outline below.

First, Laplace approximation methods even when they employ sparse linear algebra methods are typically not scalable for these problems. This has already been observed empirically for common software implementations, see for example the recent works in Gao and Owen (2017); Papaspiliopoulos et al. (2020); Goplerud (2021). Papaspiliopoulos et al. (2021) obtain various results that confirm the inappropriateness of such methods in this context. For $K = 2$ they prove that under the same conditions we assume in this chapter the complexity of methods based on sparse linear algebra is at least $\mathcal{O}(n^2/p)$. Therefore, in a recommender system-type high-dimensional regime where $p = \mathcal{O}\left(n^{1/2}\right)$, their result implies superlinear complexity. They also give concrete designs for which the complexity is provably $\mathcal{O}(p^3)$, as well numerical evidence for such complexity with completely missing at random designs. By design we mean the data are organized in the contingency table, see Section 5.2.2.

Second, the common implementations of the three iterative algorithms are not scalable either, as we discuss in Section 5.2.2. It is slightly modified, yet easily implementable, versions of these algorithms that achieve scalability.

Partial steps in the direction of our chapter were made in Papaspiliopoulos et al. (2020) and Ghosh et al. (2022) for $K = 2$. Papaspiliopoulos et al. (2020) related the rate of convergence of Gibbs sampling to that of an auxiliary random walk on a random bipartite graph and it conjectured that the latter mixes fast. Ghosh et al. (2022) established that the rate of convergence of backfitting is bounded away from 1 for a set of conditions weaker than those of Papaspiliopoulos et al. (2020). In this chapter we relate the rate of convergence of both schemes and that of CAVI to auxiliary random walks on random bipartite graphs for any $K$ using a common unifying framework for all three schemes. Additionally, we leverage very recent results for the fast mixing of these random walks (a class of results related to the so-called Alon's conjecture) to prove the conjecture in Papaspiliopoulos et al. (2020) and establish that the algorithms are scalable for all values of $K$.

## 5.2 Algorithms, models and the main result

### 5.2.1 Coordinate-wise updating algorithms for Gaussian distributions

We start with a basic observation that allows us to study all three algorithms, i.e., backfitting, Gibbs sampling and CAVI, simultaneously, and can be checked by first principles. Consider generally a multivariate Gaussian density,

$$p(x) \propto \exp\{-x^T Qx/2 + x^T b\}, \tag{5.2}$$

with $x = (x^{(1)}, \ldots, x^{(K)})$, and $b$ and $Q$ organized accordingly as block vector and matrix respectively. Throughout the chapter negative superscripts denote all components but the indexed one, for example $x^{(-1)} = (x^{(2)}, \ldots, x^{(K)})$. In the sequel we also assume that the diagonal blocks $Q^{(k,k)}$ are invertible. Backfitting amounts to coordinate-wise ascent to find $m := \arg\max_x p(x)$, and involves the updates:

$$m^{(k)} \leftarrow -(Q^{(k,k)})^{-1} \sum_{\ell \neq k} Q^{(k,\ell)} m^{(\ell)} + (Q^{(k,k)})^{-1} b^{(k)}. \tag{5.3}$$

Gibbs sampling generates a Markov chain that asymptotically samples from $p(x)$ by sampling from the conditional densities and involves the updates:

$$x^{(k)} \leftarrow -(Q^{(k,k)})^{-1} \sum_{\ell \neq k} Q^{(k,\ell)} x^{(\ell)} + (Q^{(k,k)})^{-1} b^{(k)} + \epsilon^{(k)}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, (Q^{(k,k)})^{-1}). \tag{5.4}$$

Mean field variational inference approximates $p(x)$ by a product density $q(x) = \prod_k q_k(x^{(k)})$ chosen to minimize the Kullback-Leibler divergence between $p$ and $q$ (the version of the divergence where $q$ is the dominating measure). This is equivalent to maximizing the so-called variational lower bound. CAVI does this by coordinate-wise ascent where each density $q_k$ is optimized given the other densities. Calculations show (see for example Chapter 10 of Bishop (2006)) that the optimal densities are $\mathcal{N}(\mu^{(k)}, (Q^{(k,k)})^{-1})$ and CAVI involves the following updates:

$$\mu^{(k)} \leftarrow -(Q^{(k,k)})^{-1} \sum_{\ell \neq k} Q^{(k,\ell)} \mu^{(\ell)} + (Q^{(k,k)})^{-1} b^{(k)}. \tag{5.5}$$

The derived updates show how to change a block given the values of the remaining ones. One iteration of each algorithm scans the blocks in the given order and performs the corresponding update. It is evident from the previous developments the common linear autoregressive structure in the resultant updates. If we denote by $m_t, x_t, \mu_t$ the vectors obtained after $t$ iterations of backfitting, Gibbs sampling and CAVI respectively, then we can solve the linear equations to obtain that:

$$\begin{aligned} m_t &= Bm_{t-1} + \gamma, \\ x_t &= Bx_{t-1} + \gamma + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Gamma), \\ \mu_t &= B\mu_{t-1} + \gamma, \end{aligned} \tag{5.6}$$

for certain $\gamma$ and $\Gamma$, and autoregressive matrix

$$\begin{aligned} &B = (\boldsymbol{I} - L)^{-1} U, \quad L = \text{lower}(A), \quad U = I - L \\ &A = \boldsymbol{I} - \text{diag}(Q^{(1,1)}, \dots, Q^{(K,K)})^{-1} Q. \end{aligned} \tag{5.7}$$

This leads to the following result, which we organize in the following Proposition for convenience.

*Proposition* 5.2.1. The processes $m_t, x_t, \mu_t$ as defined in (5.6) converge as $t \to \infty$ if and only if $\rho(B) < 1$, and when so $\rho(B)$ is their geometric rate of convergence, where $B$ is defined in (5.7), $\rho(B)$ is the largest modulo eigenvalue of $B$, and the mode of convergence is the vector $L^2$ norm for $m_t, \mu_t$ and the operator norm defined on the space of square integrable functions for $x_t$.

The result for $m_t, \mu_t$ can be proved from first principles, that for $x_t$ is more involved, see for example the proof of Theorem 1 in Roberts and Sahu (1997). Throughout the chapter, for any $n \in \{1, 2, \cdots\}$, we denote by $\boldsymbol{I}_n$ the identity matrix of size $n$.

Sometimes we suppress the subscript when the dimension is evident from the context.

### 5.2.2 Crossed effect models and inference algorithms

To facilitate theoretical analysis we restrict the scope of the general mixed model formulation in (5.1) and make certain assumptions. We focus on a model without continuous covariates, hence a pure analysis of variance model, and with Gaussian likelihood. Therefore, each regression coefficient is now one-dimensional, and the model can be written as:

$$\begin{aligned} y_{i_1\cdots i_K} \mid a &\sim \mathcal{N}(a^{(0)} + \sum_{k=1}^{K} a^{(k)}_{i_k}, \tau^{-1}) \\ a^{(k)}_i &\sim \mathcal{N}(0, \tau_k^{-1}), \end{aligned} \tag{5.8}$$

with $p(a^{(0)}) \propto 1$.

As we point out in the previous chapter, our findings and their methodological implications in the intercept only case are likely to apply to more general linear mixed models including continuous covariates and non-Gaussian probability distributions as seen in Papaspiliopoulos et al. (2021). The inference algorithms for such model would need to also learn the variance parameters, either by optimization, sampling or variational approximation. However, this turns out to be an easier problem in terms of complexity, see for example Ghosh et al. (2022); Papaspiliopoulos et al. (2021). This assumption also facilitates the approach we develop in this chapter, since in this case the algorithms target a high-dimensional multivariate Gaussian distribution. Thus, we focus on the problems of maximizing, sampling and mean-field approximating $p(a \mid y)$ obtained from (5.8).

Notice that in this framework $p(a \mid y)$ is a $p$-dimensional Gaussian density as in (5.2). In terms of the blocking framework of Section 5.2.1, we organize $a$ in blocks $a^{(0)}$ (one-dimensional), $a^{(1)}$ ($p_1$-dimensional), $a^{(2)}$ ($p_2$-dimensional) etc. The resultant posterior has a (typically) sparse precision matrix $Q$ with

$$Q^{(0,0)} = n\tau, \quad Q_i^{(0,k)} = n_i^{(k)}\tau, \quad Q_{i,j}^{(k,k)} = (\tau_k + n_i^{(k)}\tau)\ \mathbb{1}_{i=j}, \quad Q_{i,j}^{(k,\ell)} = n_{ij}^{(k,\ell)}\tau\ \mathbb{1}_{k\neq\ell}. \tag{5.9}$$

The expressions above involve co-occurrence counts associated to the data design. These are defined below together with a key design assumption that allows us to link the convergence of coordinate-wise algorithms to random walks on random graphs.

*Definition* 5.2.1. Let $n_{ij}^{(k,\ell)}$ be the number of observations with level $i$ on the $k$-th factor and level $j$ on the $\ell$-th factor, and $n_i^{(k)} = \sum_j n_{ij}^{(k,\ell)}$ (the equality being true for any $\ell$). We say that the data design has balanced levels when $n_i^{(k)} = n/p_k := n^{(k)}$ for each $k$. We define $C_{p_k} = \boldsymbol{I} - (1/p_k)\mathbf{1}\mathbf{1}^\intercal$ to be a $p_k \times p_k$ centering matrix.

One can also easily obtain the $b$ term in (5.2) but since it does not affect the rate of convergence we skip the details. Given that the blocks $Q^{(k,k)}$ are diagonal (regardless the balancedness assumption), it is trivial to obtain the explicit updates for the three algorithms by the general formulation of Section 5.2.1.

Our first complexity result about the iterative schemes is a negative one. The following is obtained from Theorem 3 of Papaspiliopoulos et al. (2020) for the Gibbs sampler, but due to Proposition 5.2.1 applies to all three algorithms.

*Proposition* 5.2.2. For $Q$ defined in (5.9), under balanced levels and $\rho(B)$ defined in Proposition 5.2.1,

$$\rho(B) \geq \max_k \frac{n\tau}{n\tau + p_k\tau_k}.$$

The result implies that typically their rate of convergence is not bounded away from 1 as a function of $n$ and $p$. This result has spurred methodological work to produce modified coordinate-wise updating algorithms with better complexity, see for example Papaspiliopoulos et al. (2020, 2021) for Gibbs sampling, Ghosh et al. (2022) for backfitting.

### 5.2.3 Scalable methods: collapsing, constraining and other variations

We now introduce two variations of the algorithms for which we establish that their rate not only is bounded away from 1 but actually goes to 0 as $n$ and $p$ grow. There are other variations for which such a positive result will hold. For example, the local centering introduced in Papaspiliopoulos et al. (2021) and can be easily implemented even for general mixed models and non-Gaussian likelihood. The two variations we consider here are somewhat canonical for Gaussian crossed effect models and under balanced levels they have exactly the same rate of convergence, as we prove in this Section.

The first variation is what we will call their collapsed version. This simply consists of applying the algorithms to optimize, sample and mean-field approximate the posterior density that is obtained by marginalizing out (collapsing) the one-dimensional $a^{(0)}$, that is $p(a^{(-0)} \mid y)$. (Recall the notation on superscripts established after (5.2)). The resultant marginal density is Gaussian with precision matrix that we denote by $R$, whose elements are obtained simply as $R_{i,j}^{(k,\ell)} = Q_{i,j}^{(k,\ell)} - Q_i^{(k,0)} Q_j^{(0,\ell)} / Q^{(0,0)}$. Specializing the expressions for crossed effect models and assuming balanced levels we obtain that $R$ has elements:

$$R_{i,j}^{(k,\ell)} = (\tau n_{ij}^{(k,\ell)} - \tau n/(p_k p_\ell))\ \mathbb{1}_{k\neq\ell}, \quad R_{i,i}^{(k,k)} = \tau_k + \tau n/p_k - \tau n/p_k^2, \quad R_{i,j}^{(k,k)} = -\tau n/p_k^2\ \mathbb{1}_{i\neq j}. \tag{5.10}$$

for $0 < k, \ell \leq K$.

Notice that under balanced levels:

$$R^{(k,\ell)} = C_{p_k} Q^{(k,\ell)} = Q^{(k,\ell)} C_{p_\ell}, \quad \forall\, k, \ell \tag{5.11}$$

Also $R^{(k,k)}$ is now not diagonal but it is trivially invertible, hence all schemes can be easily implemented. Specifically, the cost per iteration is comparable to that of the plain vanilla implementation. We denote by $B_{coll}$ the resulting $B$-matrix, which is defined as in (5.7) with $Q$ replaced by $R$. The precise analysis of the resultant $\rho(B_{coll})$ for $K \geq 2$ is the content of the following section. The main highlight of this analysis is the provable scalability of this scheme.

However, other implementations are also expected to be provably scalable under the same conditions and can be analyzed with arguments analogous to those we develop in this chapter. Among those we highlight one that under balanced levels turns out to have exactly the same rate of convergence as the collapsed one. It is based on imposing an identifiability constraint to the model, hence we will refer to it as the constrained implementation. A natural constraint is to require $\sum_{i=1}^{p_k} a_i^{(k)} = 0$ for every $k = 1, \ldots, K$. Notice that this is not simply a different implementation, but rather a different prior hence a different model. The separability of the constraint across $k$ makes it trivial to incorporate it at each block update, resulting again in a straightforward implementation. We have the following result proof of which reveals details about the implementation of this scheme.

