Electronic Journal of Statistics
Vol. 16 (2022) 4604–4635
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2047

Scalable logistic regression with crossed random effects

Swarnadip Ghosh, Trevor Hastie and Art B. Owen

Department of Statistics, Stanford University
e-mail: raswa281@stanford.edu; hastie@stanford.edu; owen@stanford.edu

Abstract: The cost of both generalized least squares (GLS) and Gibbs sampling in a crossed random effects model can easily grow faster than $N^{3/2}$ for N observations. Ghosh et al. (2022) develop a backfitting algorithm that reduces the cost to $O(N)$. Here we extend that 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 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 logistic regression could underestimate sampling variances by several hundred fold.

MSC 2010 subject classifications: Primary 62J12; secondary 65C60.
Keywords and phrases: Generalized linear mixed models, backfitting, Schall's algorithm, quasi-likelihood.

Received December 2021.

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. Ghosh et al. (2022) 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 $\boldsymbol{x}_{ij} \in \mathbb{R}^p$ 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 $\boldsymbol{a} = (a_1, \ldots, a_R)^\mathsf{T}$ and $\boldsymbol{b} = (b_1, \ldots, b_C)^\mathsf{T}$ the Y_{ij} are independent with

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

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 $\boldsymbol{x}_{ij}$ are nonrandom, either because they were designed, or more usually because our analysis is conditional on their observed values.

We can also write (1) as

$$Y_{ij} = \begin{cases} 1, & \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j + e_{ij} > 0 \\ 0, & \text{else} \end{cases} \tag{2}$$

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

$$\Pr(e_{ij} \leqslant w) = \pi(w) \equiv \frac{e^w}{1+e^w}. \tag{3}$$

The model (1) is a generalized linear mixed model (GLMM) owing to the appearance of both fixed effects $\boldsymbol{x}_{ij}^{\mathsf{T}}\beta$ and random effects a_i, b_j.

While the linear model $Y_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j + e_{ij}$ 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 (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. We choose to work with the Gaussian model because it is of central importance and existing approaches do not scale adequately for it.

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

$$\begin{aligned} L(\beta \,|\, \boldsymbol{a}, \boldsymbol{b}) &= \prod_{(i,j)} \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j)^{Y_{ij}} (1 - \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j))^{1-Y_{ij}} \\ &= \prod_{(i,j)} \frac{e^{(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j)Y_{ij}}}{1 + e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j}} \end{aligned} \tag{4}$$

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

$$L(\beta, \sigma_A^2, \sigma_B^2) = \int_{\mathbb{R}^{R+C}} L(\beta \,|\, \boldsymbol{a}, \boldsymbol{b}) \prod_{i=1}^{R} \frac{1}{\sigma_A} \varphi\Big(\frac{a_i}{\sigma_A}\Big) \prod_{j=1}^{C} \frac{1}{\sigma_B} \varphi\Big(\frac{b_j}{\sigma_B}\Big) \,\mathrm{d}\boldsymbol{a}\,\mathrm{d}\boldsymbol{b} \tag{5}$$

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 (5) presents a major difficulty to finding estimates and confidence intervals for β, making the generalized linear mixed model (1) much harder to work with than Gaussian linear mixed models for regression studied in Ghosh et al. (2022). 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.

A standard 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. A tensor product integration with k points per dimension would have a cost proportional to k^{R+C} which is

infeasible in our setting. 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 β under a diffuse prior that we describe below. It is thought to be close to the MLE except when some $\Pr(Y_{ij} = 1 \mid a_i, b_j)$ 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 μ 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 Ghosh et al. (2022) to develop a generalized least squares algorithm that alternates between updating a_i along with the intercept and updating b_j also with the intercept.

A related issue arises when one of the predictor variables in $\boldsymbol{x}_{ij}$, 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 β_ℓ 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 β with $\boldsymbol{a}$ when updating $\boldsymbol{a}$, and also clubs β with $\boldsymbol{b}$ when updating $\boldsymbol{b}$. The algorithm analyzed in Ghosh et al. (2022) 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.

An outline of this paper is as follows. Section 2 introduces our notation and describes the missingness mechanism. Section 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 4 describes in detail our clubbing strategy where we alternate between updating $(\hat{\beta}, \hat{\boldsymbol{a}})$ and $(\hat{\beta}, \hat{\boldsymbol{b}})$. Section 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 Ghosh et al. (2022) give 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 6 illustrates our algorithm on a data set provided by Stitch Fix. Section 7 has a discussion about the Bayesian and frequentist perspectives. In Appendix A, 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 (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. Composite likelihoods (Varin et al., 2011) lower the dimension of the integration problem by multiplying likelihoods based on selected pairs or k-tuples of observations instead of all N of them at once. 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 this paper. We have opted to build upon the penalized quasi-likelihood (PQL) 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, but they are beyond the scope of this paper. Getting proper confidence intervals for the resulting predictions would be significantly harder than handling β as we do here. Even choosing the number of latent variable interactions to include is difficult in such problems. Furthermore, quantifying the sampling uncertainty on the estimated interactions brings in difficult random matrix theory problems.

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 $(\boldsymbol{x}_{ij}, Y_{ij})$ 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 $\mathcal{Z}_A^{\mathsf{T}} \mathcal{Z}_B$ 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 Ghosh et al. (2022). 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

$$Z_{ij} \overset{\text{ind}}{\sim} \text{Bern}(p_{ij}) \quad \text{where} \quad \frac{S}{RC} \leqslant p_{ij} \leqslant \Upsilon \frac{S}{RC}. \tag{6}$$

The actual problem size N satisfies $S \leqslant \mathbb{E}(N) \leqslant \Upsilon S$. The model (6) allows unequal observation probabilities but remains more restrictive than we would like. Numerical results in Ghosh et al. (2022) show that their generalized least squares algorithm converges in $O(N)$ cost much more generally than their assumptions imply.

