The Annals of Statistics
2022, Vol. 50, No. 1, 560–583
https://doi.org/10.1214/21-AOS2121

BACKFITTING FOR LARGE SCALE CROSSED RANDOM EFFECTS REGRESSIONS

BY SWARNADIP GHOSH[*], TREVOR HASTIE[†] AND ART B. OWEN[‡]

Department of Statistics, Stanford University, [*]raswa281@stanford.edu; [†]hastie@stanford.edu; [‡]owen@stanford.edu

Regression models with crossed random effect errors can be very expensive to compute. The cost of both generalized least squares and Gibbs sampling can easily grow as $N^{3/2}$ (or worse) for N observations. Papaspiliopoulos, Roberts and Zanella (*Biometrika* **107** (2020) 25–40) present a collapsed Gibbs sampler that costs $O(N)$, but under an extremely stringent sampling model. We propose a backfitting algorithm to compute a generalized least squares estimate and prove that it costs $O(N)$. A critical part of the proof is in ensuring that the number of iterations required is $O(1)$, which follows from keeping a certain matrix norm below $1-\delta$ for some $\delta > 0$. Our conditions are greatly relaxed compared to those for the collapsed Gibbs sampler, though still strict. Empirically, the backfitting algorithm has a norm below $1-\delta$ under conditions that are less strict than those in our assumptions. We illustrate the new algorithm on a ratings data set from Stitch Fix.

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

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

The crossed random effects model we consider has

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

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

The GLS estimate $\hat{\beta}_{\text{GLS}}$ for crossed random effects can be efficiently estimated if all $R \times C$ values are available. Our motivating examples involve ratings data where R people rate C items and then it is usual that the data are very unbalanced with a haphazard observational pattern in which only $N \ll R \times C$ of the (x_{ij}, Y_{ij}) pairs are observed. The crossed random effects setting is significantly more difficult than a hierarchical model with just $a_i + e_{ij}$ but

Received August 2020; revised July 2021.
MSC2020 subject classifications. Primary 62J05; secondary 65C60.
Key words and phrases. Collapsed Gibbs sampler, mixed effect models, generalized least squares.

no b_j term. Then the observations for index j are "nested within" those for each level of index i. The result is that the covariance matrix of all observed Y_{ij} values has a block diagonal structure allowing GLS to be computed in $O(N)$ time.

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

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

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

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

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

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

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

The rest of this paper is organized as follows. Section 2 introduces our notation and assumptions for missing data. Section 3 presents the backfitting algorithm from Buja, Hastie and Tibshirani (1989). That algorithm was defined for smoothers, but we are able to cast the estimation of random effect parameters as a special kind of smoother. Section 4 proves our result about backfitting being convergent with a probability tending to one as the problem size increases. Section 5 shows numerical measures of the matrix norm of the backfitting operator. It remains bounded below and away from one under more conditions than our theory shows. We find that even one iteration of the lmer function in lme4 package Bates et al. (2015) has a cost that grows like $N^{3/2}$ in one setting and like $N^{2.1}$ in another, sparser one. The backfitting algorithm has cost $O(N)$ in both of these cases. Section 6 illustrates our GLS algorithm on some data provided to us by Stitch Fix. These are customer ratings of items of clothing on a ten-point scale. Section 7 has a discussion of these results. An Appendix contains some regression output for the Stitch Fix data.

2. Missingness. We adopt the notation from Gao and Owen (2020). We let $Z_{ij} \in \{0, 1\}$ take the value 1 if (x_{ij}, Y_{ij}) is observed and 0 otherwise, for $i = 1, \dots, R$ and $j = 1, \dots, C$. In many of the contexts, we consider, the missingness is not at random and is potentially informative. Handling such problems is outside the scope of this paper, apart from a brief discussion in Section 7. It is already a sufficient challenge to work without informative missingness.

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

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

In our regression models, we treat Z_{ij} as nonrandom. We are conditioning on the actual pattern of observations in our data. When we study the rate at which our backfitting algorithm converges, we consider Z_{ij} drawn at random. That is, the analyst is solving a GLS conditionally on the pattern of observations and missingness, while we study the convergence rates that analyst will see for data drawn from a missingness mechanism defined in Section 4.2.

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

$$\hat{\beta}_{\mathrm{OLS}} = (\mathcal{X}^{\mathsf{T}}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\mathcal{Y}. \tag{2}$$

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

$$\hat{\beta}_{\mathrm{GLS}} = (\mathcal{X}^{\mathsf{T}}\mathcal{V}^{-1}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\mathcal{V}^{-1}\mathcal{Y}, \tag{3}$$

where $\mathcal{V} \in \mathbb{R}^{N\times N}$ contains all of the $\operatorname{cov}(Y_{ij}, Y_{rs})$ in an ordering compatible with $\mathcal{X}$ and $\mathcal{Y}$. A naive algorithm costs $O(N^3)$ to solve for $\hat\beta_{\mathrm{GLS}}$. It can actually be solved through a Cholesky decomposition of an $(R+C)\times(R+C)$ matrix (Searle, Casella and McCulloch (1992)). That has cost $O(R^3+C^3)$. Now $N \le RC$, with equality only for completely observed data. Therefore, $\max(R,C) \ge \sqrt{N}$, and so $R^3+C^3 \ge N^{3/2}$. When the data are sparsely enough observed it is possible that $\min(R,C)$ grows more rapidly than $N^{1/2}$. In a numerical example, in Section 5 we have $\min(R,C)$ growing like $N^{0.70}$. In a hierarchical model, with a_i but no b_j we would find $\mathcal{V}$ to be block diagonal and then $\hat\beta_{\mathrm{GLS}}$ could be computed in $O(N)$ work.

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

We can write our crossed effects model as

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{Z}_A\boldsymbol{a} + \mathcal{Z}_B\boldsymbol{b} + \boldsymbol{e} \tag{4}$$

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

$$\mathcal{V} = \mathcal{Z}_A\mathcal{Z}_A^{\mathsf{T}}\sigma_A^2 + \mathcal{Z}_B\mathcal{Z}_B^{\mathsf{T}}\sigma_B^2 + I_N\sigma_E^2, \tag{5}$$

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

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

$$\hat\beta_{\mathrm{GLS}} = (\mathcal{X}^{\mathsf{T}}\mathcal{U})^{-1}(\mathcal{Y}^{\mathsf{T}}\mathcal{U})^{\mathsf{T}} \tag{6}$$