*Theorem* 5.2.1. Consider the coordinate-wise updating inference algorithms for the model (5.8) where the prior on the levels of each factor is further conditioned on $\sum_{i=1}^{p_k} a_i^{(k)} = 0$. Then, under balanced levels the rate of convergence of these schemes is $\rho(B_{coll})$, for $B_{coll}$ obtained from $R$ according to (5.7) and $R$ as in (5.10).

*Proof.* We establish the result for the Gibbs sampler, which given the equivalence of the rates for Gaussian distributions implies that for backfitting and CAVI. We obtain the result through a number of important observations:

1. The update of $a^{(0)}$ in the constrained version is same as in the plain vanilla and is according to the conditional distribution:

$$\mathcal{N}(\bar{y} - (1/n)\sum_k \sum_i a_i^{(k)} n_i^{(k)}, (n\tau)^{-1})$$

   where $\bar{y}$ is the sample average of the responses. Note that under balanced levels and if the factor levels are conditioned to zero-sum, this conditional becomes independent of the factor levels. Therefore the updates for $a^{(0)}$ converge after the first update, and the convergence of the algorithms depends solely on the properties of the updates of $a$.

2. A direct calculation gives that

$$(R^{(k,k)})^{-1} = (Q^{(k,k)})^{-1}(\boldsymbol{I} + \frac{n\tau}{\tau_k p_k^2}\mathbf{1}\mathbf{1}^\top)$$

and by the definitions and (5.11)

$$(R^{(k,k)})^{-1}R^{(k,\ell)} = (Q^{(k,k)})^{-1}Q^{(k,\ell)}C_{p_\ell}$$

thus, the update of $a^{(k)}$ in the collapsed implementation is according to:

$$\mathcal{N}\left(-(Q^{(k,k)})^{-1}\sum_{\ell\neq k}Q^{(k,\ell)}C_{p_\ell}a^{(\ell)} + \phi^{(k)}, \frac{p_k}{n\tau + \tau_k p_k}(I + \frac{n\tau}{\tau_k p_k^2}\mathbf{1}\mathbf{1}^\top)\right)$$

where $\phi^{(k)}$ is the appropriate intercept in the update the exact form of which is not relevant for this argument. The main thing to notice above is that it is not the $a^{(\ell)}$ that enter the formula but rather $C_{p_\ell}a^{(\ell)}$. The same properties used earlier yield that the conditional distribution of $C_{p_k}a^{(k)}$ is

$$\mathcal{N}\left(-(Q^{(k,k)})^{-1}\sum_{\ell\neq k}Q^{(k,\ell)}C_{p_\ell}a^{(\ell)} + C_{p_k}\phi^{(k)}, \frac{p_k}{n\tau + \tau_k p_k}C_{p_k}\right) \tag{5.12}$$

3. The update of $a^{(k)}$ in the constrained implementation is obtained by conditioning on the zero-sum constraint the conditional distribution in the plain vanilla implementation, and when we also use the balanced level assumption we obtain that the appropriate conditional is precisely given in (5.12). Therefore, the collapsed implementation does the same updates as those of the constrained with the difference that it adds different noise, this however is irrelevant for the convergence since in consequent steps this is cancelled by the pre-multiplication by the centering matrices. Rigorously, the equivalence of the convergence rates is established by appealing to the results about de-initializing chains, as developed in Roberts and Rosenthal (2001).

□

Therefore, although they infer slightly different models and in different ways the collapsed and the constrained implementation have exactly the same rate of convergence under the balanced levels assumption. They also have comparable costs per iteration. See also Zanella and Roberts (2021) for more discussion on alternative constraints and their computational implications.

## 5.3 Bounding the rate of convergence of coordinatewise inference algorithms by that of random walks on random bipartite graphs

The following theorem relates the spectral radius of $B_{coll}$ to the second largest eigenvalues of the transition kernels random walks on the bipartite contingency tables. More specifically, define the $p_k \times p_\ell$ stochastic matrix $P^{(k\to\ell)}$ as

$$P^{(k\to\ell)}_{i,j} = n^{(k,\ell)}_{ij}/n^{(k)}_i \qquad\qquad i = 1,\dots,p_k;\ j = 1,\dots,p_\ell$$

for every $\ell \neq k$, where we can assume $n^{(k)}_i > 0$ for every $k$ and $i$ without loss of generality. Then

$$P^{(k,\ell)} := P^{(k\to\ell)}P^{(\ell\to k)}$$

can be interpreted as the transition kernel of a Markov chain on $\{1,\dots,p_k\}$. At each iteration, such chain moves from a level of $k$-th factor to a level of the $\ell$-th factor and then back. Let $\lambda_2(P^{(k,\ell)})$ denote the second largest modulus eigenvalue of $P^{(k,\ell)}$.

*Theorem* 5.3.1. Under balanced levels, the spectral radius of $B_{coll}$ satisfies

$$\rho(B_{coll}) \leq \frac{\alpha_{coll}}{1-\alpha_{coll}} \tag{5.13}$$

where

$$\alpha_{coll} = \sum_{\substack{\ell,k=1\\k>\ell}}^{K} s_\ell s_k \lambda_2(P^{(k,\ell)})\,, \tag{5.14}$$

and $s_k = (p_k\tau_k + n\tau)^{-1}n\tau$ for $k = 1,\ldots,K$.

*Proof.* □

To begin with we work with a general $B$ from (5.7). Later on, we would use some specific properties of collapsing transformation, $B_{coll}$. We introduce a block diagonal matrix $D$ with the $i$th block being $(Q^{(i,i)})^{-\frac{1}{2}}$ and consider a transformation of $B_{coll}$, namely, $\widetilde{B} = DBD^{-1}$.

$$\begin{aligned}\widetilde{B} &= D(\boldsymbol{I}-L)^{-1}UD^{-1}\\ &= D(\boldsymbol{I}-L)^{-1}D^{-1}DUD^{-1}\\ &= (D(\boldsymbol{I}-L)D^{-1})^{-1}DUD^{-1}\\ &= (\boldsymbol{I}-DLD^{-1})^{-1}DUD^{-1}\\ &= (\boldsymbol{I}-\widetilde{L})^{-1}\widetilde{U}\end{aligned} \tag{5.15}$$

where $\widetilde{L} = DLD^{-1}$ and $\widetilde{U} = DUD^{-1}$. Defining $\widetilde{A} = DAD^{-1}$, we obtain $\widetilde{A}^{(i,i)} = 0\ \forall\ i$ and $\widetilde{A}^{(i,j)} = (Q^{(i,i)})^{-\frac{1}{2}}A^{(i,j)}(Q^{(j,j)})^{-\frac{1}{2}} = -(Q^{(i,i)})^{-\frac{1}{2}}Q^{(i,j)}(Q^{(j,j)})^{-\frac{1}{2}}\ \forall\ i \neq j$. Note that $\widetilde{A}$ is symmetric.

$$\rho(B) = \rho(\widetilde{B}) \leq \|\widetilde{B}\|_2 \leq \|(\boldsymbol{I}-\widetilde{L})^{-1}\|_2\|\widetilde{U}\|_2 \tag{5.16}$$

Observe that $\widetilde{U}$ and $\widetilde{L}$ are upper and lower (block) triangular part of the symmetric matrix $\tilde{A}$ respectively. Hence $\rho(\widetilde{L}) = \rho(\widetilde{U})$. Using results from matrix norm of block matrix, we obtain the following :

$$\|\widetilde{U}\|_2 \leq \sqrt{\sum_{ij}\|\widetilde{U}^{(i,j)}\|_2^2} = \sqrt{\sum_{j>i}\|\widetilde{A}^{(i,j)}\|_2^2} = \sqrt{\sum_{j>i}\|\widetilde{A}^{(i,j)}\widetilde{A}^{(j,i)}\|_2} := \alpha \tag{5.17}$$

In terms of precision matrix $Q$, $\alpha$ can be expressed as the following :

$$\begin{aligned}\alpha &= \sqrt{\sum_{j>i}\|(Q^{(i,i)})^{-\frac{1}{2}}Q^{(i,j)}(Q^{(j,j)})^{-1}Q^{(j,i)}(Q^{(i,i)})^{-\frac{1}{2}}\|_2}\\ &= \sqrt{\sum_{j>i}\|(Q^{(i,i)})^{-1}Q^{(i,j)}(Q^{(j,j)})^{-1}Q^{(j,i)}\|_2}\end{aligned}$$

Let us consider a generic vector $v$.

$$\begin{aligned}
&(I-\widetilde{L})^{-1}v = w\\
\Leftrightarrow &(I-\widetilde{L})w = v\\
\Leftrightarrow &\widetilde{L}w = w - v\\
\Leftrightarrow &\|\widetilde{L}\|_2\|w\|_2 \geq \|w\|_2 - \|v\|_2\\
\Leftrightarrow &\|v\|_2 \geq \|w\|_2(1-\|\widetilde{L}\|_2) \Leftrightarrow \quad \|v\|_2(1-\alpha) \geq \|w\|_2
\end{aligned}$$

Hence $\|(I-\widetilde{L})^{-1}\|_2 \leq \frac{1}{1-\alpha}$. It follows from (5.16) that $\rho(B) \leq \frac{\alpha}{1-\alpha}$.

Under balanced levels, $Q$ has following form from (5.9) :
$Q^{(k,k)} = (\tau_k + n^{(k)}\tau)\boldsymbol{I}_{p_k}, Q^{(k,\ell)} = \tau\mathcal{Z}^{(k,\ell)}$ where $\mathcal{Z}^{(k,\ell)}$ is the $p_k \times p_\ell$ dimensional co-occurrence matrix.
Using the definition in (5.7) and $Q$ above, for $0 < i < j \leq K$,

$$A^{(i,j)} = -(Q^{(i,i)})^{-1}Q^{(i,j)} = -\frac{\tau}{\tau_i + n^{(i)}\tau}\mathcal{Z}^{(i,j)} = -\frac{n^{(i)}\tau}{\tau_i + n^{(i)}\tau}P^{(i\to j)} = -\frac{n^{(i)}}{\lambda_i + n^{(i)}}P^{(i\to j)}$$

$$\widetilde{A}^{(i,j)} = -(Q^{(i,i)})^{-\frac{1}{2}}Q^{(i,j)}(Q^{(j,j)})^{-\frac{1}{2}} = -\frac{\tau}{\sqrt{(\tau_i + n^{(i)}\tau)}\sqrt{(\tau_j + n^{(j)}\tau)}}\mathcal{Z}^{(i,j)}$$

When $a^{(0)}$ is collapsed then we consider $\widetilde{A}^{(i,j)}$ for $0 < i < j \leq K$ after appropriate transformation. The corresponding $\widetilde{A}^{(i,j)}_{coll}$ can be written in terms of $R$ as the following : $-(R^{(i,i)})^{-\frac{1}{2}}R^{(i,j)}(R^{(j,j)})^{-\frac{1}{2}}$. Note that $R^{(i,i)}$ is of the form $a_i\boldsymbol{I}_{p_i} + b_i\boldsymbol{J}_{p_i}$ for $a_i = (\tau_i + \tau n/p_i)$ and some $b_i$. The exact form of $b_i$ is not relevant for proving the theorem. which would become evident from the set of equations below. Here $\boldsymbol{J}_{p_i}$ is a $p_i \times p_i$ matrix of 1's. Using $R^{(i,j)} = C_{p_i}Q^{(i,j)} = Q^{(i,j)}C_{p_j}$, we obtain the following for $1 \leq i < j \leq K$, :

$$
\begin{aligned}
\widetilde{A}^{(i,j)}_{coll} &= -(R^{(i,i)})^{-\frac{1}{2}} R^{(i,j)} (R^{(j,j)})^{-\frac{1}{2}} \\
&= -(\sqrt{(\tau_i + \tau n/p_i)^{-1}} \boldsymbol{I}_{p_i} + b_i \boldsymbol{J}_{p_i}) C_{p_i} Q^{(i,j)} (\sqrt{(\tau_j + \tau n/p_j)^{-1}} \boldsymbol{I}_{p_j} + b_j \boldsymbol{J}_{p_j}) \\
&= -\sqrt{(\tau_i + \tau n/p_i)^{-1}} C_{p_i} Q^{(i,j)} (\sqrt{(\tau_i + \tau n/p_i)^{-1}} \boldsymbol{I}_{p_j} + b_j \boldsymbol{J}_{p_j}) \quad \text{This is because } \mathbf{1}^\mathsf{T} C_{p_i} = \mathbf{0}^\mathsf{T} \\
&= -\frac{1}{\sqrt{(\tau_i + \tau n/p_i)}} Q^{(i,j)} C_{p_j} \frac{1}{\sqrt{(\tau_j + \tau n/p_j)}} \\
&= -(Q^{(i,i)})^{-\frac{1}{2}} Q^{(i,j)} C_{p_j} (Q^{(j,j)})^{-\frac{1}{2}} \\
&= -(Q^{(i,i)})^{-\frac{1}{2}} Q^{(i,j)} (Q^{(j,j)})^{-\frac{1}{2}} C_{p_j} \\
&= \widetilde{A}^{(i,j)} (\boldsymbol{I}_{p_j} - \frac{1}{p_j} \boldsymbol{J}_{p_j}) \\
&= -\frac{\tau}{\sqrt{(\tau_i + n^{(i)}\tau)} \sqrt{(\tau_j + n^{(j)}\tau)}} \mathcal{Z}^{(i,j)} (\boldsymbol{I}_{p_j} - \frac{1}{p_j} \boldsymbol{J}_{p_j}).
\end{aligned}
$$