The relevant values of ρ and κ 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 (6), Ghosh et al. (2022) show that if $2\rho + \kappa < 2$ then all rows get an adequate sample size and no single row dominates the data:

$$\lim_{S \to \infty} \Pr\big((1 - \epsilon) S^{1-\rho} \leqslant \min_i N_{i\bullet} \leqslant \max_i N_{i\bullet} \leqslant (\Upsilon + \epsilon) S^{1-\rho}\big) = 1 \tag{7}$$

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

$$\lim_{S\to\infty} \Pr\big((1-\epsilon)S^{1-\kappa} \leqslant \min_j N_{\bullet j} \leqslant \max_j N_{\bullet j} \leqslant (\Upsilon+\epsilon)S^{1-\kappa}\big) = 1 \tag{8}$$

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. Likelihood approach

We will use the iterative approach of Schall (1991) with further justification given the in the next paragraph. It alternates between estimating $\boldsymbol{a}$, $\boldsymbol{b}$ and β given σ_A^2, σ_B^2 and an overdispersion quantity ϕ and estimating the variances and overdispersion given $\boldsymbol{a}$, $\boldsymbol{b}$ and β. For the optimization over $\boldsymbol{a}$ and $\boldsymbol{b}$ and β 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 the approach in Schall (1991) and PQL of Breslow and Clayton (1993). We observe Y_{ij} together with predictor variables $\boldsymbol{x}_{ij}$. There are also unobserved random effects (a_i, b_j) associated with these data. We assume 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 $\mu_{ij}^{\boldsymbol{a},\boldsymbol{b}} \equiv \mathbb{E}(Y_{ij} \,|\, a_i, b_j) = \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j)$. The model of Breslow and Clayton (1993) includes

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

for a function $v(\cdot)$ and an overdispersion parameter $\phi > 0$. In our case $v(\mu) = \mu(1-\mu)$. Independent Bernoulli observations cannot be overdispersed (McCullagh and Nelder, 1989). However, we include an overdispersion parameter because that one parameter may help with lack of fit of our model. In the Stitch Fix data of Section 6 we obtained $\hat{\phi} = 0.75$ a mild underdispersion.

We write the estimating equations based on approximate quasi-likelihood from Breslow and Clayton (1993). The gradient in equations (7) and (8) of Breslow and Clayton (1993) involves the quantity $v(\mu)\pi'(\mu)$ which equals 1 for v and μ in our case. Translating the score equation to our context, we obtain

the following three equations:

$$\begin{aligned}\mathcal{X}^{\mathsf{T}}(\mathcal{Y}-\mu^{\boldsymbol{a},\boldsymbol{b}}) &= \mathbf{0},\\ \mathcal{Z}_A^{\mathsf{T}}(\mathcal{Y}-\mu^{\boldsymbol{a},\boldsymbol{b}}) &= \frac{\phi}{\sigma_A^2}\boldsymbol{a}, \quad \text{and}\\ \mathcal{Z}_B^{\mathsf{T}}(\mathcal{Y}-\mu^{\boldsymbol{a},\boldsymbol{b}}) &= \frac{\phi}{\sigma_B^2}\boldsymbol{b}.\end{aligned} \tag{9}$$

Using the notion of a working response z given by

$$z_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j + \frac{Y_{ij}-\mu_{ij}}{\mu_{ij}(1-\mu_{ij})} \tag{10}$$

with $\mu_{ij} = \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j)$ for the logistic CDF $\pi(\cdot)$ from equation (3), we obtain the following estimating equations from (9):

$$\begin{aligned}\mathcal{X}^{\mathsf{T}}\mathcal{W}(z-\mathcal{X}\beta-\mathcal{Z}_A\boldsymbol{a}-\mathcal{Z}_B\boldsymbol{b}) &= \mathbf{0},\\ \mathcal{Z}_A^{\mathsf{T}}\mathcal{W}(z-\mathcal{X}\beta-\mathcal{Z}_A\boldsymbol{a}-\mathcal{Z}_B\boldsymbol{b}) - \frac{1}{\sigma_A^2}\boldsymbol{a} &= \mathbf{0}\\ \mathcal{Z}_B^{\mathsf{T}}\mathcal{W}(z-\mathcal{X}\beta-\mathcal{Z}_A\boldsymbol{a}-\mathcal{Z}_B\boldsymbol{b}) - \frac{1}{\sigma_B^2}\boldsymbol{b} &= \mathbf{0}\end{aligned} \tag{11}$$

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

3.1. Schall's approach

Schall (1991) considers an iterative procedure to estimate the fixed effects, random effects and the variance components. At iteration k we have estimates $\hat\beta^{(k)}$, $\hat{\boldsymbol{a}}^{(k)}$, $\hat{\boldsymbol{b}}^{(k)}$, $\hat\sigma_A^{2(k)}$, $\hat\sigma_B^{2(k)}$ 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

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \sum_{ij} Z_{ij}\,\widehat{W}_{ij}^{(k)}\big(z_{ij}^{(k)} - \boldsymbol{x}_{ij}^{\mathsf{T}}\beta - a_i - b_j\big)^2 + \frac{\|\boldsymbol{a}\|^2}{\hat\sigma_A^{2(k)}} + \frac{\|\boldsymbol{b}\|^2}{\hat\sigma_B^{2(k)}}, \tag{12}$$

with weights $\widehat{W}_{ij}^{(k)} = \hat\mu_{ij}^{(k)}(1-\hat\mu_{ij}^{(k)})/\hat\phi^{(k)}$ and $\hat\mu_{ij}^{(k)} = \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\hat\beta^{(k)} + \hat a_i^{(k)} + \hat b_j^{(k)})$. This optimization problem leads to exactly the same estimating equations (11) that we found for the approximate quasi-likelihood. A straightforward solution of equation (11) costs $O((R+C+p)^3)$ which is infeasible and so we develop a backfitting iteration for it in Subsection 3.2.