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

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

3. Backfitting algorithms. Our first goal is to develop computationally efficient ways to solve the GLS problem (6) for the linear mixed model (4). We use the backfitting algorithm that Hastie and Tibshirani (1990) and Buja, Hastie and Tibshirani (1989) use to fit additive models. We write $\mathcal{V}$ in (5) as $\sigma_E^2(\mathcal{Z}_A\mathcal{Z}_A^{\mathsf{T}}/\lambda_A + \mathcal{Z}_B\mathcal{Z}_B^{\mathsf{T}}/\lambda_B + I_N)$ with $\lambda_A = \sigma_E^2/\sigma_A^2$ and $\lambda_B = \sigma_E^2/\sigma_B^2$, and define $\mathcal{W} = \sigma_E^2\mathcal{V}^{-1}$. Then the GLS estimate of β is

$$\hat\beta_{\mathrm{GLS}} = \arg\min_\beta (\mathcal{Y}-\mathcal{X}\beta)^{\mathsf{T}}\mathcal{W}(\mathcal{Y}-\mathcal{X}\beta) = (\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{Y} \tag{7}$$

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

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

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \|\mathcal{Y}-\mathcal{X}\beta-\mathcal{Z}_A\boldsymbol{a}-\mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{8}$$

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

3.1. *One factor.* For a single factor, we simply drop the $\mathcal{Z}_B\boldsymbol{b}$ term from (4) to get

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

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

$$\min_{\beta,\boldsymbol{a}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a}\|^2 + \lambda_A\|\boldsymbol{a}\|^2. \tag{9}$$

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

The normal equations from (9) yield

$$\mathbf{0} = \mathcal{X}^\mathsf{T}(\mathcal{Y} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A\hat{\boldsymbol{a}}) \quad \text{and} \tag{10}$$

$$\mathbf{0} = \mathcal{Z}_A^\mathsf{T}(\mathcal{Y} - \mathcal{X}\hat{\beta} - \mathcal{Z}_A\hat{\boldsymbol{a}}) - \lambda_A\hat{\boldsymbol{a}}. \tag{11}$$

Solving (11) for $\hat{\boldsymbol{a}}$ and multiplying the solution by $\mathcal{Z}_A$ yields

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

for an $N \times N$ ridge regression "smoother matrix" $\mathcal{S}_A$. As we explain below, this smoother matrix implements shrunken within-group means. Then substituting $\mathcal{Z}_A\hat{\boldsymbol{a}}$ into equation (10) yields

$$\hat{\beta} = \big(\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_A)\mathcal{X}\big)^{-1}\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_A)\mathcal{Y}. \tag{12}$$

Using the Sherman–Morrison–Woodbury (SMW) identity, one can show that $\mathcal{W} = I_N - \mathcal{S}_A$ and hence $\hat{\beta}$ above equals $\hat{\beta}_{\mathrm{GLS}}$ from (7). This is not in itself a new discovery; see, for example, Robinson (1991) or Hastie and Tibshirani (1990), Section 5.3.3.

To compute the solution in (12), we need to compute $\mathcal{S}_A\mathcal{Y}$ and $\mathcal{S}_A\mathcal{X}$. The heart of the computation in $\mathcal{S}_A\mathcal{Y}$ is $(\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A + \lambda_A I_R)^{-1}\mathcal{Z}_A^\mathsf{T}\mathcal{Y}$. But $\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A = \operatorname{diag}(N_{1\bullet}, N_{2\bullet}, \ldots, N_{R\bullet})$ and we see that all we are doing is computing an R-vector of shrunken means of the elements of $\mathcal{Y}$ at each level of the factor A; the ith element is $\sum_j Z_{ij}Y_{ij}/(N_{i\bullet} + \lambda_A)$. This involves a single pass through the N elements of Y, accumulating the sums into R registers, followed by an elementwise scaling of the R components. Then premultiplication by $\mathcal{Z}_A$ simply puts these R shrunken means back into an N-vector in the appropriate positions. The total cost is $O(N)$. Likewise $\mathcal{S}_A\mathcal{X}$ does the same separately for each of the columns of $\mathcal{X}$. Hence the entire computational cost for (12) is $O(Np^2)$, the same order as regression on $\mathcal{X}$.

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

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

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

3.2. *Two factors.* With two factors, we face the problem of incompatible block diagonal matrices discussed in Section 2. Define $\mathcal{Z}_G = (\mathcal{Z}_A : \mathcal{Z}_B)$ ($R + C$ columns), $\mathcal{D}_\lambda = \operatorname{diag}(\lambda_A I_R, \lambda_B I_C)$, and $\boldsymbol{g}^\mathsf{T} = (\boldsymbol{a}^\mathsf{T}, \boldsymbol{b}^\mathsf{T})$. Then solving (8) is equivalent to

$$\min_{\beta,\boldsymbol{g}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_G\boldsymbol{g}\|^2 + \boldsymbol{g}^\mathsf{T}\mathcal{D}_\lambda\boldsymbol{g}. \tag{13}$$

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

$$\hat{\beta} = H_{\mathrm{GLS}}\mathcal{Y} \quad \text{for } H_{\mathrm{GLS}} = \big(\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_G)\mathcal{X}\big)^{-1}\mathcal{X}^\mathsf{T}(I_N - \mathcal{S}_G), \tag{14}$$

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

$$\mathcal{S}_G = \mathcal{Z}_G\big(\mathcal{Z}_G^{\mathsf{T}}\mathcal{Z}_G + \mathcal{D}_\lambda\big)^{-1}\mathcal{Z}_G^{\mathsf{T}}. \tag{15}$$

We can again use SMW to show that $I_N - \mathcal{S}_G = \mathcal{W}$ and hence the solution $\hat\beta = \hat\beta_{\mathrm{GLS}}$ in (7). But in applying $\mathcal{S}_G$ we do not enjoy the computational simplifications that occurred in the one factor case, because

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

where $Z \in \{0,1\}^{R\times C}$ is the observation matrix which has no special structure. Therefore, we need to invert an $(R+C)\times(R+C)$ matrix to apply $\mathcal{S}_G$, and hence to solve (14), at a cost of at least $O(N^{3/2})$ (see Section 2).

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

$$\min_{\boldsymbol{a},\boldsymbol{b}} \|\mathcal{R} - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{16}$$

Using $\mathcal{S}_G$ defined at (15) in terms of the indicator variables $\mathcal{Z}_G \in \{0,1\}^{N\times(R+C)}$, it is clear that the fitted values are given by $\widehat{\mathcal{R}} = \mathcal{S}_G\mathcal{R}$. Solving (16) would result in two blocks of estimating equations similar to equations (10) and (11). These can be written

$$\begin{aligned} \mathcal{Z}_A\hat{\boldsymbol{a}} &= \mathcal{S}_A(\mathcal{R} - \mathcal{Z}_B\hat{\boldsymbol{b}}) \quad \text{and} \\ \mathcal{Z}_B\hat{\boldsymbol{b}} &= \mathcal{S}_B(\mathcal{R} - \mathcal{Z}_A\hat{\boldsymbol{a}}), \end{aligned} \tag{17}$$

where $\mathcal{S}_A = \mathcal{Z}_A(\mathcal{Z}_A^{\mathsf{T}}\mathcal{Z}_A + \lambda_A I_R)^{-1}\mathcal{Z}_A^{\mathsf{T}}$ is again the ridge regression smoothing matrix for row effects and similarly $\mathcal{S}_B = \mathcal{Z}_B(\mathcal{Z}_B^{\mathsf{T}}\mathcal{Z}_B + \lambda_B I_C)^{-1}\mathcal{Z}_B^{\mathsf{T}}$ the smoothing matrix for column effects. We solve these equations iteratively by block coordinate descent, also known as backfitting. The iterations converge to the solution of (16) (Buja, Hastie and Tibshirani (1989), Hastie and Tibshirani (1990)).