$$
\begin{aligned}
\|\widetilde{A}^{(i,j)}_{coll}\|_2^2 &= \rho(\frac{n^{(i)}\tau}{(\tau_i + n^{(i)}\tau)} \frac{n^{(j)}\tau}{(\tau_j + n^{(j)}\tau)} P^{(i\to j)} (\boldsymbol{I}_{p_j} - \frac{1}{p_j} \boldsymbol{J}_{p_j}) (\boldsymbol{I}_{p_j} - \frac{1}{p_j} \boldsymbol{J}_{p_j}) P^{(j\to i)}) \\
&= \frac{n^{(i)}\tau}{(\tau_i + n^{(i)}\tau)} \frac{n^{(j)}\tau}{(\tau_j + n^{(j)}\tau)} \rho((\boldsymbol{I}_{p_j} - \frac{1}{p_j} \boldsymbol{J}_{p_j}) P^{(j\to i)} P^{(i\to j)}) \\
&= \frac{n\tau}{(p_i\tau_i + n\tau)} \frac{n\tau}{(p_j\tau_j + n\tau)} \lambda_2(P^{(j\to i)} P^{(i\to j)}) \\
&= s_i s_j \lambda_2(P^{(j,i)})
\end{aligned} \tag{5.18}
$$

For the plain vanilla version $\|\widetilde{A}^{(i,j)}\|_2^2 = s_i s_j$ for $1 \le i < j \le K$ and $\|\widetilde{A}^{(0,j)}\|_2^2 = s_j$ for $1 \le j \le K$. Note that $s_i$'s converge to 1 in the asymptotic regime we are interested in.

Combining (5.17) and (5.18), for collapsing we obtain

$$
\alpha_{coll} = \sum_{\substack{i,j=1 \\ j>i}}^{K} s_i s_j \lambda_2(P^{(j,i)}) \le \sum_{\substack{i,j=1 \\ j>i}}^{K} \lambda_2(P^{(j,i)}) := \eta_{coll}
$$

Using the result from the first part of the proof we obtain that

$$\rho(B_{coll}) \leq \frac{\alpha_{coll}}{1-\alpha_{coll}} \leq \frac{\eta_{coll}}{1-\eta_{coll}} := \beta_{coll} \tag{5.19}$$

For a fixed $K$, in any regime where $\mathbb{P}(|\lambda_2(P^{(j,i)})| < \epsilon) > 1-\delta \;\forall\; 1 \leq i < j \leq K$, we have $\mathbb{P}\Big(|\rho(B_{coll})| < \frac{K(K-1)}{2}\epsilon\Big) > 1 - \frac{K(K-1)}{2}\delta$. We thus need to study $\lambda_2(P^{(j,i)})$, where $P^{(j,i)}$ is a random walk on a random bipartite bi-regular graphs. The second largest eigenvalue of the adjacency matrix of random (regular) graphs is well studied in the random graph literature. In particular, the Alon's conjecture (aka Friedman's second eigenvalue theorem) provides sharp characterizations which show that the corresponding random walks are have near-optimal mixing properties. Such analyses have recently been extended to random bipartite bi-regular graphs (see Brito et al. (2022)) and roughly imply that $\lambda_2(P^{(j,i)}) \approx 2(\min\{d_i, d_j\})^{-1}$ when $n \to \infty$ while $d_i := n_1^{(i)} = n/p_i$ and $d_j := n_1^{(j)} = n/p_j$ remain fixed. Such results have been extended to cases where the degrees $d_i$ and $d_j$ are allowed to grow with $n$, which is our context of interest, but this required additional and potentially restrictive assumptions on the rate at which this happens (see Zhu (2020)). For instance, we may provide a restrictive regularity condition under which $\lambda_2(P^{(j,i)}) \to 0$ in probability as $d_i, d_j, p_i, p_j \to \infty \;\forall\; 1 \leq i \neq j \leq K$. If $p_i \approx N^{\rho}$, $p_j \approx N^{\kappa}$ and $(\rho, \kappa)$ is in $\mathcal{D}_1 \cup \mathcal{D}_2$ for $\mathcal{D}_1 = \{0 < \rho \leq \kappa < 1, 2\rho + 3\kappa > 3\}$, $\mathcal{D}_2 = \{0 < \kappa \leq \rho < 1, 3\rho + 2\kappa > 3\}$, then $\lambda_2(P^{(j,i)}) = o(1)$ with high probability. Figure 5.1 depicts this domain $\mathcal{D}$.

We find empirically that $\rho(B_{coll})$ is bounded away from 1 in a broader regime. The regime for $(\rho, \kappa)$ described by $2\rho + 3\kappa > 3$ and $3\rho + 2\kappa > 3$ is a sparse regime whereas the regime described by $2\rho + 3\kappa < 3$, $3\rho + 2\kappa < 3$ and $\rho + \kappa > 1$ (the white triangle in figure 5.1) is a comparatively dense regime. Increasing the amount of data per level corresponds to adding more edges to the conditional independence graph, resulting in a larger connectivity in the state space for the auxiliary process facilitating better mixing. Thus we believe that $\lambda_2(P^{(j,i)}) \to 0$ even for $(\rho, \kappa)$ in the white triangle, thus having a good mixing for almost any asymptotic regime. To give empirical evidence, we calculate $\rho(B_{coll})$ and the upper bound $\beta_{coll}$ in (5.19), when the levels of the three factors are generated at random. We also enforce $n_{i_1 i_2 i_3}$ is either 0 or 1. We consider

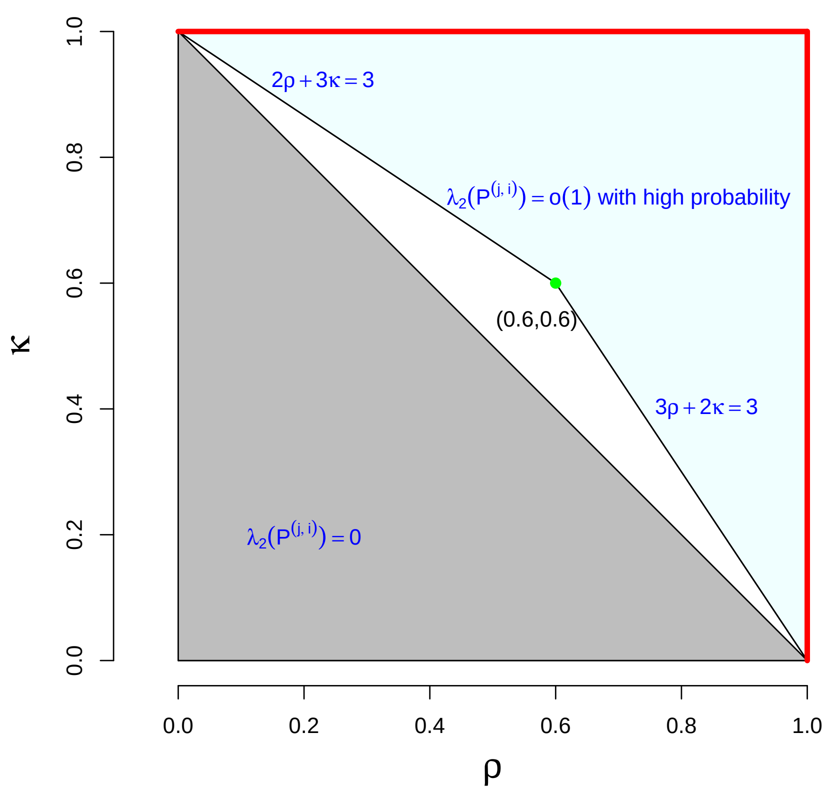



Figure 5.1: The light blue shaded region is the domain $\mathcal{D}$ where the spectral gap of $P^{(j,i)}$ is close to 1.

simulated data with $K = 3$, $p_1 = N^{\rho}$, $p_2 = N^{\kappa}$, $p_3 = N^{\chi}$ for $\rho = \kappa = \chi = 0.56$ and $\rho = \kappa = \chi = 0.64$. 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 $\rho(B_{coll})$ and the upper bound in (5.19).

Figure 5.2 displays the behavior of the $\rho(B_{coll})$ and the upper bound in an asymptotic regime, where both the number of data points and factor levels increase. In our spectral radius and upper bound computations, we took precision terms $\tau = \tau_1 = \tau_2 = \tau_3 = 1$. We take $\log_{10}(N)$ in the range from 4 to 5. We observe that $\rho(B_{coll})$ is well below 1 and decreasing with problem size showing scalability of a

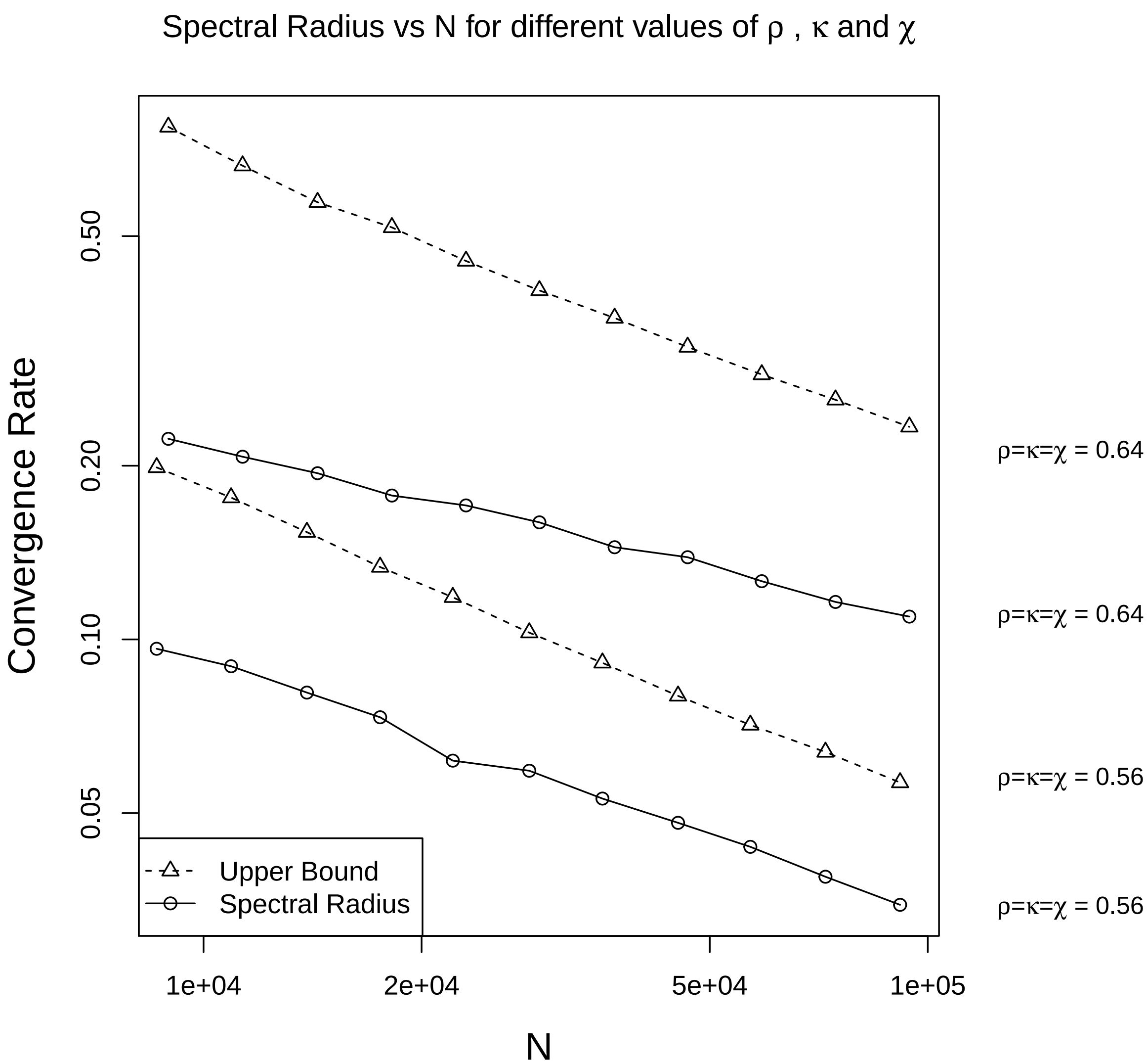



Figure 5.2: Spectral radius of collapsed Gibbs sampler $\rho(B_{coll})$ and upper bound $\beta_{coll}$ in (5.19) versus problem size $N$ for $K = 3$, $(\rho, \kappa, \chi) = (0.56, 0.56, 0.56)$ and $(\rho, \kappa, \chi) = (0.64, 0.64, 0.64)$.

collapsed Gibbs sampler. This suggests that the theory developed previously is relevant beyond cases in which the balancing levels are strictly satisfied. We also observe that the spectral radius for $(\rho, \kappa, \chi) = (0.56, 0.56, 0.56)$ is dominated by that of $(\rho, \kappa, \chi) = (0.64, 0.64, 0.64)$ as expected.