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

$$T^{(k)} = \begin{pmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ T_{21}^{(k)} & T_{22}^{(k)} \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_A^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}\mathcal{Z}_A + \hat\sigma_A^{-2(k)} I_R & \mathcal{Z}_A^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}\mathcal{Z}_B \\ \mathcal{Z}_B^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}\mathcal{Z}_A & \mathcal{Z}_B^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}\mathcal{Z}_B + \hat\sigma_B^{-2(k)} I_C \end{pmatrix} \tag{13}$$

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

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

Schall's update evaluates (see Algorithm 2 in Sect 2. of Schall (1991))

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

and then sets

$$\hat{\sigma}_A^{2(k+1)} = \frac{\|\hat{\boldsymbol{a}}^{(k+1)}\|^2}{R - \nu_A^{(k+1)}} \quad \text{and} \quad \hat{\sigma}_B^{2(k+1)} = \frac{\|\hat{\boldsymbol{b}}^{(k+1)}\|^2}{C - \nu_B^{(k+1)}}. \tag{14}$$

The derivation of this formula is technical and can be found in Harville (1977) (equation 6.5) for the gaussian case. Schall (1991) adapted the derivation for generalized linear models. Upon differentiating the log determinant in the quasi-likelihood, a trace appears that we see in the denominator. The trace term also occurs in Fellner (1986) and Breslow and Clayton (1993). From these formulas it is clear that the ν parameters function as degrees of freedom estimates. Finally, we let

$$\hat{\phi}^{(k+1)} = \frac{\sum_{ij} Z_{ij}\hat{\mu}_{ij}^{(k+1)}(1-\hat{\mu}_{ij}^{(k+1)})\left(z_{ij}^{(k+1)} - x_{ij}^{\mathsf{T}}\hat{\beta}^{(k+1)} - \hat{a}_i^{(k+1)} - \hat{b}_j^{(k+1)}\right)^2}{N - p - (R - \nu_A^{(k+1)}) - (C - \nu_B^{(k+1)})} \tag{15}$$

where p is the number of parameters in β including the intercept.

The quantity Schall computes is almost but not quite the maximum likelihood estimate of β. It is a quantity studied by Stiratelli et al. (1984). The update (12) for β, $\boldsymbol{a}$ and $\boldsymbol{b}$ is a Fisher scoring iteration to maximize their posterior density under a diffuse prior for β independent of a zero-mean Gaussian prior for $(\boldsymbol{a}, \boldsymbol{b})$ with plugged in variance components (see Schall, 1991). 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 updates (14) and (15) for σ_A^2, σ_B^2 and ϕ 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.

3.2. Modified Schall Approach

We want to solve the equation (12). It is instructive to begin with the case of just one factor. Then we obtain the following optimization problem:

$$\min_{\beta,\boldsymbol{a}} \sum_{i,j} Z_{ij}\widehat{W}_{ij}^{(k)}\left(z_{ij}^{(k)} - \boldsymbol{x}_{ij}^{\mathsf{T}}\beta - a_i\right)^2 + \frac{\|\boldsymbol{a}\|^2}{\hat{\sigma}_A^{2(k)}}. \tag{16}$$

The normal equations from (16) yield

$$\mathbf{0} = \mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(z^{(k)} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A\hat{\boldsymbol{a}}), \quad \text{and} \tag{17}$$

$$\mathbf{0} = \mathcal{Z}_A^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(z^{(k)} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A\hat{\boldsymbol{a}}) - \hat{\boldsymbol{a}}/\hat{\sigma}_A^{2(k)}. \tag{18}$$

Solving (18) for $\hat{\boldsymbol{a}}$ and multiplying the solution by $\mathcal{Z}_A$ yields

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

for an $N \times N$ ridge regression "smoother matrix" $\mathcal{S}_A^{(k)}$. (We have rederived one of the estimating equations in (11).) This smoother matrix implements weighted shrunken within-group means. Substituting $\mathcal{Z}_A\hat{\boldsymbol{a}}$ into equation (17) yields

$$\hat{\beta} = \big(\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_A^{(k)})\mathcal{X}\big)^{-1}\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_A^{(k)})z^{(k)}. \tag{19}$$

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

$$\widehat{\operatorname{cov}}(\hat{\beta}) = (\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)} \cdot \widehat{\Sigma} \cdot \widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X}(\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}_{\mathcal{S}_A}^{(k)}\mathcal{X})^{-1}, \tag{20}$$

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 (19) is

$$\hat{\beta} = (\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_{AB}^{(k)})\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}^{(k)}(I_N - \mathcal{S}_{AB}^{(k)})z^{(k)} \tag{21}$$

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 $D^{(k)} = \operatorname{diag}(\hat{\sigma}_A^{2(k)}I_R, \hat{\sigma}_B^{2(k)}I_C)$. Hence we would need to invert an $(R + C) \times (R + C)$ matrix $T^{(k)}$ in (13) to apply $\mathcal{S}_{AB}^{(k)}$ and thereby solve (21), incurring a cost far greater than $O(N)$.

However, in order to solve (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

$$\min_{\boldsymbol{a},\boldsymbol{b}} \|r - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|_{\widehat{\mathcal{W}}^{(k)}}^2 + \hat{\sigma}_A^{-2(k)}\|\boldsymbol{a}\|^2 + \hat{\sigma}_B^{-2(k)}\|\boldsymbol{b}\|^2. \tag{22}$$

It is clear that the fitted values are given by $\widehat{r} = \mathcal{S}_{AB}^{(k)}r$. Solving (22) leads to the following two blocks of estimating equations:

$$\begin{aligned}\mathcal{Z}_A\hat{\boldsymbol{a}} &= \mathcal{S}_A^{(k)}(r - \mathcal{Z}_B\hat{\boldsymbol{b}}), \\ \mathcal{Z}_B\hat{\boldsymbol{b}} &= \mathcal{S}_B^{(k)}(r - \mathcal{Z}_A\hat{\boldsymbol{a}}).\end{aligned} \tag{23}$$

We can solve these equations iteratively by block coordinate descent (backfitting) as done in Ghosh et al. (2022) (see Sect. 3.2). 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 (20), we obtain the covariance estimate for the two factor case.

$$\widehat{\mathrm{cov}}(\hat{\beta}) = (\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)} \cdot \widehat{\Sigma} \cdot \widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X}(\mathcal{X}^\mathsf{T}\widehat{\mathcal{W}}_{\mathcal{S}_{AB}}^{(k)}\mathcal{X})^{-1}, \tag{24}$$

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 (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 4.