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

Here, the simplifications we enjoyed in the one-factor case once again apply. Each step applies its operator to a vector (the terms in parentheses on the right-hand side in (17)). For both $\mathcal{S}_A$ and $\mathcal{S}_B$, these are simply the shrunken-mean operations described for the one-factor case, separately for factor A and B each time. As before, we do not need to actually construct $\mathcal{Z}_B$, but simply use a factor f_B that records the level of factor B for each of the N observations.

The above description holds for a generic response $\mathcal{R}$; we apply that algorithm (in parallel) to $\mathcal{Y}$ and each column of $\mathcal{X}$ to obtain the quantities $\mathcal{S}_G\mathcal{X}$ and $\mathcal{S}_G\mathcal{Y}$ that we need to compute $H_{\mathrm{GLS}}\mathcal{Y}$ in (14). Now solving (14) is $O(Np^2)$ plus a negligible $O(p^3)$ cost. These computations deliver $\hat\beta_{\mathrm{GLS}}$; if the BLUP estimates $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are also required, the same algorithm can be applied to the response $\mathcal{Y} - \mathcal{X}\hat\beta_{\mathrm{GLS}}$, retaining the $\boldsymbol{a}$ and $\boldsymbol{b}$ at the final iteration. We can also write

$$\mathrm{cov}(\hat\beta_{\mathrm{GLS}}) = \sigma_E^2\big(\mathcal{X}^{\mathsf{T}}(I_N - \mathcal{S}_G)\mathcal{X}\big)^{-1}. \tag{18}$$

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

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

$$\min_{\beta,\boldsymbol{a},\boldsymbol{b}} \|\mathcal{Y} - \mathcal{X}\beta - \mathcal{Z}_A\boldsymbol{a} - \mathcal{Z}_B\boldsymbol{b}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 + \lambda_B\|\boldsymbol{b}\|^2. \tag{19}$$

LEMMA 3.1. *If $\mathcal{X}$ in model (19) includes a column of ones (intercept), and $\lambda_A > 0$ and $\lambda_B > 0$, then the solutions for $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy $\sum_{i=1}^R a_i = 0$ and $\sum_{j=1}^C b_j = 0$.*

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

$$\min_{\beta,\boldsymbol{a}} \|\mathcal{Y} - \mathbf{1}\beta - \mathcal{Z}_A\boldsymbol{a}\|^2 + \lambda_A\|\boldsymbol{a}\|^2. \tag{20}$$

Notice that for any candidate solution $(\beta, \{a_i\}_1^R)$, the alternative solution $(\beta + c, \{a_i - c\}_1^R)$ leaves the loss part of (20) unchanged, since the row sums of $\mathcal{Z}_A$ are all one. Hence, if $\lambda_A > 0$, we would always improve $\boldsymbol{a}$ by picking c to minimize the penalty term $\sum_{i=1}^R (a_i - c)^2$, or $c = (1/R)\sum_{i=1}^R a_i$. □

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

THEOREM 3.2. *Consider the generic optimization problem*

$$\min_{\boldsymbol{a}} \|\mathcal{R} - \mathcal{Z}_A\boldsymbol{a}\|^2 + \lambda_A\|\boldsymbol{a}\|^2 \quad \textit{subject to } \sum_{i=1}^R a_i = 0. \tag{21}$$

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

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

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

$$\hat{a}_i = \frac{\mathcal{R}_i^+ - \sum_r w_r\mathcal{R}_r^+}{N_{i\bullet} + \lambda_A}, \quad i = 1, \ldots, R. \tag{22}$$

Moreover, the fit is given by

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

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

The computations are a simple modification of the noncentered case.

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

$$\min_{\tilde{\boldsymbol{\gamma}}} \|\mathcal{R} - \tilde{\mathcal{G}}\tilde{\boldsymbol{\gamma}}\|^2 + \lambda_A\|\tilde{\boldsymbol{\gamma}}\|^2 \quad \text{subject to } \tilde{\gamma}_1 = 0. \tag{23}$$

To solve (23), we simply drop the first column of $\tilde{\mathcal{G}}$. Let $\mathcal{G} = \mathcal{Z}_A Q$ where Q is the matrix M omitting the first column, and $\boldsymbol{\gamma}$ the corresponding subvector of $\tilde{\boldsymbol{\gamma}}$ having $R-1$ components. We now solve

$$\min_{\tilde{\boldsymbol{\gamma}}} \|\mathcal{R} - \mathcal{G}\boldsymbol{\gamma}\|^2 + \lambda_A \|\tilde{\boldsymbol{\gamma}}\|^2 \tag{24}$$

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

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

$$\begin{aligned}\mathcal{G}\hat{\gamma} &= \mathcal{Z}_A Q\big(Q^\mathsf{T}\mathcal{Z}_A^\mathsf{T}\mathcal{Z}_A Q + \lambda_A I_{R-1}\big)^{-1} Q^\mathsf{T}\mathcal{Z}_A^\mathsf{T}\mathcal{R} \\ &= \mathcal{Z}_A Q\big(Q^\mathsf{T} D Q + \lambda_A I_{R-1}\big)^{-1} Q^\mathsf{T}\mathcal{R}^+ \\ &= \mathcal{Z}_A \hat{\boldsymbol{a}},\end{aligned} \tag{25}$$

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

$$\tilde{H} = (D + \lambda_A I_R)^{\frac{1}{2}} H (D + \lambda_A I_R)^{\frac{1}{2}} = \tilde{Q}\big(\tilde{Q}^\mathsf{T}\tilde{Q}\big)^{-1}\tilde{Q}^\mathsf{T}. \tag{26}$$

Now (26) is a projection matrix of $\mathbb{R}^R$ onto a $R-1$ dimensional subspace. Let $\tilde{q} = (D + \lambda_A I_R)^{-\frac{1}{2}}\mathbf{1}$. Then $\tilde{q}^\mathsf{T}\tilde{Q} = \mathbf{0}$, and so