## 5.4 Discussion

In this chapter we have presented an insightful connection between the convergence of three coordinate wise inference algorithms and the relaxation time of an auxiliary Markov chain on a discrete state-space. We are also able to establish scalability for $K > 2$ under the balanced design assumption. However, the conditions in Zhu (2020) may not be necessary. The heuristic argument presented before suggests that such results hold true for almost any asymptotic regime. We show empirical evidence of scalability for designs without balanced level assumption for $K = 3$ and $(\rho, \kappa, \chi)$ outside the regime in Zhu (2020). We conjecture if the co-occurrence matrix $n^{(i,j)} = (n^{(i,j)}_{k,\ell})$ is uniformly distributed among the set of $p_i \times p_j$ binary matrices with balanced levels and $n = p_i d_i = p_j d_j$ non-zero entries, then $\lambda_2(P^{(j,i)}) \to 0$ in probability as $d_i, d_j, p_i, p_j \to \infty$. One immediate direction this work can move forward is to provide rigorous arguments for the same. Greenhill et al. (2004) considers decompositions of random bipartite regular graphs for a fixed degree. An “averaging argument” adapted for increasing degree under mild conditions on the growth rates could be useful.

# Appendix A

# Linear Mixed Model with Crossed Random Effect

Table A.1 shows results of OLS and GLS regression for the Stitch Fix data in Section 2.6. OLS is estimated to be naive when $\widehat{\mathrm{SE}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}}) < \widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})$ and inefficient when $\widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}}) > \widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS}})$. Estimates that are more than double their corresponding standard error get an asterisk.

Table A.1: Stitch Fix Regression Results

| | $\hat{\beta}_{\text{OLS}}$ | $\widehat{\text{SE}}_{\text{OLS}}(\hat{\beta}_{\text{OLS}})$ | $\widehat{\text{SE}}_{\text{GLS}}(\hat{\beta}_{\text{OLS}})$ | $\hat{\beta}_{\text{GLS}}$ | $\widehat{\text{SE}}_{\text{GLS}}(\hat{\beta}_{\text{GLS}})$ |
|---|---|---|---|---|---|
| Intercept | 4.635* | 0.005397 | 0.05148 | 5.103* | 0.01092 |
| Match | 5.048* | 0.01174 | 0.1297 | 3.442* | 0.01823 |
| $\mathbb{I}\{$ client edgy $\}$ | 0.001020 | 0.002443 | 0.004444 | 0.003041 | 0.003550 |
| $\mathbb{I}\{$ item edgy $\}$ | −0.3358* | 0.004253 | 0.03307 | −0.3515* | 0.01375 |
| $\mathbb{I}\{$ client edgy $\}$ $*\mathbb{I}\{$ item edgy $\}$ | 0.3925* | 0.006229 | 0.01233 | 0.3793* | 0.005916 |
| $\mathbb{I}\{$ client boho $\}$ | 0.1386* | 0.002264 | 0.004211 | 0.1296* | 0.003356 |
| $\mathbb{I}\{$ item boho $\}$ | −0.5499* | 0.005981 | 0.02713 | −0.6266* | 0.01485 |
| $\mathbb{I}\{$ client boho $\}$ $*\mathbb{I}\{$ item boho $\}$ | 0.3822* | 0.007566 | 0.01001 | 0.3763* | 0.007123 |
| Acrylic | −0.06482* | 0.003778 | 0.03371 | −0.005360 | 0.01909 |
| Angora | −0.01262 | 0.007848 | 0.08530 | 0.07486 | 0.05177 |
| Bamboo | −0.04593 | 0.06215 | 0.2096 | 0.03251 | 0.1535 |
| Cashmere | −0.1955* | 0.02484 | 0.1414 | 0.008930 | 0.1048 |
| Cotton | 0.1752* | 0.003172 | 0.04220 | 0.1033* | 0.01612 |
| Cupro | 0.5979* | 0.3016 | 0.4519 | 0.2089 | 0.4363 |
| Faux Fur | 0.2759* | 0.02008 | 0.07694 | 0.2749* | 0.06691 |
| Fur | −0.2021* | 0.03121 | 0.1388 | −0.07924 | 0.1182 |
| Leather | 0.2677* | 0.02482 | 0.07759 | 0.1674* | 0.06545 |
| Linen | −0.3844* | 0.05632 | 0.2429 | −0.08658 | 0.1499 |
| Modal | 0.002587 | 0.009775 | 0.1816 | 0.1388* | 0.05804 |
| Nylon | 0.03349* | 0.01552 | 0.08878 | 0.08174 | 0.05751 |
| Patent Leather | −0.2359 | 0.1800 | 0.3838 | −0.3764 | 0.3771 |
| Pleather | 0.4163* | 0.008916 | 0.08774 | 0.3292* | 0.04468 |
| PU | 0.4160* | 0.008225 | 0.07989 | 0.4579* | 0.03737 |
| PVC | 0.6574* | 0.06545 | 0.3462 | 0.9688* | 0.3441 |
| Rayon | −0.01109* | 0.002951 | 0.04074 | 0.05155* | 0.01329 |
| Silk | −0.1422* | 0.01317 | 0.08907 | −0.1828* | 0.04871 |
| Spandex | −0.3916* | 0.00931 | 0.1373 | 0.4140* | 0.1141 |
| Tencel | 0.4966* | 0.01729 | 0.1712 | 0.1234* | 0.05982 |
| Viscose | 0.04066* | 0.006953 | 0.08519 | −0.02259 | 0.03145 |
| Wool | −0.06021* | 0.006611 | 0.07211 | −0.05883 | 0.03319 |

# Appendix B

# Generalized Linear Mixed Model with Crossed Random Effect

## B.1 Some proofs

### B.1.1 Proof of Lemma 3.3.1

The first derivative of $T^{-1}$ with respect to $\eta$ is

$$\frac{\mathrm{d}}{\mathrm{d}\eta}T(\eta)^{-1} = -T(\eta)^{-1}T'(\eta)T(\eta)^{-1} \quad \text{for} \quad T'(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}T(\eta) = \begin{pmatrix} 0 & B \\ B^{\mathsf{T}} & 0 \end{pmatrix}.$$

Higher order derivatives of $T(\eta)^{-1}$ simplify greatly because $T''(\eta) = 0$. For integers $k \geqslant 1$ we find by induction that

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}\eta^k}T(\eta)^{-1} &= (-1)^k k!\big(T(\eta)^{-1}T'(\eta)\big)^k T(\eta)^{-1} \\ &= (-1)^k k! T(\eta)^{-1/2}\Big(T(\eta)^{-1/2}T'(\eta)T(\eta)^{-1/2}\Big)^k T(\eta)^{-1/2}.\end{aligned}$$

Next, for $B_* = A^{-1/2}BC^{-1/2}$, we find that

$$T(0)^{-1/2}T'(0)T(0)^{-1/2} = \begin{pmatrix} 0 & -B_* \\ -B_*^{\mathsf{T}} & 0 \end{pmatrix}$$

and then taking a Taylor series around $\eta = 0$ we have

$$T(\eta)^{-1} = T(0)^{-1} + T(0)^{-1/2} \sum_{k \geqslant 1} \begin{pmatrix} 0 & -\eta B_* \\ -\eta B_*^\mathsf{T} & 0 \end{pmatrix}^k T(0)^{-1/2}. \tag{B.1}$$

This Taylor series converges for $|\eta| < 1/\rho$.

Therefore, for $0 \leqslant \eta < 1/\rho$,

$$\begin{aligned} \operatorname{tr}(T(\eta)^{-1}) &= \operatorname{tr}(T(0)^{-1}) + \operatorname{tr}\bigl(T(0)^{-1/2} E(\eta) T(0)^{-1/2}\bigr) \\ &= \operatorname{tr}\bigl(T(0)^{-1}(I + E(\eta))\bigr), \quad \text{for} \\ E(\eta) &= \sum_{k \geqslant 1} \begin{pmatrix} 0 & -\eta B_* \\ -\eta B_*^\mathsf{T} & 0 \end{pmatrix}^k. \end{aligned}$$

Recalling that $T(0)$ is diagonal, the $k = 1$ term does not contribute to the trace of the inverse. That is also true for any every odd integer $k \geqslant 1$. It follows that

$$\begin{aligned} \operatorname{tr}(T(\eta)^{-1}) &= \operatorname{tr}\bigl(T(0)^{-1}(I + E_2(\eta))\bigr), \quad \text{for} \\ E_2(\eta) &= \sum_{k \geqslant 1} \begin{pmatrix} 0 & -\eta B_* \\ -\eta B_*^\mathsf{T} & 0 \end{pmatrix}^{2k} = \sum_{k \geqslant 1} \begin{pmatrix} \eta^2 B_* B_*^\mathsf{T} & 0 \\ 0 & \eta^2 B_*^\mathsf{T} B_* \end{pmatrix}^k. \end{aligned}$$

Now we can write the difference $\operatorname{tr}(T(\eta)^{-1}) - \operatorname{tr}(T(0)^{-1})$ as

$$\sum_{k \geqslant 1} \eta^{2k} \Bigl( \operatorname{tr}\bigl(A^{-1}(B_* B_*^\mathsf{T})^k\bigr) + \operatorname{tr}\bigl(C^{-1}(B_*^\mathsf{T} B_*)^k\bigr) \Bigr).$$

### B.1.2 Proof of Lemma 3.3.2

First $\rho^2$ is no larger than the spectral radius of

$$\begin{pmatrix} 0 & B_* \\ B_* & 0 \end{pmatrix}^2 = \begin{pmatrix} B_* B_*^\mathsf{T} & 0 \\ 0 & B_*^\mathsf{T} B_* \end{pmatrix}.$$

The eigenvalues of this matrix are the same as those of $B_*^\mathsf{T} B_*$ and of

$$B_* B_*^\mathsf{T} = A^{-1/2} B C^{-1} B^\mathsf{T} A^{-1/2}$$

which has by similarity, the same eigenvalues as $BC^{-1}B^\mathsf{T}A^{-1}$.

Now $BC^{-1}$ is a nonnegative matrix and so its largest eigenvalue is a real number no larger than its largest row sum, which in this case is $\max_i \sum_j B_{ij}/C_{jj}$. Applying the same to $B^\mathsf{T}A^{-1}$ gives us a bound on $\rho^2$ that is the square of the claimed bound for $\rho$.

### B.1.3 Proof of Theorem 3.3.1

The $is$ element of $B_* B_*^\mathsf{T}$ is

$$(B_* B_*^\mathsf{T})_{is} = \sum_j \frac{Z_{ij} W_{ij}}{\sqrt{W_{i\bullet} + \sigma_A^{-2}}\sqrt{W_{\bullet j} + \sigma_B^{-2}}} \frac{Z_{sj} W_{sj}}{\sqrt{W_{s\bullet} + \sigma_A^{-2}}\sqrt{W_{\bullet j} + \sigma_B^{-2}}}.$$

Therefore the $i$'th diagonal element of $B_* B_*^\mathsf{T}$ is

$$(B_* B_*^\mathsf{T})_{ii} = \sum_j \frac{Z_{ij} W_{ij}^2}{(W_{i\bullet} + \sigma_A^{-2})(W_{\bullet j} + \sigma_B^{-2})} \leqslant \frac{1}{W_{i\bullet}} \sum_j \frac{Z_{ij} W_{ij}^2}{W_{\bullet j}}.$$

With probability tending to 1

$$(B_* B_*^\mathsf{T})_{ii} \leqslant S^{2\psi} \bar{\omega}^2 \frac{1}{N_{i\bullet}} \sum_j \frac{Z_{ij}}{N_{\bullet j}} \leqslant S^{2\psi} \bar{\omega}^2 (1-\epsilon)^{-2} S^{\rho+\kappa-2} \sum_j Z_{ij}$$

by equations (3.7) and (3.8). Now $\sum_j Z_{ij} = N_{i\bullet} \leqslant (\Upsilon + \epsilon) S^{1-\rho}$ by equation (3.7). Therefore with probability approaching one,

$$\mathrm{tr}(B_* B_*^\mathsf{T}) \leqslant \bar{\omega}^2 (1-\epsilon)^{-2} (\Upsilon + \epsilon) S^{2\psi+\rho+\kappa-1} = o(R)$$

because $R = S^\rho$ and $\kappa < 1$ and we can choose any $\psi > 0$. Now none of the eigenvalues of $B_* B_*^\mathsf{T}$ can be negative. If there are $R^\alpha = S^{\alpha\rho}$ eigenvalues larger than $\delta$ then

$$\mathrm{tr}(B_* B_*^\mathsf{T}) > \delta S^{\alpha\rho}.$$

That cannot hold for large $S$ if

$$\alpha > \underline{\alpha} \equiv 1 + \frac{2\psi + \kappa - 1}{\rho} < 1$$

for small enough $\psi$. Therefore any $\alpha \in (\underline{\alpha}, 1)$ satisfies the condition in (3.32).