Now we consider the second step of Schall's method. 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

$$\nu_A^{(k+1)} = \mathrm{tr}((T_{11}^{(k)})^{-1})/\hat{\sigma}_A^{2(k)} \quad \text{and} \quad \nu_B^{(k+1)} = \mathrm{tr}((T_{22}^{(k)})^{-1})/\hat{\sigma}_B^{2(k)} \tag{25}$$

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

$$T = \begin{pmatrix} \mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_A + \sigma_A^{-2}I_R & \mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_B \\ \mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_A & \mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_B + \sigma_B^{-2}I_C \end{pmatrix} \tag{26}$$

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

$$\mathrm{tr}((\mathcal{Z}_A^\mathsf{T}\mathcal{W}\mathcal{Z}_A + \sigma_A^{-2}I_R)^{-1}) \quad \text{and} \quad \mathrm{tr}((\mathcal{Z}_B^\mathsf{T}\mathcal{W}\mathcal{Z}_B + \sigma_B^{-2}I_C)^{-1}) \tag{27}$$

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

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

we see that the traces in (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

$$T(\eta) = \begin{pmatrix} A & \eta B \\ \eta B^{\mathsf{T}} & C \end{pmatrix} \tag{28}$$

where matrices A, B and C are defined so that $T(1)$ is Schall's matrix T from (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 3.1 we will see that $\operatorname{tr}(T(\eta)^{-1}) - \operatorname{tr}(T(0)^{-1}) = O(\eta^2)$ so that small perturbations ηB do not bring a large approximation error. However we need to study $\eta = 1$ so a small η analysis is not quite enough, so the following Lemma 3.2 shows that $\eta = 1$ is within the radius of convergence.

Lemma 3.1. *Let $T(\eta) \in \mathbb{R}^{(R+C)\times(R+C)}$ have the form (28) for positive definite diagonal matrices A and C. Define $B_* = A^{-1/2} B C^{-1/2}$ and let ρ be the spectral radius of*

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

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

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

Proof. See Appendix A.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

$$\operatorname{tr}\Big(A^{-1} \sum_{k \geqslant 1} (B_* B_*^{\mathsf{T}})^k\Big) \quad \text{and} \quad \operatorname{tr}\Big(C^{-1} \sum_{k \geqslant 1} (B_*^{\mathsf{T}} B_*)^k\Big) \tag{29}$$

after setting $\eta = 1$. To use Lemma 3.1 with $\eta = 1$ we need to verify that the given spectral radius ρ is below 1 for our setting. We use Lemma 3.2 to control the spectral radius.

Lemma 3.2. *In the notation of Lemma 3.1, let B have nonnegative entries. Then the spectral radius ρ satisfies*

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

Proof. See Appendix A.2. □

In our present context $A = \operatorname{diag}(W_{i\bullet} + \sigma_A^{-2})$, $C = \operatorname{diag}(W_{\bullet j} + \sigma_B^{-2})$ 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 $|\boldsymbol{x}_{ij}^{\mathsf{T}}\beta|$ is bounded away from infinity for all (i,j), even the unobserved (i,j) pairs. However, our model uses $\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j$ 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

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

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

$$\lim_{S\to\infty} \Pr\Big(\min_{1\leqslant i\leqslant R} \min_{1\leqslant j\leqslant C} W_{ij} < S^{-\psi}\Big) = 0. \tag{30}$$

We need to bound the largest eigenvalue of $B_* B_*^{\mathsf{T}}$ from equation (29). In our context that matrix equals

$$\operatorname{diag}(W_{i\bullet} + \sigma_A^{-2})^{-1/2} W \operatorname{diag}(W_{\bullet j} + \sigma_B^{-2})^{-1} W^{\mathsf{T}} \operatorname{diag}(W_{i\bullet} + \sigma_A^{-2})^{-1/2}. \tag{31}$$

Proposition 1. *Let R, C and Z_{ij} for $1 \leqslant i \leqslant R$ and $1 \leqslant j \leqslant C$ be sampled as in Section 2 with $\rho, \kappa \in (0,1)$ and $\max(2\rho+\kappa, \rho+2\kappa) < 2$. Let λ_1 be the largest eigenvalue of the matrix at (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 (7) and (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 $\epsilon > 0$

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

Theorem 3.3. *Let $R = S^\rho$, $C = S^\kappa$ and Z_{ij} follow the sampling model from Section 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 (30). Let the matrix $B_*B_*^{\mathsf{T}}$ in (31) have eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_R \geqslant 0$. Then there exists $\alpha < 1$ such that for any $\delta > 0$*

$$\lim_{S\to\infty} \Pr\left(\sum_{i=1}^{R} 1\{\lambda_i > \delta\} > R^\alpha\right) = 0. \tag{32}$$

Proof. See Appendix A.3. $\square$

Theorem 3.4. *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 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 (30). Then our approximation error from (29) is*

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

and for any $\gamma > 0$

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

Proof. See Appendix A.4. $\square$

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 β 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 ψ 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 σ_A^2, σ_B^2 and ϕ 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 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 λ_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 λ_1 in Proposition 1. Very large values are a priori implausible for our target setting. For instance, enormous σ_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 3.4, we only need $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ to be $o(S^\delta)$ for any $\delta > 0$.

4. Clubbed backfitting

The iterative method in Section 3.2 to solve (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 (16):

$$\min_{\beta,\boldsymbol{a}} \|z - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a}\|_{\boldsymbol{w}}^2 + \sigma_A^{-2}\|\boldsymbol{a}\|^2. \tag{33}$$

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 $\ell_1, \ell_2, \ldots, \ell_k$. 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 (33).

Lemma 4.1. *Suppose that column q of $\mathcal{X}$ in problem (33) equals the sum of $k \geqslant 1$ distinct columns $\mathcal{I} = \{i_1, i_2, \ldots, i_k\}$ 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

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

If we evaluate the quadratic in (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 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 (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

$$\mathrm{pl}(\beta,\boldsymbol{a},\boldsymbol{b}) = \sum_{i,j} Z_{ij}\widehat{W}_{ij}\left(z_{ij} - \boldsymbol{x}_{ij}^{\mathsf{T}}\beta - a_i - b_j\right)^2 + \hat{\sigma}_A^{-2}\|\boldsymbol{a}\|^2 + \hat{\sigma}_B^{-2}\|\boldsymbol{b}\|^2. \quad (34)$$

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 β 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 β 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

$$\begin{aligned}\mathcal{X}\beta^{(k+\frac{1}{2})} &= (\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}(I_N-\mathcal{S}_A)\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}(I_N-\mathcal{S}_A)(z-\mathcal{Z}_B\boldsymbol{b}^{(k)}), \quad \text{and}\\ \mathcal{Z}_A\boldsymbol{a}^{(k+1)} &= \mathcal{S}_A(z-\mathcal{X}\beta^{(k+\frac{1}{2})}-\mathcal{Z}_B\boldsymbol{b}^{(k)})\end{aligned} \quad (35)$$