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

Unraveling this expression, we get

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

With $\hat{\boldsymbol{a}} = H\mathcal{R}^+$ in (25), this gives the expressions for each $\hat{a}_i$ in (22). Finally, $\tilde{\mathcal{S}}_A = \mathcal{Z}_A H \mathcal{Z}_A^\mathsf{T}$ is symmetric. □

3.4. *Covariance matrix for $\hat{\beta}_{\mathrm{GLS}}$ with centered operators.* In Section 3.2, we saw in (18) that we get a simple expression for $\operatorname{cov}(\hat{\beta}_{\mathrm{GLS}})$. This simplicity relies on the fact that $I_N - \mathcal{S}_G = \mathcal{W} = \sigma_E^2\mathcal{V}^{-1}$, and the usual cancellation occurs when we use the sandwich formula to compute this covariance. When we backfit with our centered smoothers, we get a modified residual operator $I_N - \widetilde{\mathcal{S}}_G$ such that the analog of (14) still gives us the required coefficient estimate:

$$\hat{\beta}_{\mathrm{GLS}} = \big(\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{X}\big)^{-1}\mathcal{X}^\mathsf{T}(I_N - \widetilde{\mathcal{S}}_G)\mathcal{Y}. \tag{27}$$

However, $I_N - \widetilde{\mathcal{S}}_G \neq \sigma_E^2\mathcal{V}^{-1}$, and so now we need to resort to the sandwich formula $\operatorname{cov}(\hat{\beta}_{\mathrm{GLS}}) = H_{\mathrm{GLS}}\mathcal{V}H_{\mathrm{GLS}}^\mathsf{T}$ with H_{GLS} from (14). Expanding this, we find that $\operatorname{cov}(\hat{\beta}_{\mathrm{GLS}})$ equals

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

While this expression might appear daunting, the computations are simple. Note first that while $\hat{\beta}_{\mathrm{GLS}}$ can be computed via $\tilde{\mathcal{S}}_G\mathcal{X}$ and $\tilde{\mathcal{S}}_G\mathcal{Y}$ this expression for $\operatorname{cov}(\hat{\beta}_{\mathrm{GLS}})$ also involves $\mathcal{X}^\mathsf{T}\tilde{\mathcal{S}}_G$. When we use the centered operator from Theorem 3.2 we get a symmetric matrix $\tilde{\mathcal{S}}_G$. Let $\widetilde{\mathcal{X}} = (I_N - \widetilde{\mathcal{S}}_G)\mathcal{X}$, the residual matrix after backfitting each column of $\mathcal{X}$ using these centered operators. Then because $\widetilde{\mathcal{S}}_G$ is symmetric, we have

$$\begin{aligned}\hat{\beta}_{\mathrm{GLS}} &= \big(\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}}\big)^{-1}\widetilde{\mathcal{X}}^\mathsf{T}\mathcal{Y} \quad \text{and} \\ \operatorname{cov}(\hat{\beta}_{\mathrm{GLS}}) &= \big(\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}}\big)^{-1}\widetilde{\mathcal{X}}^\mathsf{T}\cdot\mathcal{V}\cdot\widetilde{\mathcal{X}}\big(\mathcal{X}^\mathsf{T}\widetilde{\mathcal{X}}\big)^{-1}.\end{aligned} \tag{28}$$

Since $\mathcal{V} = \sigma_E^2(\mathcal{Z}_A\mathcal{Z}_A^\mathsf{T}/\lambda_A + \mathcal{Z}_B\mathcal{Z}_B^\mathsf{T}/\lambda_B + I_N)$ (two low-rank matrices plus the identity), we can compute $\mathcal{V}\cdot\widetilde{\mathcal{X}}$ very efficiently, and hence also the covariance matrix in (28). The entire algorithm is summarized in Section 6.3.

4. Convergence of the matrix norm. In this section, we prove a bound on the norm of the matrix that implements backfitting for our random effects $\boldsymbol{a}$ and $\boldsymbol{b}$ and show how this controls the number of iterations required. In our algorithm, backfitting is applied to $\mathcal{Y}$ as well as to each nonintercept column of $\mathcal{X}$ so we do not need to consider the updates for $\mathcal{X}\hat{\beta}$. It is useful to take account of intercept adjustments in backfitting, by the centerings described in Section 3 because the space spanned by $a_1, \ldots, a_R$ intersects the space spanned by $b_1, \ldots, b_C$ since both include an intercept column of ones.

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

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

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

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

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

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

Now for any $1 \le p \le \infty$, we have

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

The explicit formula

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

makes the matrix L_1 matrix norm very tractable theoretically and so that is the one we study. We look at this and some other measures numerically in Section 5.

4.1. *Updates.* Recall that $Z \in \{0,1\}^{R\times C}$ describes the pattern of observations. In a model with no intercept, centering the responses and then taking shrunken means as in (17) would yield these updates

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

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

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

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

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

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

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

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

$$M^{(1)}_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r \frac{Z_{rs}(Z_{rj} - N_{\bullet j}/R)}{N_{r\bullet} + \lambda_A} \tag{29}$$

after replacing a sum over i by an equivalent one over r.

In practice, we prefer to use the weighted centering from Section 3.3 to center the a_i because it provides a symmetric smoother $\tilde{\mathcal{S}}_G$ that supports computation of $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$. While it is more complicated to analyze it is easily computable and it satisfies the optimality condition in Theorem 3.2. The algorithm is for a generic response $\mathcal{R} \in \mathbb{R}^N$ such as $\mathcal{Y}$ or a column of $\mathcal{X}$. Let us illustrate it for the case $\mathcal{R} = \mathcal{Y}$. We begin with vector of N values $Y_{ij} - b_j$ and so $Y_i^+ = \sum_s Z_{is}(Y_{is} - b_s)$. Then $w_i = (N_{i\bullet} + \lambda_A)^{-1} / \sum_r (N_{r\bullet} + \lambda_A)^{-1}$ and the updated a_r is

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

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

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

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

$$M^{(2)}_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_r \frac{Z_{rj}}{N_{r\bullet} + \lambda_A} \left(Z_{rs} - \frac{\sum_i \frac{Z_{is}}{N_{i\bullet} + \lambda_A}}{\sum_i \frac{1}{N_{i\bullet} + \lambda_A}} \right). \tag{30}$$

Our preferred algorithm applies the optimal update from Theorem 3.2 to both $\boldsymbol{a}$ and $\boldsymbol{b}$ updates. With that choice we do not need to decide beforehand which random effects to center and which to leave uncentered to contain the intercept. We call the corresponding matrix $M^{(3)}$. Our theory below analyzes $\|M^{(1)}\|_1$ and $\|M^{(2)}\|_1$, which have simpler expressions than $\|M^{(3)}\|_1$.