### B.1.4 Proof of Theorem 3.3.2

Let $B_* B_*^\mathsf{T}$ have eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_R \geqslant 0$ with corresponding unit norm eigenvectors $v_i$. Then

$$\mathrm{Err} = \mathrm{tr}\Big(A^{-1} \sum_{k \geqslant 1} \sum_{i=1}^{R} \lambda_i^k v_i v_i^\mathsf{T}\Big) = \sum_{k \geqslant 1} \sum_{i=1}^{R} \lambda_i^k \mathrm{tr}(v_i^\mathsf{T} A^{-1} v_i).$$

Now with probability tending to 1 as $S \to \infty$,

$$\mathrm{tr}(v_i^\mathsf{T} A^{-1} v_i) \leqslant \lambda_{\max}(A^{-1}) = \max_s (W_{s\bullet} + \sigma_A^{-2})^{-1} \leqslant S^{\psi}(1-\epsilon)^{-1} S^{\rho-1}$$

because $W_{s\bullet} \geqslant S^{-\psi} \min_i N_{i\bullet} \geqslant S^{-\psi}(1-\epsilon) S^{1-\rho}$ with overwhelming probability. Therefore with probability tending to 1 as $S \to \infty$

$$\mathrm{Err} \leqslant S^{\psi+\rho-1}(1-\epsilon)^{-1} \sum_{i=1}^{R} \frac{\lambda_i}{1-\lambda_i}.$$

Now choose the $\alpha < 1$ provided by Theorem 3.3.1 and suppose that for $\delta > 0$ that fewer than $R^\alpha$ of the $\lambda_i$ are larger than $\delta$. Under this event which has probability tending to one,

$$\mathrm{Err} \leqslant S^{\psi+\rho-1}(1-\epsilon)^{-1} \Big( \frac{\delta \lambda_1 R^\alpha}{1-\lambda_1} + \frac{R}{1-\delta} \Big).$$

The second term above is $O(S^{\psi+\rho-1}R) = o(R)$ because $\rho < 1$ and we can choose any $\psi > 0$.

To control the first term, we use Proposition 3.3.1 to bound $1/(1-\lambda_1)$. The bound there takes the form $\lambda_1 \leqslant (1+a)^{-1}(1+b)^{-1}$ for $a = S^{\rho-1}/[(\Upsilon+\epsilon)\bar\omega\sigma_A^2]$ and $b = S^{\kappa-1}/[(\Upsilon+\epsilon)\bar\omega\sigma_B^2]$. Therefore

$$\begin{aligned}\frac{1}{(1/\lambda_1 - 1)} \leqslant \frac{1}{a+b+ab} &\leqslant \min\Bigl(\frac{1}{a}, \frac{1}{b}\Bigr)\\ &= \min(S^{1-\rho}\sigma_A^2, S^{1-\kappa}\sigma_B^2)\bar\omega(\Upsilon+\epsilon).\end{aligned}$$

It now follows that the first term in Err is $O(S^{\psi}R^{\alpha}) = o(R)$ because $\alpha < 1$ and we can choose any $\psi > 0$.

### B.1.5 Proof of Lemma 3.4.2

Our iterative algorithm is designed to minimize

$$\mathrm{pl}(\beta, \boldsymbol{a}, \boldsymbol{b}) = \sum_{ij} Z_{ij}\widehat{W}_{ij}(z_{ij} - x_{ij}^{\mathsf{T}}\beta - a_i - b_j)^2 + \hat\sigma_A^{-2}\|\boldsymbol{a}\|^2 + \hat\sigma_B^{-2}\|\boldsymbol{b}\|^2,$$

where all $\widehat{W}_{ij} > 0$, $\min(\hat\sigma_A^{-2}, \hat\sigma_B^{-2}) > 0$, the matrix $\mathcal{X}$ with rows $x_{ij}^{\mathsf{T}}$ has full rank, and $z_{ij}$ are fixed numbers. This quadratic function has a positive definite Hessian and a unique global minimum $(\beta_*, \boldsymbol{a}_*, \boldsymbol{b}_*)$.

We introduce $\mathrm{pl}(\beta, \boldsymbol{a}; \boldsymbol{b})$ which for any $\boldsymbol{b} \in \mathbb{R}^C$ gives us a function of $\beta \in \mathbb{R}^p$ and $\boldsymbol{a} \in \mathbb{R}^C$. Similarly, we introduce $\mathrm{pl}(\beta, \boldsymbol{b}; \boldsymbol{a})$ where this time $\boldsymbol{a}$ is the parameter while $\beta$ and $\boldsymbol{b}$ are the function's arguments. Given $\beta^{(k)}$, $\boldsymbol{a}^{(k)}$ and $\boldsymbol{b}^{(k)}$ we minimize $\mathrm{pl}(\beta, \boldsymbol{a}; \boldsymbol{b}^{(k)})$ over $\beta$ and $\boldsymbol{a}$ to get $\beta^{(k+\frac{1}{2})}$ and $\boldsymbol{a}^{(k+1)}$. Then we minimize $\mathrm{pl}(\beta, \boldsymbol{b}; \boldsymbol{a}^{(k+1)})$ over $\beta$ and $\boldsymbol{b}$ to get $\beta^{(k+1)}$ and $\boldsymbol{b}^{(k+1)}$. Both minimizations choose arguments that make their respective gradients equal to zero. Therefore

$$\begin{aligned}&\mathrm{pl}(\beta^{(k)}, \boldsymbol{a}^{(k)}; \boldsymbol{b}^{(k)}) - \mathrm{pl}(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)}; \boldsymbol{b}^{(k)})\\ &= \frac{1}{2}\begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})}\\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix}^{\mathsf{T}} H(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)}; \boldsymbol{b}^{(k)}) \begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})}\\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix}\end{aligned}$$

using $H$ to denote the Hessian of the parameterized function. Considering both steps at once we get

$$\begin{aligned}\Delta^{(k+1)} &\equiv \mathrm{pl}(\beta^{(k)}, \boldsymbol{a}^{(k)}, \boldsymbol{b}^{(k)}) - \mathrm{pl}(\beta^{(k+1)}, \boldsymbol{a}^{(k+1)}, \boldsymbol{b}^{(k+1)}) \\ &= \frac{1}{2}\begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})} \\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix}^{\mathsf{T}} H(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)}; \boldsymbol{b}^{(k)}) \begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})} \\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix} \\ &+ \frac{1}{2}\begin{pmatrix}\beta^{(k+\frac{1}{2})} - \beta^{(k+1)} \\ \boldsymbol{b}^{(k)} - \boldsymbol{b}^{(k+1)}\end{pmatrix}^{\mathsf{T}} H(\beta^{(k+1)}, \boldsymbol{b}^{(k+1)}; \boldsymbol{a}^{(k+1)}) \begin{pmatrix}\beta^{(k+\frac{1}{2})} - \beta^{(k+1)} \\ \boldsymbol{b}^{(k)} - \boldsymbol{b}^{(k+1)}\end{pmatrix}.\end{aligned}$$

Now $\Delta^{(k+1)} \geqslant 0$ and

$$\sum_{k=0}^{\infty} \Delta^{(k+1)} \leqslant \mathrm{pl}(\beta^{(0)}, \boldsymbol{a}^{(0)}, \boldsymbol{b}^{(0)}) - \mathrm{pl}(\beta_*, \boldsymbol{a}_*, \boldsymbol{b}_*) < \infty$$

and so this sum converges. Noting that $H(\beta, \boldsymbol{a}; \boldsymbol{b})$ and $H(\beta, \boldsymbol{b}; \boldsymbol{a})$ are independent of $(\beta, \boldsymbol{a}, \boldsymbol{b})$ and strictly positive definite, we conclude that $(\beta^{(k)}, \boldsymbol{a}^{(k)}, \boldsymbol{b}^{(k)})$ converges. At the limit point, the gradients of both $\mathrm{pl}(\beta, \boldsymbol{a}; \boldsymbol{b})$ and $\mathrm{pl}(\beta, \boldsymbol{b}; \boldsymbol{a})$ must vanish and so therefore the gradient of $\mathrm{pl}(\beta, \boldsymbol{a}, \boldsymbol{b})$ also vanishes there.

### B.1.6 Results of the binary regression in Section 3.6

Table B.1 shows coefficient estimates and standard errors for plain logistic regression and a generalized linear mixed model logistic regression for the Stitch Fix data in Section 3.6. Logistic is estimated to be naive when $\widehat{\mathrm{SE}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}}) < \widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})$ and inefficient when $\widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}}) > \widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GLMM}})$. Estimates that are more than double their corresponding standard error get an asterisk.

For the Stitch Fix data we obtained $\hat{\sigma}_A^2 = 0.68$ (customers), $\hat{\sigma}_B^2 = 0.21$ (items).

Table B.1: Stitch Fix Binary Regression Results

| | $\hat{\beta}_{\mathrm{LR}}$ | $\widehat{\mathrm{SE}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}})$ | $\widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})$ | $\hat{\beta}_{\mathrm{GLMM}}$ | $\widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GLMM}})$ |
|---|---|---|---|---|---|
| Intercept | $0.677^*$ | 0.022 | 0.048 | $0.584^*$ | 0.036 |
| $\mathbb{I}\{$client fit profile missing$\}$ | $0.099^*$ | 0.015 | 0.025 | $0.127^*$ | 0.023 |
| $\mathbb{I}\{$client fit profile loose$\}$ | $0.0155^*$ | 0.002 | 0.006 | $0.208^*$ | 0.004 |
| $\mathbb{I}\{$client fit profile oversize$\}$ | $-0.0940^*$ | 0.010 | 0.019 | $-0.0477^*$ | 0.015 |
| $\mathbb{I}\{$client fit profile straight$\}$ | $0.125^*$ | 0.003 | 0.005 | $0.154^*$ | 0.004 |
| $\mathbb{I}\{$client fit profile tight$\}$ | $-0.229^*$ | 0.011 | 0.1297 | $-0.250^*$ | 0.009 |
| $\mathbb{I}\{$client edgy$\}$ | $-0.042^*$ | 0.002 | 0.005 | $-0.044^*$ | 0.003 |
| $\mathbb{I}\{$client boho$\}$ | $0.147^*$ | 0.002 | 0.006 | $0.194^*$ | 0.003 |
| $\mathbb{I}\{$client chest size$\}$ | $-0.008^*$ | 0.0006 | 0.001 | $-0.011^*$ | 0.001 |
| $\mathbb{I}\{$client dress size$\}$ | $-0.002^*$ | 0.0003 | 0.0007 | $-0.0002$ | 0.0005 |
| $\mathbb{I}\{$item fit profile missing$\}$ | $-0.281^*$ | 0.024 | 0.188 | $-0.079$ | 0.151 |
| $\mathbb{I}\{$item fit profile fitted$\}$ | $0.061^*$ | 0.004 | 0.089 | $0.277^*$ | 0.029 |
| $\mathbb{I}\{$item fit profile loose$\}$ | $0.209^*$ | 0.003 | 0.082 | 0.038 | 0.029 |
| $\mathbb{I}\{$item fit profile oversize$\}$ | $-0.066^*$ | 0.016 | 0.135 | $-0.020$ | 0.111 |
| $\mathbb{I}\{$item fit profile tight$\}$ | 0.186 | 0.0098 | 0.431 | 0.251 | 0.359 |
| $\mathbb{I}\{$material missing$\}$ | $-0.203^*$ | 0.022 | 0.190 | $-0.227$ | 0.135 |
| Acrylic | $-0.187^*$ | 0.005 | 0.060 | $-0.206^*$ | 0.031 |
| Angora | $-0.305^*$ | 0.024 | 0.222 | $-0.279$ | 0.151 |
| Cashmere | $0.367^*$ | 0.057 | 0.462 | $-0.744$ | 0.399 |
| Cotton | $-0.246^*$ | 0.004 | 0.060 | $-0.245^*$ | 0.028 |
| Faux Fur | 0.321 | 0.250 | 0.516 | 0.547 | 0.479 |
| Fur | $-0.459^*$ | 0.128 | 0.489 | $-0.619$ | 0.480 |
| Linen | $-0.494^*$ | 0.025 | 0.182 | $-0.374^*$ | 0.163 |
| Modal | $0.035^*$ | 0.007 | 0.262 | $-0.057$ | 0.080 |
| Nylon | $0.096^*$ | 0.013 | 0.188 | 0.070 | 0.108 |
| Patent Leather | $-0.789^*$ | 0.110 | 0.454 | $-0.531$ | 0.379 |
| Pleather | $-0.215^*$ | 0.020 | 0.194 | $-0.121$ | 0.130 |
| PU | $0.390^*$ | 0.042 | 0.0462 | 0.509 | 0.0351 |
| Rayon | $-0.033^*$ | 0.002 | 0.068 | $-0.013$ | 0.020 |
| Silk | $-0.041^*$ | 0.009 | 0.133 | 0.053 | 0.068 |
| Spandex | 0.025 | 0.050 | 0.355 | 0.210 | 0.335 |
| Tencel | $-0.107^*$ | 0.025 | 0.181 | $-0.041$ | 0.158 |
| Viscose | $-0.081^*$ | 0.008 | 0.086 | $-0.084$ | 0.054 |
| Wool | $-0.217^*$ | 0.024 | 0.235 | $-0.194$ | 0.150 |