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 (35) solve the following minimization problem,

$$\min_{\beta,\boldsymbol{a}} \sum_{i,j} Z_{ij}\widehat{W}_{ij}\left(z_{ij} - b_j^{(k)} - \boldsymbol{x}_{ij}^{\mathsf{T}}\beta - a_i\right)^2 + \hat{\sigma}_A^{-2}\|\boldsymbol{a}\|^2. \quad (36)$$

If we absorb $b_j^{(k)}$ into z_{ij} we see that (36) has the form (33) and so by Lemma 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 β and $\boldsymbol{b}$, via

$$\begin{aligned}\mathcal{X}\beta^{(k+1)} &= \mathcal{X}(\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}(I_N-\mathcal{S}_B)\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\widehat{\mathcal{W}}(I_N-\mathcal{S}_B)(z-\mathcal{Z}_A\boldsymbol{a}^{(k+1)}), \quad \text{and}\\ \mathcal{Z}_B\boldsymbol{b}^{(k+1)} &= \mathcal{S}_B(z-\mathcal{X}\beta^{(k+1)}-\mathcal{Z}_A\boldsymbol{a}^{(k+1)})\end{aligned} \quad (37)$$

for a smoother matrix $\mathcal{S}_B = \mathcal{Z}_B(\mathcal{Z}_B^{\mathsf{T}}\widehat{\mathcal{W}}\mathcal{Z}_B + \hat{\sigma}_B^{-2}I_C)^{-1}\mathcal{Z}_B^{\mathsf{T}}\widehat{\mathcal{W}}$. Again Lemma 4.1 applies to the solution.

Lemma 4.2. *If $\mathcal{X}$ is of full rank, $\max(\hat{\sigma}_A^2,\hat{\sigma}_B^2) < \infty$ and the weights $\widehat{W}_{ij}$ are positive, then the iterative algorithm converges to a global minimum of the equation (34).*

Proof. See Appendix A.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 ζ 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 ℓ 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 (35) and (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 ν from (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 (24). As a consequence of Lemma 4.1, if $\mathcal{X}$ contains the intercept the penalized least squares problem in (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 Ghosh et al., 2022) $\tilde{\mathcal{S}}_A^{(k)}$ and $\tilde{\mathcal{S}}_B^{(k)}$ in (23) instead of $\mathcal{S}_A^{(k)}$ and $\mathcal{S}_B^{(k)}$.

Our numerical results in the Section 5 use $\epsilon = 10^{-8}$. Open-source R code at https://github.com/G28Sw/scalable-logistic-regression-with-crossed-random-effects does these computations.

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 β 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 2. We look at timings and we also consider accuracy. These experiments were carried out in R language (R Core Team, 2021) on a computer with the macOS operating system, 16 GB of memory and an Intel i7 processor.

5.1. Timings

To study how the computation time varies with N we generated data over a range of sample sizes using $\rho = \kappa = 0.56$. We used $\Upsilon = 1.2716$ in our simulations to generate Z_{ij}. Our predictors were $\boldsymbol{x}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ 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 β:

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 $\beta_0 = -2$.

We took $\sigma_A = 0.8$ and $\sigma_B = 0.4$.

The rationale for our choices is as follows. The point $(\rho, \kappa) = (0.56, 0.56)$ is just barely inside the region in Figure 1 of Ghosh et al. (2022) where it is proved that backfitting for generalized least squares converges in $O(1)$ iterations. Larger values would severely raise the computational times for glmer that we simulate. The value of $\Upsilon = 1.2716$ is the largest one for which Ghosh et al. (2022) were able to prove that $O(1)$ iterations are required. We do not expect the computational time should depend on β. We took $\beta_0 < 0$ to reflect that our cases of interest have $\mathbb{E}(Y) < 1/2$. We are interested in studying the accuracy of estimation for intercept and other coefficients separately. The cases we chose show different behavior in comparison to naive logistic regression in Section 5.3. The variance component values we chose for simulation are close to those we have fit to the Stitch Fix data where we obtained $\hat{\sigma}_A^2 = 0.68$ (customers), and $\hat{\sigma}_B^2 = 0.21$ (items).

We included our backfit iteratively reweighted least squares algorithm, a naive logistic regression ignoring random effects 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.

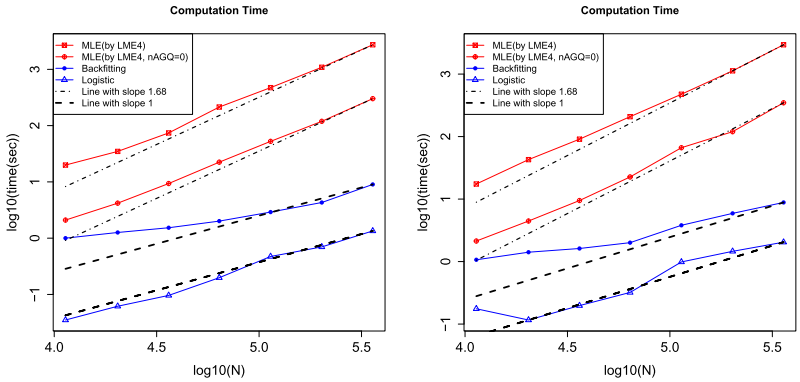

FIG 1. *Time versus N at $(\rho, \kappa) = (0.56, 0.56)$ for $\beta_\ell = 0$ (other than intercept) and the linearly trending β_ℓ. The left panel plots the computational time for logistic regression when true $\beta_0 = -2$ and rest of the β's are zero. The right panel has trending β. The cost for glmer is roughly $O(N^{\max(3\rho, 3\kappa)})$. The cost for naive logistic regression and backfitting is $O(N)$.*

Figure 1 shows our timings for two choices of β. 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 the choices of β in a) and b), backfitting took 10–12 iterations to converge for the smaller values of N and 8–9 iterations for the larger ones. Our convergence criterion was a relative change of 10^{-8} as described in Section 4. At the largest values of N, the timings follow the same asymptotic rates, approximately $O(N^{1.68})$ and $O(N)$. 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.