Update $M^{(0)}$ uses symmetric smoothers for both A and B: they are both shrunken averages. The naive centering update $M^{(1)}$ uses a nonsymmetric smoother $\mathcal{Z}_A(I_R - \mathbf{1}_R \mathbf{1}_R^{\mathsf{T}})(\mathcal{Z}_A^{\mathsf{T}} \mathcal{Z}_A + \lambda_A I_R)^{-1} \mathcal{Z}_A^{\mathsf{T}}$ on the a_i with a symmetric smoother on b_j, and hence it does not generally produce a symmetric smoother needed for efficient computation of $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$. The update $M^{(2)}$ uses two symmetric smoothers, one optimal and one a simple shrunken mean. The update $M^{(3)}$ takes the optimal smoother for both A and B. Thus both $M^{(2)}$ and $M^{(3)}$ support efficient computation of $\widehat{\mathrm{cov}}(\hat{\beta}_{\mathrm{GLS}})$.

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

LEMMA 4.1. *If $X \sim \mathrm{Bin}(n, p)$, then for any $t \geq 0$,*

$$\Pr(X \geq np + t) \leq \exp(-2t^2/n) \quad \textit{and}$$
$$\Pr(X \leq np - t) \leq \exp(-2t^2/n).$$

PROOF. This follows from Hoeffding's theorem. □

LEMMA 4.2. *Let $X_i \sim \mathrm{Bin}(n, p)$ for $i = 1, \ldots, m$, not necessarily independent. Then, for any $t \geq 0$,*

$$\Pr\Big(\max_{1\leq i\leq m} X_i \geq np + t\Big) \leq m \exp(-2t^2/n) \quad \textit{and}$$
$$\Pr\Big(\min_{1\leq i\leq m} X_i \leq np - t\Big) \leq m \exp(-2t^2/n).$$

PROOF. This is from the union bound applied to Lemma 4.1. □

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

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

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

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

$$\frac{S}{RC} \leq p_{ij} \leq \Upsilon \frac{S}{RC} \quad \text{for } 1 \leq \Upsilon < \infty. \tag{31}$$

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

4.3. *Bounds for row and column size.* Letting $X \preccurlyeq Y$ mean that X is stochastically smaller than Y, we know that

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

By Lemma 4.1, if $t \geq 0$, then

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

Therefore, if $2\rho + \kappa < 2$, we find using using Lemma 4.2 that

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

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

$$\lim_{S\to\infty} \Pr\Big((1-\epsilon)S^{1-\rho} \le \min_i N_{i\bullet} \le \max_i N_{i\bullet} \le (\Upsilon+\epsilon)S^{1-\rho}\Big) = 1. \tag{32}$$

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

$$\lim_{S\to\infty} \Pr\Big((1-\epsilon)S^{1-\kappa} \le \min_j N_{\bullet j} \le \max_j N_{\bullet j} \le (\Upsilon+\epsilon)S^{1-\kappa}\Big) = 1. \tag{33}$$

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

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

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

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

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

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

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

4.5. *Coobservation.* Recall that the coobservation matrices are $Z^\mathsf{T}Z \in \{0, 1\}^{C\times C}$ and $ZZ^\mathsf{T} \in \{0, 1\}^{R\times R}$. If $s \ne j$, then

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

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

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

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

$$\Pr\Big(\max_s \max_{j\ne s} (Z^\mathsf{T}Z)_{sj} \ge (\Upsilon^2 + \epsilon)S^{2-\rho-2\kappa}\Big) \to 0 \quad \text{and}$$

$$\Pr\Big(\min_s \min_{j\ne s} (Z^\mathsf{T}Z)_{sj} \le (1-\epsilon)S^{2-\rho-2\kappa}\Big) \to 0,$$

for any $\epsilon > 0$.

4.6. *Asymptotic bounds for* $\|M\|_1$. Here, we prove upper bounds for $\|M^{(k)}\|_1$ for $k = 1, 2$ of equations (29) and (30), respectively. The bounds depend on Υ and there are values of $\Upsilon > 1$ for which these norms are bounded strictly below one, with probability tending to one.

THEOREM 4.3. *Let Z_{ij} follow the model from Section 4.2 with $\rho, \kappa \in (0,1)$, that satisfy $\rho+\kappa>1$, $2\rho+\kappa<2$ and $3\rho+4\kappa<4$. Then, for any $\epsilon>0$,*

$$\Pr\bigl(\|M^{(1)}\|_1 \le \Upsilon^2 - \Upsilon^{-2} + \epsilon\bigr) \to 1 \quad \textit{and} \tag{34}$$

$$\Pr\bigl(\|M^{(2)}\|_1 \le \Upsilon^2 - \Upsilon^{-2} + \epsilon\bigr) \to 1 \tag{35}$$

as $S \to \infty$.

PROOF. Without loss of generality, we assume that $\epsilon < 1$. We begin with (35). Let $M = M^{(2)}$. When $j \ne s$,

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

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

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

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

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

Similarly,

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

and so

$$\frac{\bar{Z}_{\bullet s}}{N_{\bullet j}+\lambda_B} \sum_r \frac{Z_{rj}}{N_{r\bullet}+\lambda_A} \in S^{-\kappa} \frac{[1-\epsilon, \Upsilon+\epsilon]^2}{[1-\epsilon, \Upsilon+\epsilon]^2} \subseteq \frac{1}{C}\biggl[\biggl(\frac{1-\epsilon}{\Upsilon+\epsilon}\biggr)^2, \biggl(\frac{\Upsilon+\epsilon}{1-\epsilon}\biggr)^2\biggr]. \tag{36}$$

Next, using bounds on the coobservation counts,

$$\frac{1}{N_{\bullet j}+\lambda_B} \sum_r \frac{Z_{rj}Z_{rs}}{N_{r\bullet}+\lambda_A} \in \frac{S^{\rho+\kappa-2}(Z^{\mathsf{T}}Z)_{sj}}{[1-\epsilon, \Upsilon+\epsilon]^2} \subseteq \frac{1}{C}\frac{[1-\epsilon, \Upsilon^2+\epsilon]}{[1-\epsilon, \Upsilon+\epsilon]^2}. \tag{37}$$

Combining (36) and (37),

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

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

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

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

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

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

This bound holds for all $s \in \{1, 2, \ldots, C\}$, establishing (35).