# Appendix C

# Convergence of Collapsed Gibbs for Crossed Random Effect Models

## C.1 Some proofs

### C.1.1 Proof of Proposition 4.2.1

Observe that

$$
\begin{aligned}
& E\left(a_i^{(1)}(t+1) \mid a^{(0)}(t+1), \boldsymbol{a}^{(2)}(t)\right) \\
= \quad & s_i^{(1)}\left(\tilde{y}_i^{(1)} - a^{(0)}(t+1) - \frac{1}{N_{i\bullet}}\sum_j a_j^{(2)}(t)Z[i,j]\right).
\end{aligned} \tag{C.1}
$$

Marginalizing over $a^{(0)}(t+1)$, we obtain

$$
\begin{aligned}
& E\left(a_i^{(1)}(t+1) \mid \boldsymbol{a}^{(2)}(t)\right) \\
= \quad & s_i^{(1)}\left(\tilde{y}_i^{(1)} - \mathbb{E}(a^{(0)}(t+1) \mid \boldsymbol{a}^{(2)}(t)) - \frac{1}{N_{i\bullet}}\sum_j a_j^{(2)}(t)Z[i,j]\right).
\end{aligned} \tag{C.2}
$$

We also note that

$$
E\left(a^{(0)}(t+1) \mid \boldsymbol{a}^{(2)}(t)\right)
$$

$$= \frac{1}{\sum_{i'=1}^{R} s_{i'}^{(1)}} \sum_{i=1}^{R} s_i^{(1)} \left( \tilde{y}_i^{(1)} - \frac{1}{N_{i\bullet}} \sum_j a_j^{(2)}(t) Z[i,j] \right). \tag{C.3}$$

Substituting (C.3) in (C.2), we conclude

$$\begin{aligned} E\left(a_i^{(1)}(t+1) \mid \boldsymbol{a}^{(2)}(t)\right) &= \frac{s_i^{(1)}}{\sum_l s_l^{(1)}} \sum_{i'=1}^{R} \frac{s_{i'}^{(1)}}{N_{i'\bullet}} \sum_j Z[i',j] a_j^{(2)}(t) \\ &\quad - \sum_{j=1}^{C} \frac{s_i^{(1)} Z[i,j]}{N_{i\bullet}} a_j^{(2)}(t) + C, \end{aligned} \tag{C.4}$$

where $C$ is a constant. From the definition of $B_1$, we know that $B_1[i,j]$ is the coefficient of $a_j^{(2)}(t)$ in the conditional expectation of $a_i^{(1)}(t+1)$ given $\boldsymbol{a}^{(2)}(t)$. Thus (C.4) implies that

$$B_1[i,j] = s_i^{(1)} \left[ \frac{\sum_{i'=1}^{R} Z[i',j](s_{i'}^{(1)}/N_{i'\bullet})}{\sum_{l=1}^{R} s_l^{(1)}} - \frac{Z[i,j]}{N_{i\bullet}} \right].$$

Noting that $\frac{s_i^{(1)}}{N_{i\bullet}} = \frac{1}{N_{i\bullet}+\lambda_A}$ and $w_i^{(1)} = \frac{s_i^{(1)}}{\sum_{i'} s_{i'}^{(1)}}$, we simplify the expression as

$$B_1[i,j] = -\frac{Z[i,j]}{N_{i\bullet}+\lambda_A} + w_i^{(1)} \sum_{i'=1}^{R} \frac{Z[i',j]}{N_{i'\bullet}+\lambda_A} = -\frac{Z[i,j]}{N_{i\bullet}+\lambda_A} + w_i^{(1)} u_j \tag{C.5}$$

for $u_j = \sum_{i'=1}^{R} \frac{Z[i',j]}{N_{i'\bullet}+\lambda_A}$. Similarly,

$$B_2[j,i] = -\frac{Z^{\mathsf{T}}[j,i]}{N_{\bullet j}+\lambda_B} + w_j^{(2)} \sum_{j'=1}^{C} \frac{Z^{\mathsf{T}}[j',i]}{N_{\bullet j'}+\lambda_B} = -\frac{Z^{\mathsf{T}}[j,i]}{N_{\bullet j}+\lambda_B} + w_j^{(2)} l_i \tag{C.6}$$

for $l_i = \sum_{j'=1}^{C} \frac{Z^{\mathsf{T}}[j',i]}{N_{\bullet j'}+\lambda_B}$.

A simple computation gives that $\boldsymbol{a}^{(2)}(t)$ is a Gaussian autoregressive process with autoregression matrix $B_2 B_1$. Combining expressions in (C.5) and (C.6) we obtain

$$(B_2 B_1)_{js} = \sum_{i=1}^{R} \frac{Z^{\mathsf{T}}[j,i]}{N_{\bullet j}+\lambda_B} \left( \frac{Z[i,s]}{N_{i\bullet}+\lambda_A} - w_i^{(1)} u_s \right)$$

$$- \quad w_j^{(2)} \times \sum_{i=1}^{R} l_i \left( \frac{Z[i,s]}{N_{i\bullet} + \lambda_A} - w_i^{(1)} u_s \right). \tag{C.7}$$

To analyze the convergence rate of the two component Gibbs sampler it suffices to study the convergence of one chain (see Roberts and Rosenthal (2001)). Then the autoregression matrix for collapsed Gibbs sampler is obtained upon translating (C.7) in matrix notation :

$$M = (I_C - \boldsymbol{w}^{(2)} \mathbf{1}_C^{\mathsf{T}}) M_0.$$

By (Roberts and Sahu, 1997, Theorem 1), we have

$$t_{rel} = \frac{1}{1 - \rho(M)}. \tag{C.8}$$

### C.1.2 Proof of Lemma 4.3.2

Letting $X \preccurlyeq Y$ mean that $X$ is stochastically smaller than $Y$, we know that

$$\begin{aligned}
\mathrm{Bin}(R, \frac{1}{\Upsilon'} S^{1-\rho-\kappa}) \preccurlyeq \sum_i Z_{ij} \preccurlyeq \mathrm{Bin}(R, \Upsilon S^{1-\rho-\kappa}), \quad \text{and} \\
\mathrm{Bin}(C, \frac{1}{\Upsilon'} S^{1-\rho-\kappa}) \preccurlyeq \sum_j Z_{ij} \preccurlyeq \mathrm{Bin}(C, \Upsilon S^{1-\rho-\kappa}).
\end{aligned}$$

By Lemma 4.3.1, if $\psi \geqslant 0$, then

$$\begin{aligned}
\mathbb{P}(\sum_j Z_{ij} \geqslant S^{1-\rho}(\Upsilon + \psi)) &\leqslant \mathbb{P}\big(\mathrm{Bin}(C, \Upsilon S^{1-\rho-\kappa}) \geqslant S^{1-\rho}(\Upsilon + \psi)\big) \\
&\leqslant \exp(-2(S^{1-\rho}\psi)^2/C),
\end{aligned}$$

$$\mathbb{P}\Big(\max_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \geqslant S^{1-\rho}(\Upsilon + \psi)\Big) \leqslant R \exp(-2S^{2-\kappa-2\rho}\psi^2).$$

Similarly,

$$\mathbb{P}\Big(\min_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \leqslant S^{1-\rho}(\frac{1}{\Upsilon'} - \psi)\Big) \leqslant R \exp(-2S^{2-\kappa-2\rho}\psi^2).$$

Therefore for any $\psi > 0$,

$$\mathbb{P}\Big(S^{1-\rho}(\frac{1}{\Upsilon'} - \psi) \leqslant \min_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \leqslant \max_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \leqslant S^{1-\rho}(\Upsilon + \psi)\Big)$$
$$\geq 1 - 2R \exp(-2S^{2-\kappa-2\rho}\psi^2).$$

If $\rho + \frac{1}{2}\kappa < 1$, then for any fixed $\psi > 0$,

$$\lim_{S\to\infty} \mathbb{P}\Big(S^{1-\rho}(\frac{1}{\Upsilon'} - \psi) \leqslant \min_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \leqslant \max_{1 \leqslant i \leqslant R} \sum_j Z_{ij} \leqslant S^{1-\rho}(\Upsilon + \psi)\Big) = 1.$$

Similarly, for any $\psi > 0$,

$$\mathbb{P}\Big(S^{1-\kappa}(\frac{1}{\Upsilon'} - \psi) \leqslant \min_{1 \leq j \leq C} \sum_i Z_{ij} \leqslant \max_{1 \leq j \leq C} \sum_i Z_{ij} \leqslant S^{1-\kappa}(\Upsilon + \psi)\Big)$$
$$\geq 1 - 2C \exp(-2S^{2-\rho-2\kappa}\psi^2).$$

If $\frac{1}{2}\rho + \kappa < 1$, then for any fixed $\psi > 0$,

$$\lim_{S\to\infty} \mathbb{P}\Big(S^{1-\kappa}(\frac{1}{\Upsilon'} - \psi) \leqslant \min_{1 \leq j \leq C} \sum_i Z_{ij} \leqslant \max_{1 \leq j \leq C} \sum_i Z_{ij} \leqslant S^{1-\kappa}(\Upsilon + \psi)\Big) = 1.$$

### C.1.3 Proof of Lemma 4.3.3

Using the inequality $\|.\|_2 \leq \sqrt{\|.\|_1 \|.\|_\infty}$, we obtain $\|Z\|_2 \leq \sqrt{\max_i N_{i\bullet} \max_j N_{\bullet j}}$. Using lemma 4.3.2, under the model (4.10) and the condition in (4.6), we obtain $\|Z\|_2 \leq (\Upsilon + 1)\frac{S}{\sqrt{RC}}$ with high probability.

## C.2 Some additional figures

Below we give the trace plot, effective sample size and autocorrelation plot for $(\rho, \kappa) = (0.36, 0.66)$. Between the two factors, the mixing and effective sample size seem to be poorer for the factor with lower number of levels.

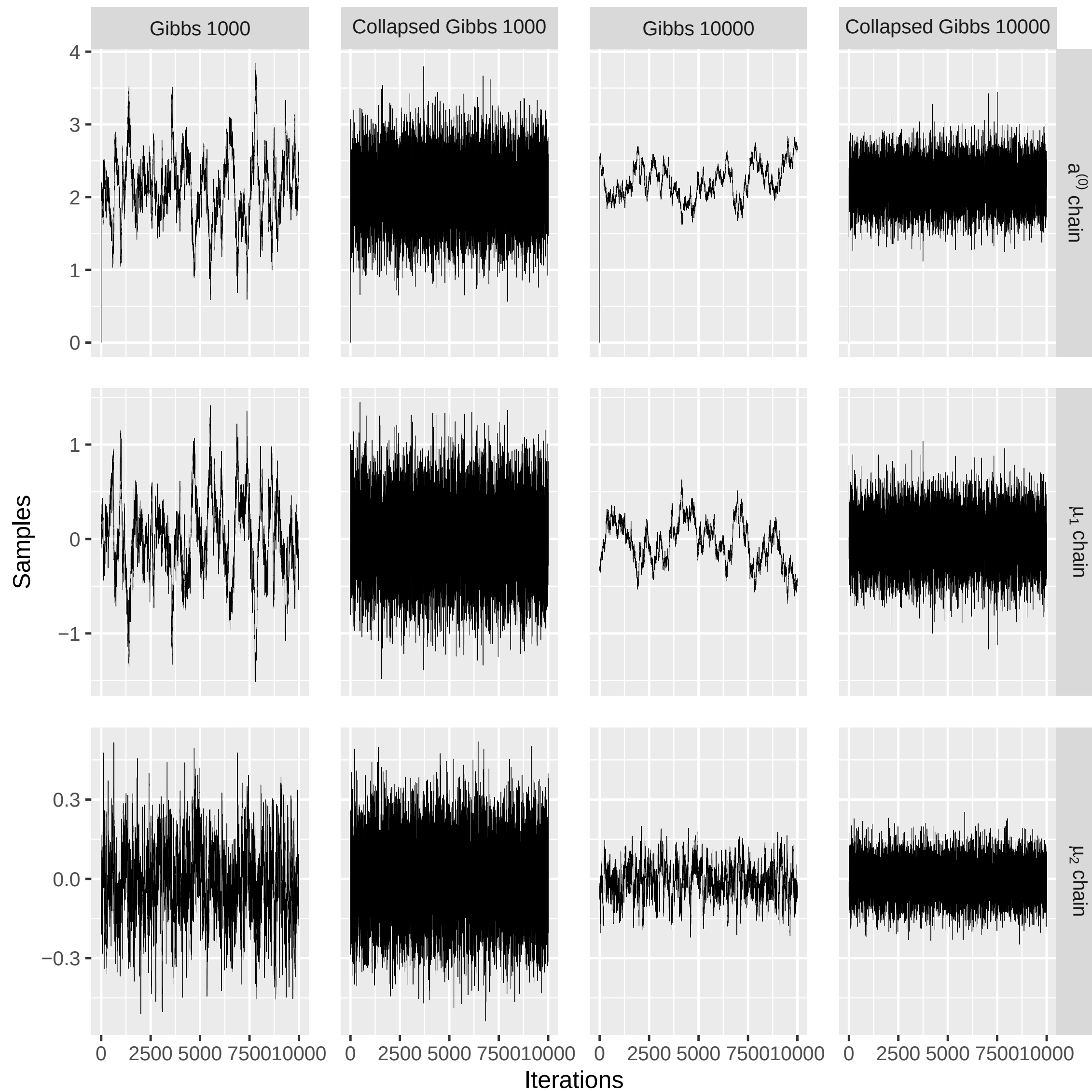



Figure C.1: Trace plots of Gibbs sampler and collapsed Gibbs sampler for mean parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.36} \rceil, C = \lceil S^{0.66} \rceil$. This indicates poor mixing of Gibbs sampler and increasing complexity with sample size. The mixing seems to be poorer for the factor with lower number of levels.