5.2. *Accuracy*

It is also important to compare the accuracy of these algorithms. We computed estimates of β using logistic regression, clubbed backfitting, and glmer from Bates et al. (2015) with maximum number of function evaluations set at 1000. For each we computed the estimates and compute the mean squared error (MSE) of the parameters over 100 replicates.

For accuracy we chose S to get $\log_{10}(N)$ in the range from 3 to 4.5 approximately for glmer algorithms. The value of N varies only slightly given S with $S \leqslant \mathbb{E}(N) \leqslant \Upsilon S$ and we plot the average N for each S that we used. Because naive logistic regression and backfitting have cost $O(N)$ we were able to study them over a greater sample size range, up to $\log_{10}(N) \approx 5.5$.

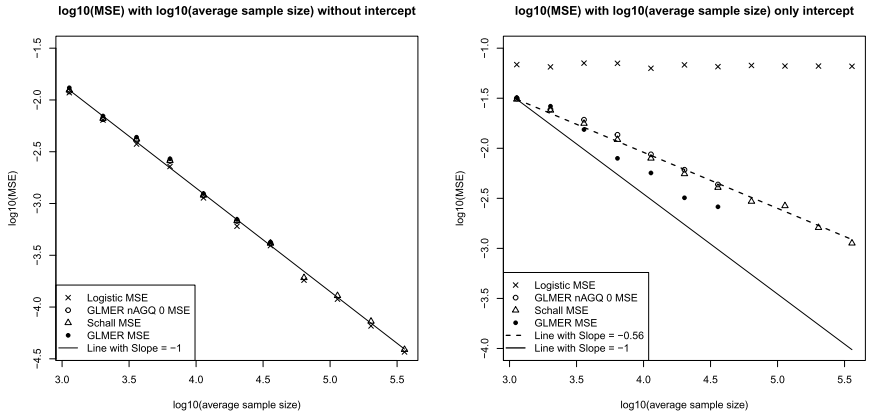

FIG 2. *MSE of $\hat{\beta}$ vs average sample size at $(\rho, \kappa) = (0.56, 0.56)$ for $\beta_\ell = 0$ (other than the intercept). The left panel plots the MSE of non intercept β with true value being zero. The right panel plots the MSE for the intercept with true intercept being -2. MSE for glmer, 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 regression appears to be inconsistent.*

Figure 2 handles case a) above with all $\beta_\ell = 0$ except for the intercept. The left panel shows the MSE for all β_ℓ 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 intercept really is $o(N^{-1/2})$ for default glmer, then we are unable to explain that.

Figure 3 shows case b) with nonzero β_ℓ. 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.

5.3. About naive logistic regression

This section provides an explanation for some of the biases we see for naive logistic regression. It may seem odd to consider plain logistic regression in this random effects problem. We suspect however that naive logistic regression is widely used, because logistic regression is a very popular algorithm in electronic

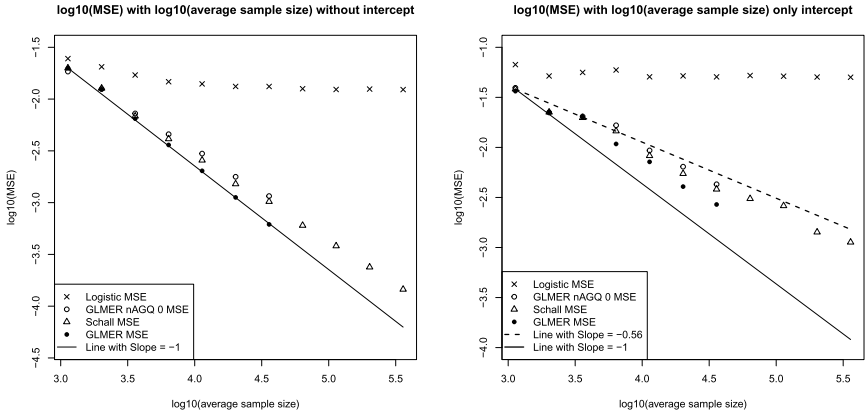

FIG 3. *MSE of $\hat{\beta}$ vs average sample size at $(\rho, \kappa) = (0.56, 0.56)$ for trending β. The left panel plots the MSE of non intercept β. The right panel plots the MSE for the intercept with true intercept being -2. MSE for glmer and backfitting seems to scale slightly worse than $O(1/N)$ whereas logistic regression appears to be inconsistent. The MSE for the intercept scales at a much slower rate than $O(1/N)$.*

commerce where random effects models have hitherto been computationally infeasible. The following analysis preceded the simulations and it influenced our choice of simulation settings.

Suppose we have a mixed effects model with one random effect (no crossing). The responses Y_{ij} satisfy $\Pr(Y_{ij} = 1 \,|\, \boldsymbol{x}_{ij}, a_i) = \pi(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i)$ for $i = 1, \ldots, R$. Since the columns are not linked we let $j = 1, \ldots, N_{i\bullet}$ The random effects are $a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The $\boldsymbol{x}_{ij}$ are fixed either by design or by running a conditional analysis. We will suppose at first that σ is 'small'.

If an analyst ignores the random effects and uses likelihood

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

then their log likelihood function is

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

Their naive score function is

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

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

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

For small σ, we anticipate that small a will dominate, so we make a Taylor approximation

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

and using this approximation

$$\mathbb{E}\biggl(\frac{\partial\tilde{\ell}}{\partial\beta}\biggr) \approx \frac{\sigma^2}{2}\sum_{i=1}^{R}\sum_{j=1}^{N_{i\bullet}}\pi''(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta)\boldsymbol{x}_{ij}. \tag{38}$$

In light of equation (38), the analyst needs each variable $x_{ij\ell}$ to be nearly orthogonal to $\pi''(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta)$ with the true unknown β. This is virtually impossible to arrange.

In one special case we could have β equal to zero apart from a negative intercept component. Then $\pi''(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta)$ is a positive constant and for any centered variables $x_{ij\ell}$ the ℓ'th component of (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.

6. Stitch Fix data example

Stitch Fix is an online personal styling service. One of their business models involves sending customers a sample of clothing items. The customer may keep and purchase any of those items and return the others. They have provided us with some of their client ratings data. That data 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. The binary response of interest 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 $\max_i N_{i\bullet}/N \doteq 1.24\times10^{-5}$ 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

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

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 β 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 A.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 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 β 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 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.