The proof of (34) is similar. The quantity $\bar{Z}_{\bullet s}$ is replaced by $(1/R)\sum_i Z_{is}/(N_{i\bullet}+\lambda_A)$. □

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

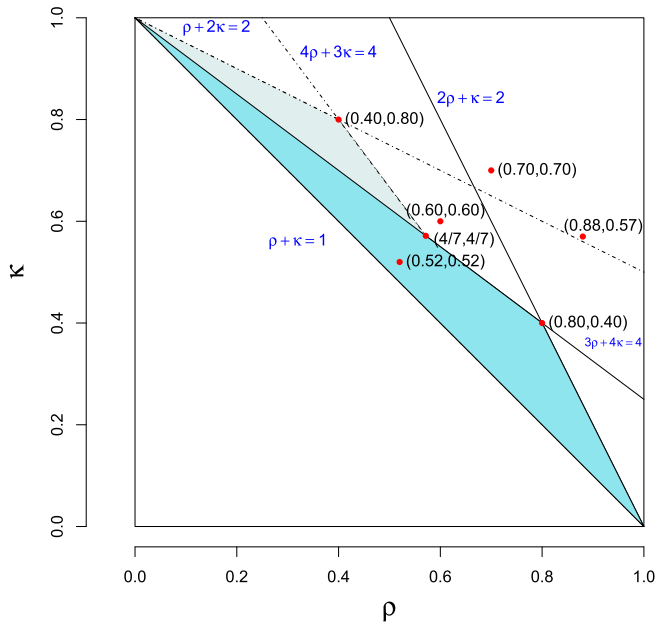


FIG. 1. *The large shaded triangle is the domain of interest $\mathcal{D}$ for Theorem 4.3. The smaller shaded triangle shows a region where the analogous update to* $\boldsymbol{a}$ *would have acceptable norm. The points marked are the ones we look at numerically, including* $(0.88, 0.57)$*, which corresponds to the Stitch Fix data in Section 6.*

5. Convergence and computation. In this section, we make some computations on synthetic data following the probability model from Section 4. First, we study the norms of our update matrix $M^{(2)}$, which affects the number of iterations to convergence. In addition to $\|\cdot\|_1$ covered in Theorem 4.3, we also consider $\|\cdot\|_2$, $\|\cdot\|_\infty$ and $\lambda_{\max}(\cdot)$. Then we compare the cost to compute $\hat{\beta}_{\mathrm{GLS}}$ by our backfitting method with that of lmer (Bates et al. (2015)).

The problem size is indexed by S. Indices i go from 1 to $R = \lceil S^\rho \rceil$ and indices j go from 1 to $C = \lceil S^\kappa \rceil$. Reasonable parameter values have $\rho, \kappa \in (0, 1)$ with $\rho + \kappa > 1$. Theorem 4.3 applies when $2\rho + \kappa < 2$ and $3\rho + 4\kappa < 4$. Figure 1 depicts this triangular domain of interest $\mathcal{D}$. There is another triangle $\mathcal{D}'$ where a corresponding update for $\boldsymbol{a}$ would satisfy the conditions of Theorem 4.3. Then $\mathcal{D} \cup \mathcal{D}'$ is a nonconvex polygon of five sides. Figure 1 also shows $\mathcal{D}' \setminus \mathcal{D}$ as a second triangular region. For points (ρ, κ) near the line $\rho + \kappa = 1$, the matrix Z will be mostly ones unless S is very large. For points (ρ, κ) near the upper corner $(1, 1)$, the matrix Z will be extremely sparse with each $N_{i\bullet}$ and $N_{\bullet j}$ having nearly a Poisson distribution with mean between 1 and Υ. The fraction of potential values that have been observed is $O(S^{1-\rho-\kappa})$.

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

The following (ρ, κ) combinations are of interest. First, $(4/5, 2/5)$ is the closest vertex of the domain of interest to the point $(1, 1)$. Second, $(2/5, 4/5)$ is outside the domain of interest for the $\boldsymbol{b}$ but within the domain for the analogous $\boldsymbol{a}$ update. Third, among points with $\rho = \kappa$, the value $(4/7, 4/7)$ is the farthest one from the origin that is in the domain of interest. We also look at some points on the 45 degree line that are outside the domain of interest because the sufficient conditions in Theorem 4.3 might not be necessary.

In our matrix norm computations, we took $\lambda_A = \lambda_B = 0$. This completely removes shrinkage and will make it harder for the algorithm to converge than would be the case for the

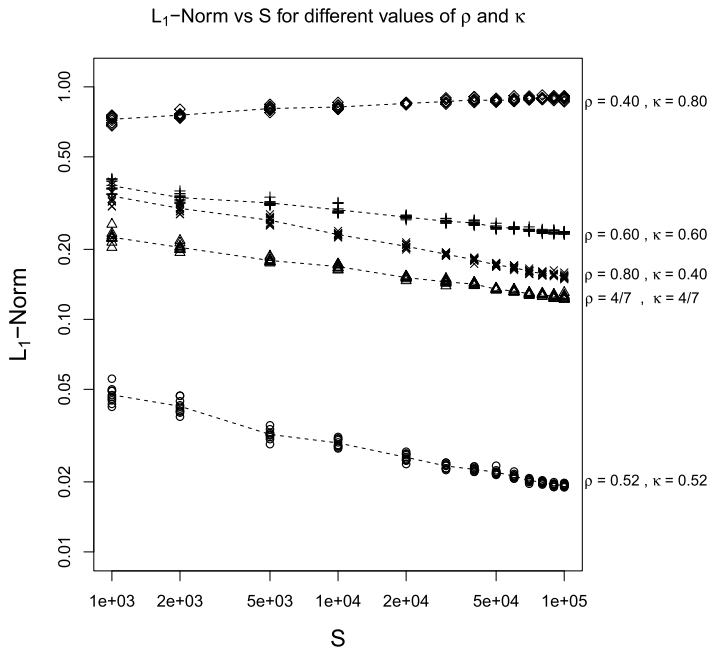


FIG. 2. *Norm $\|M^{(2)}\|_1$ of centered update matrix versus problem size S for different (ρ, κ).*

positive λ_A and λ_B that hold in real data. The values of λ_A and λ_B appear in expressions $N_{i\bullet} + \lambda_A$ and $N_{\bullet j} + \lambda_B$ where their contribution is asymptotically negligible, so conservatively setting them to zero will nonetheless be realistic for large data sets.

We sampled from the model 10 times at various values of S and have plotted $\|M^{(2)}\|_1$ versus S on a logarithmic scale. Figure 2 shows the results. We observe that $\|M^{(2)}\|_1$ is below 1 and decreasing with S for all the examples $(\rho, \kappa) \in \mathcal{D}$. This holds also for $(\rho, \kappa) = (0.60, 0.60) \notin \mathcal{D}$. We chose that point because it is on the convex hull of $\mathcal{D} \cup \mathcal{D}'$.

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

The point $(\rho, \kappa) = (0.70, 0.70)$ is not in the domain $\mathcal{D}$ covered by Theorem 4.3 and we see that $\|M^{(2)}\|_1 > 1$ and generally increasing with S as shown in Figure 3, which uses 10 replicates. This does not mean that backfitting must fail to converge. Here, we find that $\|M^{(2)}\|_2 < 1$ and generally decreases as S increases. This is a strong indication that the number of backfitting iterations required will not grow with S for this (ρ, κ) combination. We cannot tell whether $\|M^{(2)}\|_2$ will decrease to zero but that is what appears to happen.