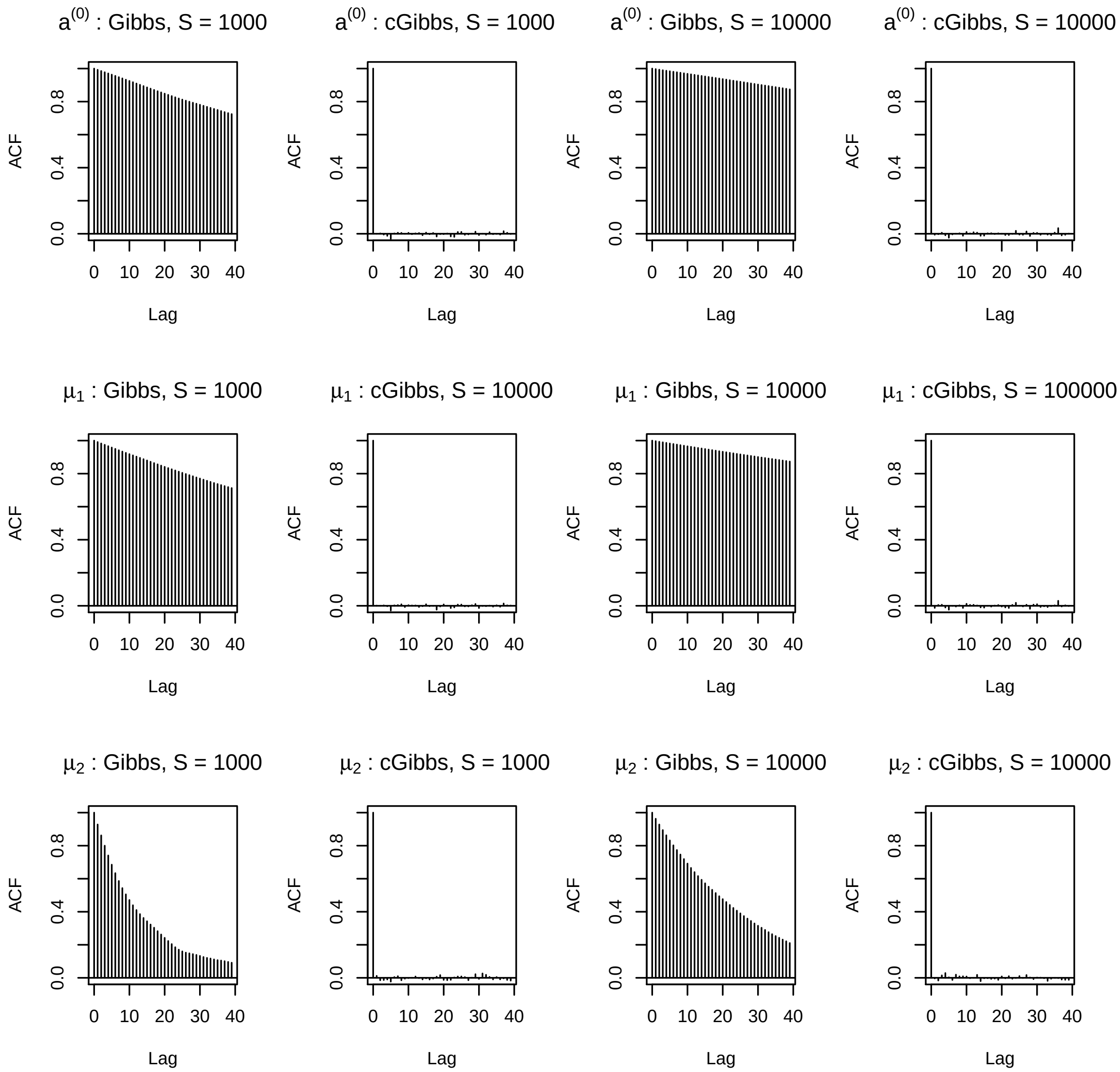



Figure C.2: Autocorrelation plot of Gibbs sampler and collapsed Gibbs (cGibbs) sampler for mean parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.36} \rceil, C = \lceil S^{0.66} \rceil$. This indicates that the autocorrelation of the Gibbs sampler is significantly higher than that of the collapsed Gibbs sampler.

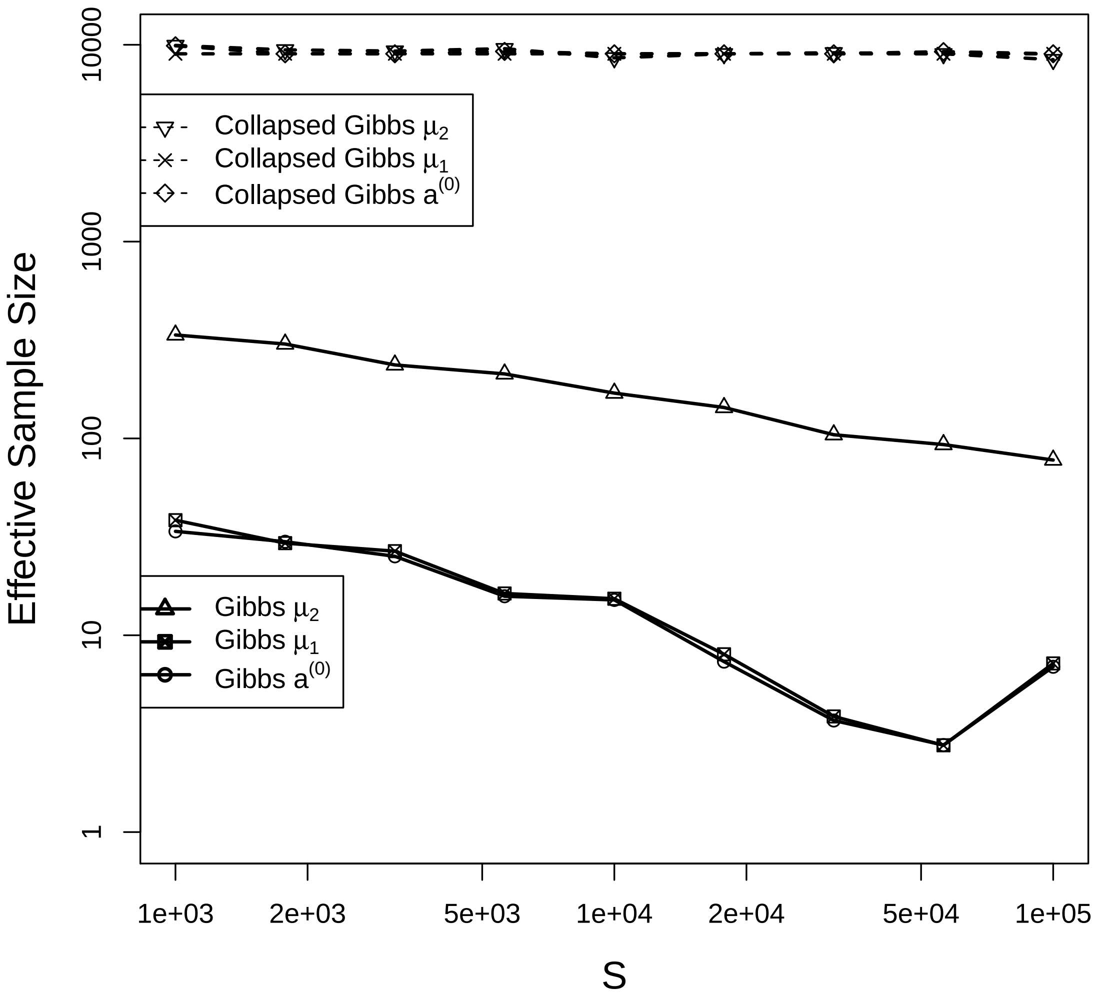



Figure C.3: Effective sample size for mean parameters of crossed random effects of Gibbs and collapsed Gibbs sampler versus problem sizes ($S$) under the model (4.68) with $R = \lceil S^{0.36} \rceil, C = \lceil S^{0.66} \rceil$. The plot has a logarithmic horizontal and vertical scale.

We observe that for Gibbs sampler, the effective sample size exhibits a decreasing trend overall, with the exception of the last data point for $a^{(0)}$ and $\mu_1$ (which could be due to the randomness).

For the figures above and in Sect. 4.4 we assumed precision parameters to be known. We observe a similar behavior of trace plot, autocorrelation and effective sample size as we did for the unknown precision parameter case. For brevity, we give the trace plot for mean and precision parameters for the simulation set up in model (4.68) with flat prior on ${\tau_1}^{-\frac{1}{2}}$, ${\tau_2}^{-\frac{1}{2}}$ and ${\tau_E}^{-\frac{1}{2}}$.

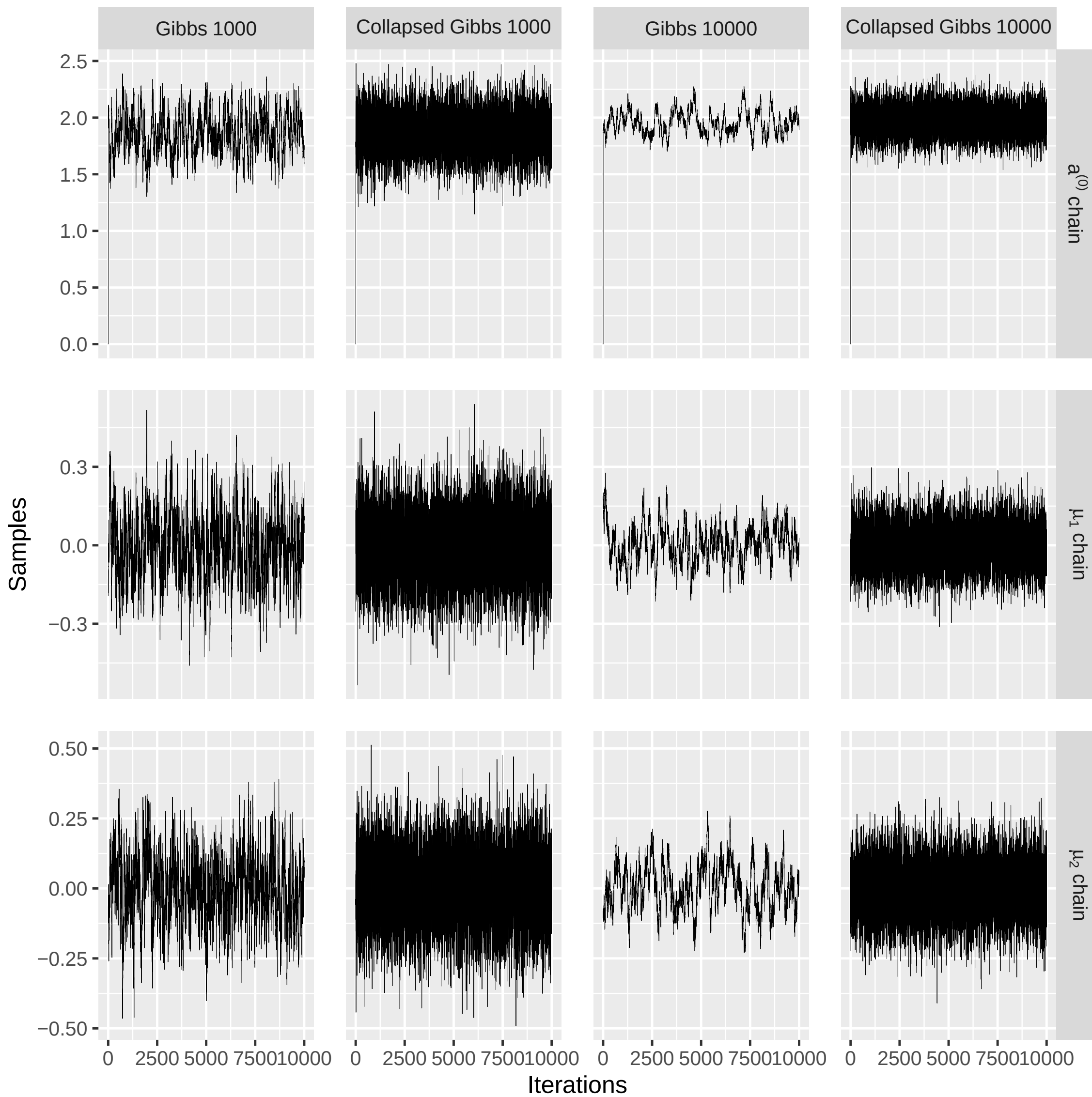



Figure C.4: Trace plots of Gibbs sampler and collapsed Gibbs sampler for mean parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.52} \rceil, C = \lceil S^{0.52} \rceil$. This indicates poor mixing of Gibbs sampler and increasing complexity with sample size. The mixing seems to be poorer for the factor with lower number of levels.