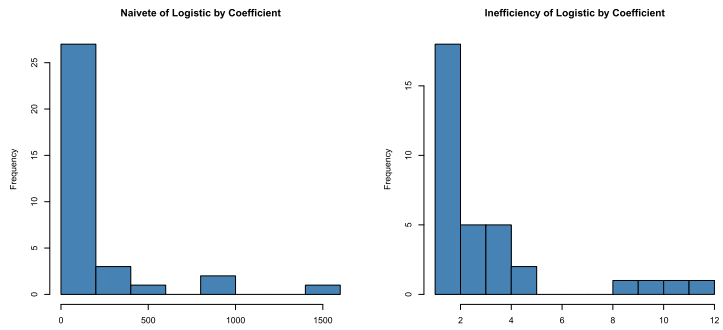

FIG 4. *The left panel is a histogram of naivete of logistic regression quantified by* $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR},\ell})/\widehat{\mathrm{cov}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR},\ell})$ *for coefficients* β_ℓ *in the Stitch Fix data. The right panel is the inefficiency* $\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR},\ell})/\widehat{\mathrm{cov}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GLMM},\ell})$.

Figure 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. This trend was also seen for the naive and inefficient ordinary least squares in Ghosh et al. (2022).

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, leading 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.

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

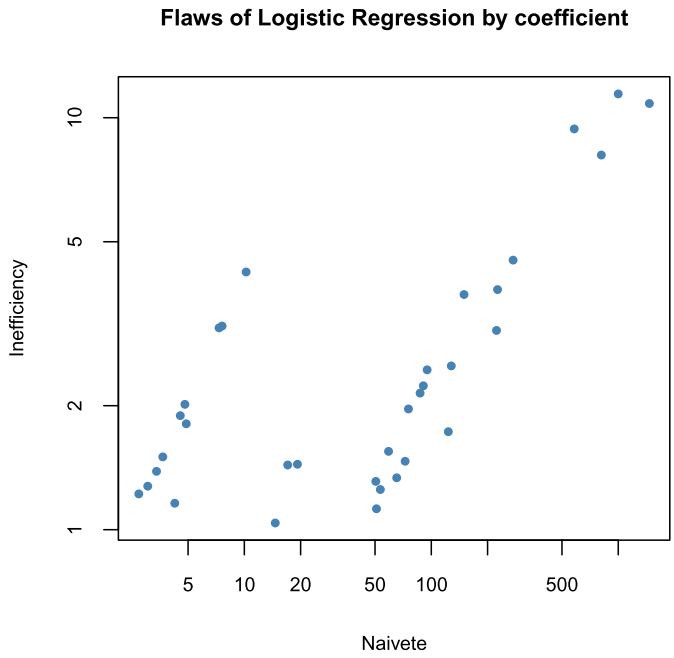

FIG 5. *Inefficiency vs naivete for logistic regression coefficients in the Stitch Fix data.*

is to find more explanations of this phenomenon and sufficient conditions to bound the iteration count.

Appendix A: Some proofs

A.1. Proof of Lemma 3.1

The first derivative of T^{-1} with respect to η 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

$$T(\eta)^{-1} = T(0)^{-1} + T(0)^{-1/2} \sum_{k \geqslant 1} \begin{pmatrix} 0 & -\eta B_* \\ -\eta B_*^{\mathsf{T}} & 0 \end{pmatrix}^k T(0)^{-1/2}. \tag{39}$$

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

$$\sum_{k \geqslant 1} \eta^{2k} \Bigl(\operatorname{tr}\bigl(A^{-1}(B_* B_*^{\mathsf{T}})^k\bigr) + \operatorname{tr}\bigl(C^{-1}(B_*^{\mathsf{T}} B_*)^k\bigr) \Bigr).$$

A.2. Proof of Lemma 3.2

First ρ^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 ρ^2 that is the square of the claimed bound for ρ.

A.3. Proof of Theorem 3.3

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 (7) and (8). Now $\sum_j Z_{ij} = N_{i\bullet} \leqslant (\Upsilon+\epsilon)S^{1-\rho}$ by equation (7). Therefore with probability approaching one,

$$\operatorname{tr}(B_*B_*^{\mathsf{T}}) \leqslant \bar{\omega}^2(1-\epsilon)^{-2}(\Upsilon+\epsilon)S^{2\psi+\rho+\kappa-1} = o(R)$$

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 δ then

$$\operatorname{tr}(B_*B_*^{\mathsf{T}}) > \delta S^{\alpha\rho}.$$

That cannot hold for large S if

$$\alpha > \underline{\alpha} \equiv 1 + \frac{2\psi+\kappa-1}{\rho} < 1$$

for small enough ψ. Therefore any $\alpha \in (\underline{\alpha}, 1)$ satisfies the condition in (32).

A.4. Proof of Theorem 3.4

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

$$\text{Err} = \operatorname{tr}\Big(A^{-1}\sum_{k\geqslant 1}\sum_{i=1}^{R}\lambda_i^k v_i v_i^{\mathsf{T}}\Big) = \sum_{k\geqslant 1}\sum_{i=1}^{R}\lambda_i^k \operatorname{tr}(v_i^{\mathsf{T}}A^{-1}v_i).$$

Now with probability tending to 1 as $S \to \infty$,

$$\operatorname{tr}(v_i^{\mathsf{T}}A^{-1}v_i) \leqslant \lambda_{\max}(A^{-1}) = \max_s (W_{s\bullet}+\sigma_A^{-2})^{-1} \leqslant S^{\psi}(1-\epsilon)^{-1}S^{\rho-1}$$

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 and suppose that for $\delta > 0$ that fewer than R^α of the λ_i are larger than δ. Under this event which has probability tending to one,

$$\mathrm{Err} \leqslant S^{\psi+\rho-1}(1-\epsilon)^{-1} \left(\frac{\delta \lambda_1 R^\alpha}{1-\lambda_1} + \frac{R}{1-\delta} \right).$$

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 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

$$\begin{aligned} \frac{1}{(1/\lambda_1 - 1)} &\leqslant \frac{1}{a+b+ab} \leqslant \min\left(\frac{1}{a}, \frac{1}{b}\right) \\ &= \min(S^{1-\rho}\sigma_A^2, S^{1-\kappa}\sigma_B^2)\bar{\omega}(\Upsilon+\epsilon). \end{aligned}$$

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 $\psi > 0$.