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

We believe that backfitting will have a spectral radius well below 1 for more cases than we can as yet prove. In addition to the previous figures showing matrix norms as S increases for certain special values of (ρ, κ), we have computed contour maps of those norms over $(\rho, \kappa) \in [0, 1]$ for $S = 10{,}000$. See Figure 4. Each point in our contour plots was based on an average of 10 independent replicates.

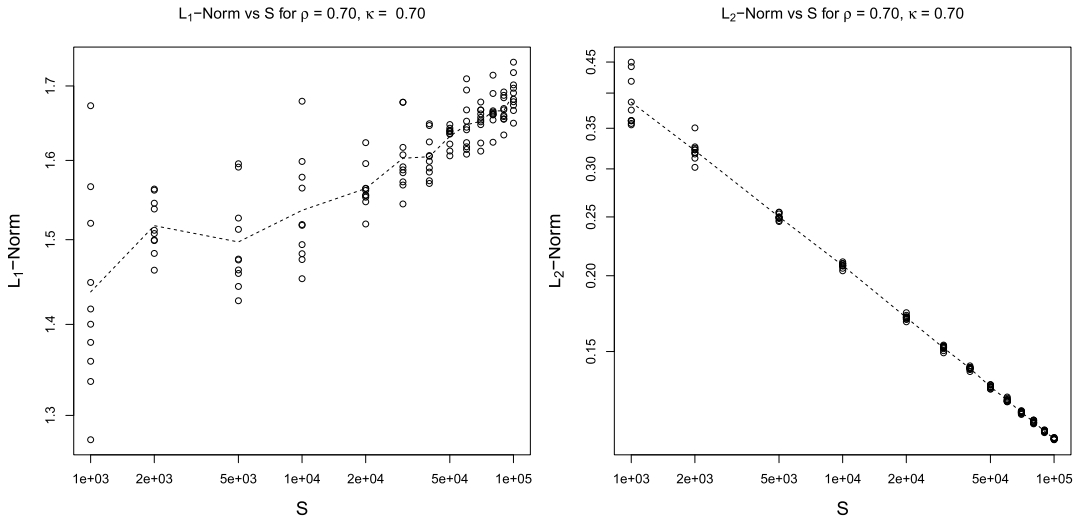


FIG. 3. *The left panel shows* $\|M^{(2)}\|_1$ *versus* S. *The right panel shows* $\|M^{(2)}\|_2$ *versus* S *with a logarithmic vertical scale. Both have* $(\rho, \kappa) = (0.7, 0.7)$.

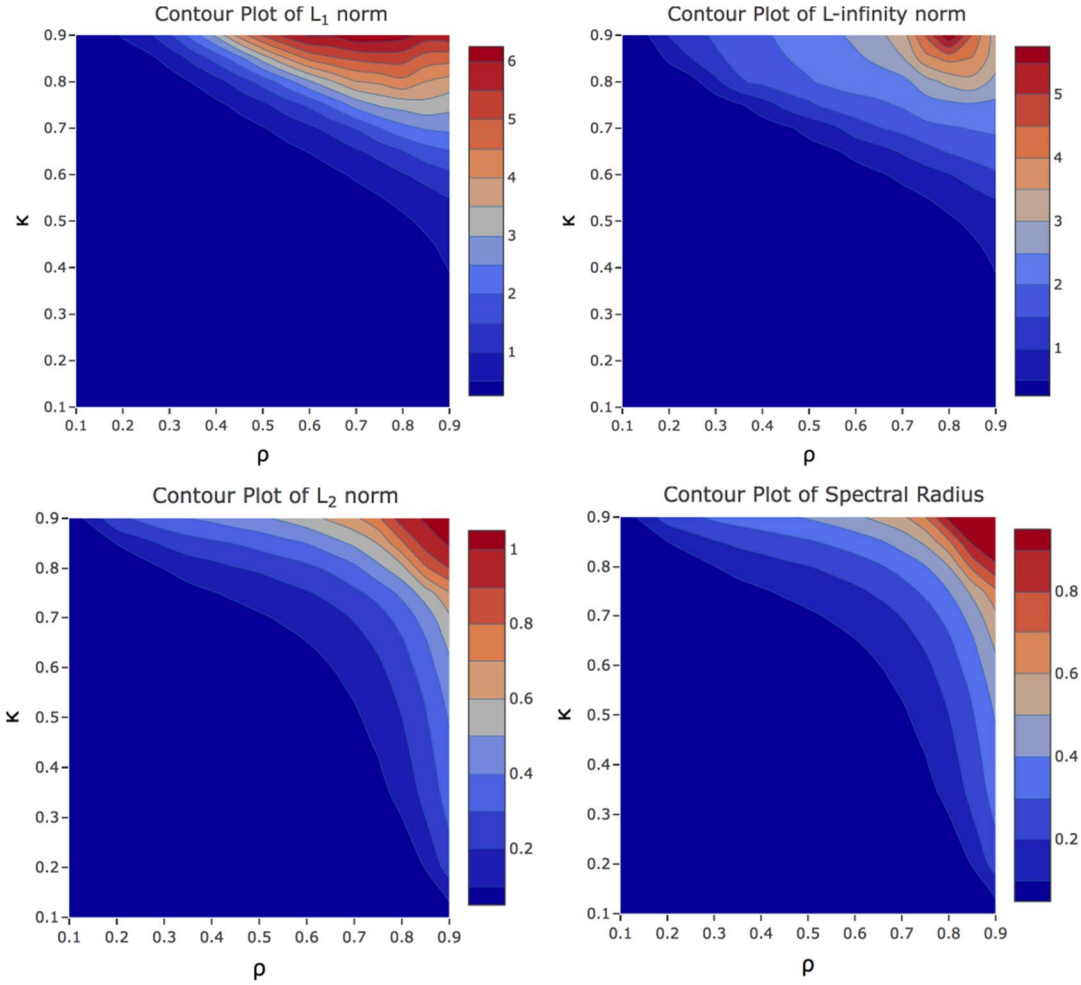
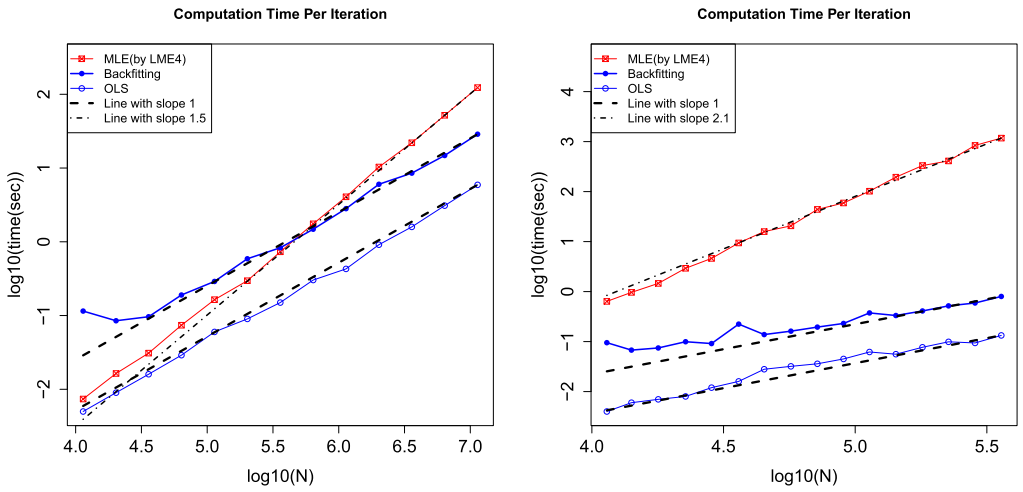