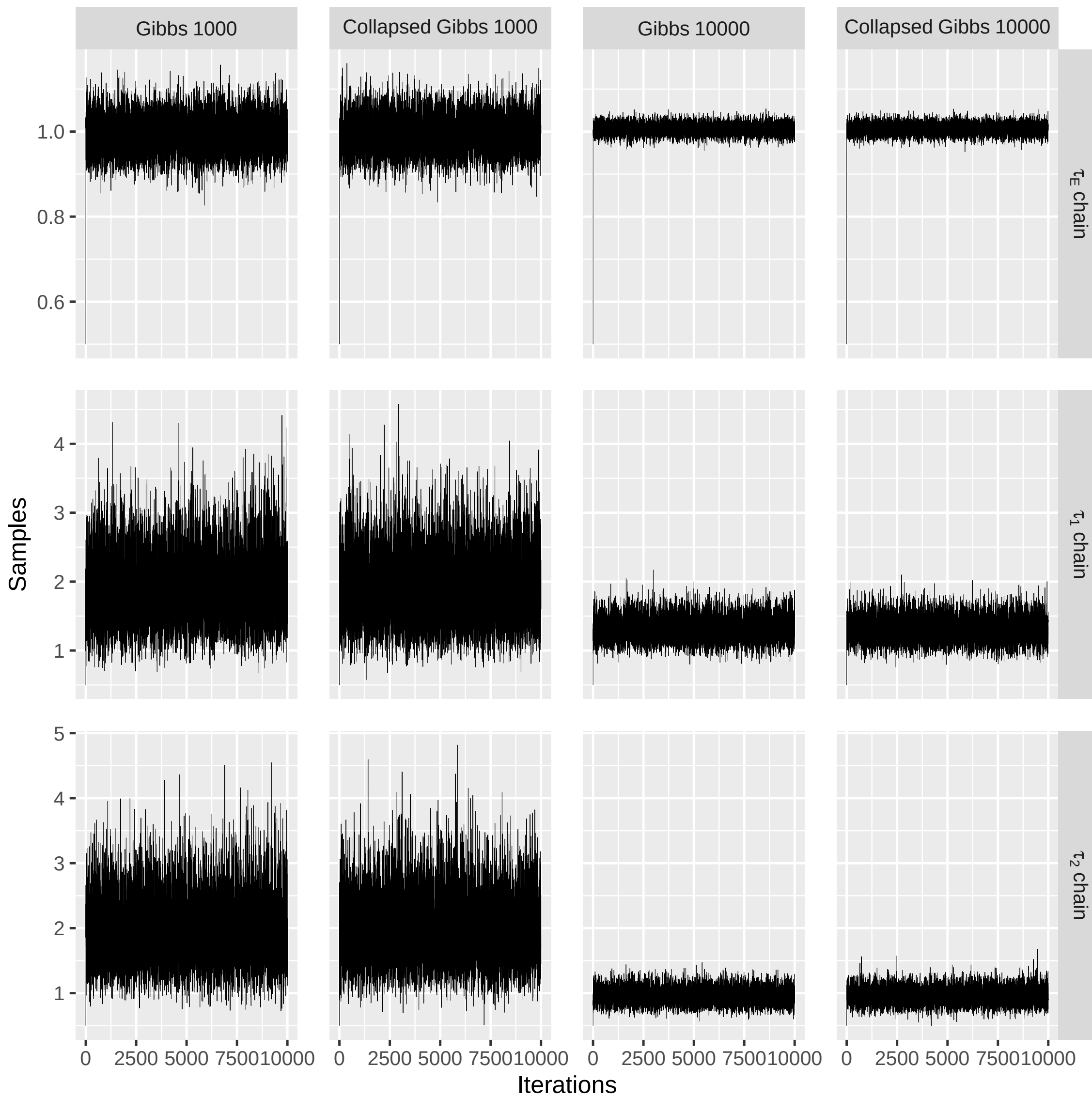



Figure C.5: Trace plots of Gibbs sampler and collapsed Gibbs sampler for precision parameters of crossed random effects model in (4.68) for two problem sizes (S) 1000 and 10000 with $R = \lceil S^{0.52} \rceil, C = \lceil S^{0.52} \rceil$. This indicates similar mixing of Gibbs sampler and collapsed Gibbs sampler.

## Correlation of posterior sample

Below we give the scatter plot of the posterior sample of $a^{(0)}$ and $\mu_2$ for Gibbs sampler on Stitch Fix data. Figure C.6 explains why we see a similar effective sample size for $a^{(0)}$ and $\mu_2$ chain.

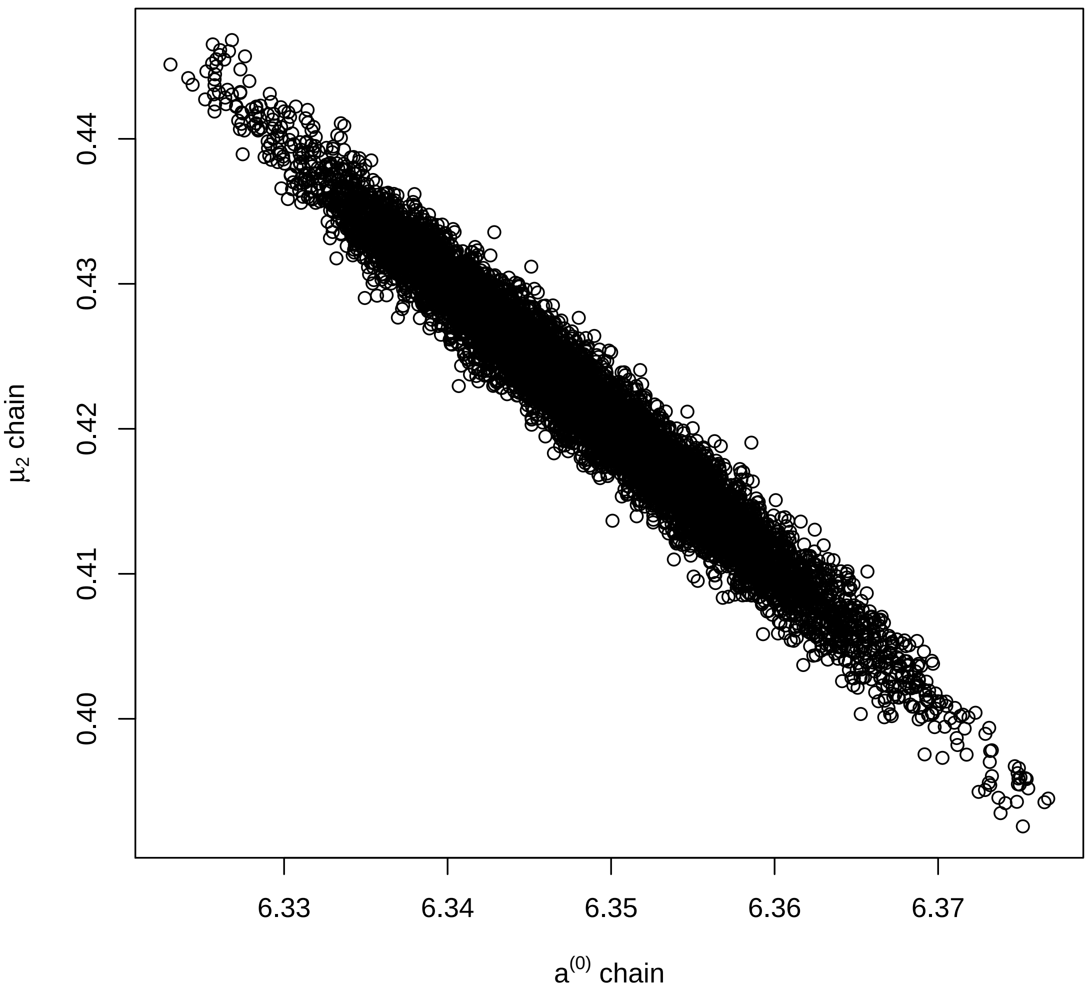



Figure C.6: Scatter plot of $a^{(0)}$ and $\mu_2$ chain.

# Bibliography


Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Bates, D. (2014). Computational methods for mixed models. Technical report, Department of Statistics, University of Wisconsin–Madison. `https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf`.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5(3):217–227.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H., and White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear models. *Journal of the American Statistical Association*, 88(421):9–25.

Brito, G., Dumitriu, I., and Harris, K. (2022). Spectral gap in random bipartite biregular graphs and applications. *Combinatorics, Probability and Computing*, 31(2):229–267.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics*, 17(2):453–510.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Coey, D. and Bailey, M. (2016). People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1103–1111. International World Wide Web Conferences Steering Committee.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Dick, J., Kuo, F. Y., and Sloan, I. H. (2013). High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288.

Gao, K. (2017). *Scalable Estimation and Inference for Massive Linear Mixed Models with Crossed Random Effects*. PhD thesis, Stanford University.

Gao, K. and Owen, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics*, 11(1):1235–1296.

Gao, K. and Owen, A. B. (2020). Estimation and inference for very large linear mixed effects models. *Statistica Sinica*, 30:1741–1771.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The annals of statistics*, 33(1):1–53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.

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

Gelman, A. and Shirley, K. (2011). *Inference from simulations and monitoring convergence*, volume 6. CRC Press Boca Raton, FL.

Geyer, C. J. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, New York.

Ghosh, S., Hastie, T., and Owen, A. B. (2021). Scalable logistic regression with crossed random effects. *arXiv preprint arXiv:2105.13747*.

Ghosh, S., Hastie, T., and Owen, A. B. (2022). Backfitting for large scale crossed random effects regressions. *Annals of Statistics*, 50(1):560–583.

Ghosh, S. and Zhong, C. (2021). Convergence rate of a collapsed gibbs sampler for crossed random effects models. *arXiv preprint arxiv:2109.02849*.

Gil, P. (2021). 54 predictions about the state of data in 2021. `https://www.forbes.com/sites/gilpress/2021/12/30/54-predictions-about-the-state-of-data-in-2021`. [Online; accessed: 2022-05-07].

Goodman, J. and Sokal, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Physical Review D*, 40(6):2035–2071.

Goplerud, M. (2021). Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models Using Variational Inference. *Bayesian Analysis*, pages 1 – 28.

Greenhill, C., Kim, J. H., and Wormald, N. C. (2004). Hamiltonian decompositions of random bipartite regular graphs. *Journal of Combinatorial Theory, Series B*, 90(2):195–222.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, Boca Raton, FL.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111.

Heiss, F. and Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144(1):62–80.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Jiang, J. (2013). The subset argument and consistency of mle in glmm: Answer to an open problem and beyond. *The Annals of Statistics*, 41(1):177–195.

Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Ferres, J. L., and Melamed, T. (2009). Online experimentation at microsoft. *Data Mining Case Studies*, 11.

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94(4):939–752.

Kuo, F. Y., Dunsmuir, W. T. M., Sloan, I. H., Wand, M. P., and Womersley, R. S. (2008). Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability*, 10(2):239–275.

Laird, N. M. (1978). Empirical bayes methods for two-way contingency tables. *Biometrika*, 65(3):581–590.

Latała, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 37:23–37.

Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

McCulloch, C. E. and Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1):270–279.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, PA.

Owen, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411.

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2020). Scalable inference for crossed random effects models. *Biometrika*, 107(1):25–40.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, 18(4):321–349.

Roberts, G. O. and Rosenthal, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28(3):489–504.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317.

Robinson, G. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6(1):15–51.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Searle, S. R., Casella, G., and McCulloch, C. E. (2006). *Variance components*. John Wiley & Sons.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971.

Stitch Fix, Inc. (2011). `https://www.stitchfix.com/`. Accessed: 2022-05-07.

Strassen, V. (1969). Gaussian elimination is not optimal. *Numerische mathematik*, 13(4):354–356.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95.

Zanella, G. and Roberts, G. (2021). Multilevel linear models, Gibbs samplers and multigrid decompositions. *Bayesian Analysis*, pages 1 – 35.

Zhu, Y. (2020). On the second eigenvalue of random bipartite biregular graphs. Technical report, arXiv:2005.08103.