A.5. Proof of Lemma 4.2

Our iterative algorithm is designed to minimize

$$\mathrm{pl}(\beta, \boldsymbol{a}, \boldsymbol{b}) = \sum_{ij} Z_{ij}\widehat{W}_{ij}(z_{ij} - \boldsymbol{x}_{ij}^{\mathsf{T}}\beta - a_i - b_j)^2 + \hat{\sigma}_A^{-2}\|\boldsymbol{a}\|^2 + \hat{\sigma}_B^{-2}\|\boldsymbol{b}\|^2,$$

where all $\widehat{W}_{ij} > 0$, $\min(\hat{\sigma}_A^{-2}, \hat{\sigma}_B^{-2}) > 0$, the matrix $\mathcal{X}$ with rows $\boldsymbol{x}_{ij}^{\mathsf{T}}$ 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 β 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 β 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 β 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

$$\mathrm{pl}(\beta^{(k)}, \boldsymbol{a}^{(k)}; \boldsymbol{b}^{(k)}) - \mathrm{pl}(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)}; \boldsymbol{b}^{(k)})$$

$$= \frac{1}{2}\begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})} \\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix}^{\mathsf{T}} H(\beta^{(k+\frac{1}{2})}, \boldsymbol{a}^{(k+1)}; \boldsymbol{b}^{(k)}) \begin{pmatrix}\beta^{(k)} - \beta^{(k+\frac{1}{2})} \\ \boldsymbol{a}^{(k)} - \boldsymbol{a}^{(k+1)}\end{pmatrix}$$

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.

A.6. Results of the binary regression in Section 6

Table 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 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).

Acknowledgments

We thank Stitch Fix for sharing some data with us and are especially grateful to Bradley Klingenberg and Sven Schmit. This work was supported in part by a National Science Foundation BIGDATA grant IIS-1837931.

TABLE 1
Stitch Fix Binary Regression Results

	$\hat{\beta}_{\mathrm{LR}}$	$\widehat{\mathrm{SE}}_{\mathrm{LR}}(\hat{\beta}_{\mathrm{LR}})$	$\widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{LR}})$	$\hat{\beta}_{\mathrm{GL}}$	$\widehat{\mathrm{SE}}_{\mathrm{GLMM}}(\hat{\beta}_{\mathrm{GL}})$
Intercept	0.677*	0.022	0.048	0.584*	0.036
𝕀{CFP missing}	0.099*	0.015	0.025	0.127*	0.023
𝕀{CFP loose}	0.016*	0.002	0.006	0.208*	0.004
𝕀{CFP oversize}	−0.094*	0.010	0.019	−0.048*	0.015
𝕀{CFP straight}	0.125*	0.003	0.005	0.154*	0.004
𝕀{CFP tight}	−0.229*	0.011	0.130	−0.250*	0.009
𝕀{client edgy}	−0.042*	0.002	0.005	−0.044*	0.003
𝕀{client boho}	0.147*	0.002	0.006	0.194*	0.003
𝕀{client chest size}	−0.008*	0.001	0.001	−0.011*	0.001
𝕀{client dress size}	−0.002*	0.000	0.001	−0.000	0.001
𝕀{IFP missing}	−0.281*	0.024	0.188	−0.079	0.151
𝕀{IFP fitted}	0.061*	0.004	0.089	0.277*	0.029
𝕀{IFP loose}	0.209*	0.003	0.082	0.038	0.029
𝕀{IFP oversize}	−0.066*	0.016	0.135	−0.020	0.111
𝕀{IFP tight}	0.186	0.010	0.431	0.251	0.359
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	0.046	0.509	0.035
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

In the table, $\hat{\beta}_{\mathrm{GL}}$ denotes the GLMM estimates of β obtained by clubbed backfitting. The abbreviations CFP and IFP are used for client fit profile and item fit profile respectively.

References

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language 59*(4), 390–412.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. Stevens, and J. S. White (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution 24*(3), 127–135.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized

linear models. *Journal of the American Statistical Association 88*(421), 9–25. MR1394064

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics 17*(2), 453–510. MR0994249

Dick, J., F. Y. Kuo, and I. H. Sloan (2013). High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica 22*, 133–288. MR3038697

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics 28*(1), 51–60. MR0824731

Gao, K. (2017). *Scalable Estimation and Inference for Massive Linear Mixed Models with Crossed Random Effects*. Ph.D. thesis, Stanford University. MR4257222

Gao, K. and A. B. Owen (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects designs. *Electronic Journal of Statistics 11*(1), 1235–1296. MR3635913

Gao, K. and A. B. Owen (2020). Estimation and inference for very large linear mixed effects models. *Statistica Sinica 30*, 1741–1771. MR4260743

Ghosh, S., T. Hastie, and A. B. Owen (2022). Backfitting for large scale crossed random effects regressions. *Annals of Statistics 50*(1), 560–583. MR4382028

Ghosh, S. and C. Zhong (2021). Convergence rate of a collapsed gibbs sampler for crossed random effects models. Technical report, arXiv:2109.02849.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338. MR0451550

Heiss, F. and V. Winschel (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics 144*(1), 62–80. MR2439922

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics 9*(2), 226–252. MR0055650

Kuo, F. Y., W. T. M. Dunsmuir, I. H. Sloan, M. P. Wand, and R. S. Womersley (2008). Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability 10*(2), 239–275. MR2399683

Laird, N. M. (1978, December). Empirical bayes methods for two-way contingency tables. *Biometrika 65*(3), 581–590. MR0521825

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological 37*, 23–37. MR0383628

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). Boca Raton, FL: Chapman and Hall. MR3223057

McCulloch, C. E. and J. M. Neuhaus (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics 67*(1), 270–279. MR2898839

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: SIAM. MR1172997

Papaspiliopoulos, O., G. O. Roberts, and G. Zanella (2020). Scalable inference for crossed random effects models. *Biometrika 107*(1), 25–40. MR4064138

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Searle, S. R., G. Casella, and C. E. McCulloch (2006). *Variance Components*. John Wiley & Sons. MR2298115

Stiratelli, R., N. Laird, and J. H. Ware (1984). Random-effects models for serial observations with binary response. *Biometrics 40*, 961–971.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42. MR2796852