FIG. 4. *Numerically computed matrix norms for* $M^{(2)}$ *using* $S = 10{,}000$. *The color code varies with the subfigures.*



(a) $(\rho, \kappa) = (0.52, 0.52)$ (b) $(\rho, \kappa) = (0.70, 0.70)$

FIG. 5. *Time for one iteration versus the number of observations, N at two points (ρ, κ). The cost for lmer is roughly $O(N^{3/2})$ in the left panel and $O(N^{2.1})$ in the right panel. The costs for OLS and backfitting are $O(N)$.*

To compare the computation times for algorithms we generated Z_{ij} as above and also took $x_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_7)$. That gives 7 fixed effect parameters and then the intercept makes it $p = 8$. Although backfitting can run with $\lambda_A = \lambda_B = 0$, lmer cannot do so for numerical reasons. So we took $\sigma_A^2 = \sigma_B^2 = 1$ and $\sigma_E^2 = 1$ corresponding to $\lambda_A = \lambda_B = 1$. The cost per iteration does not depend on Y_{ij}, and hence not on β either. We used $\beta = 0$.

Figure 5 shows computation times for one single iteration when $(\rho, \kappa) = (0.52, 0.52)$ and when $(\rho, \kappa) = (0.70, 0.70)$. The time to do one iteration in lmer grows roughly like $N^{3/2}$ in the first case. For the second case, it appears to grow at the even faster rate of $N^{2.1}$. Solving a system of $S^\kappa \times S^\kappa$ equations would cost $S^{3\kappa} = S^{2.1} = O(N^{2.1})$, which explains the observed rate. This analysis would predict $O(N^{1.56})$ for $\rho = \kappa = 0.52$ but that is only minimally different from $O(N^{3/2})$. These experiments were carried out in R on a computer with the macOS operating system, 16 GB of memory and an Intel i7 processor. Each backfitting iteration entails solving (17) along with the fixed effects.

The cost per iteration for backfitting follows closely to the $O(N)$ rate predicted by the theory. OLS only takes one iteration and it is also of $O(N)$ cost. In both of these cases, $\|M^{(2)}\|_2$ is bounded away from one so the number of backfitting iterations does not grow with S. For $\rho = \kappa = 0.52$, backfitting took 4 iterations to converge for the smaller values of S and 3 iterations for the larger ones. For $\rho = \kappa = 0.70$, backfitting took 6 iterations for smaller S and 4 or 5 iterations for larger S. In each case, our convergence criterion was a relative change of 10^{-8} as described in Section 6.3. Further backfitting to compute BLUPs $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ given $\hat{\beta}_{\text{GLS}}$ took at most 5 iterations for $\rho = \kappa = 0.52$ and at most 10 iterations for $\rho = \kappa = 0.7$. In the second example, lme4 did not reach convergence in our time window so we ran it for just 4 iterations to measure its cost per iteration.

6. Example: Ratings from Stitch Fix. We illustrate backfitting for GLS on some data from Stitch Fix. Stitch Fix sells clothing. They mail their customers a sample of items. The customers may keep and purchase any of those items that they want, while returning the others. It is valuable to predict the extent to which a customer will like an item, not just whether they will purchase it. Stitch Fix has provided us with some of their client ratings data. It was anonymized, void of personally identifying information, and as a sample it does

not reflect their total numbers of clients or items at the time they provided it. It is also from 2015. While it does not describe their current business, it is a valuable data set for illustrative purposes.

The sample sizes for this data are as follows. We received $N = 5{,}000{,}000$ ratings by $R = 762{,}752$ customers on $C = 6318$ items. These values of R and C correspond to the point $(0.88, 0.57)$ in Figure 1. Thus $C/N \doteq 0.00126$ and $R/N \doteq 0.153$. The data are not dominated by a single row or column because $\max_i N_{i\bullet}/N \doteq 9 \times 10^{-6}$ and $\max_j N_{\bullet j}/N \doteq 0.0143$. The data are sparse because $N/(RC) \doteq 0.001$.

6.1. *An illustrative linear model.* The response Y_{ij} is a rating on a ten-point scale of the satisfaction of customer i with item j. The data come with features about the clients and items. In a business setting, one would fit and compare possibly dozens of different regression models to understand the data. Our purpose here is to study large scale GLS and compare it to ordinary least squares (OLS) and so we use just one model, not necessarily one that we would have settled on. For that purpose, we use the same model that was used in Gao and Owen (2020). It is not chosen to make OLS look as bad as possible. Instead it is potentially the first model one might look at in a data analysis. For client i and item j,

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

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

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

$$U_A = \sum_i \sum_j Z_{ij} \Bigg(Y_{ij} - \frac{1}{N_{i\bullet}} \sum_{j'} Z_{ij'} Y_{ij'} \Bigg)^2,$$

$$U_B = \sum_j \sum_i Z_{ij} \Bigg(Y_{ij} - \frac{1}{N_{\bullet j}} \sum_{i'} Z_{i'j} Y_{i'j} \Bigg)^2 \quad \text{and}$$

$$U_E = N \sum_{ij} Z_{ij} \Bigg(Y_{ij} - \frac{1}{N} \sum_{i'j'} Z_{i'j'} Y_{i'j'} \Bigg)^2.$$

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

$$\mathbb{E}(U_A) = \big(\sigma_B^2 + \sigma_E^2\big)(N - R),$$

$$\mathbb{E}(U_B) = \big(\sigma_A^2 + \sigma_E^2\big)(N - C) \quad \text{and}$$

$$\mathbb{E}(U_E) = \sigma_A^2\Big(N^2 - \sum_i N_{i\bullet}^2\Big) + \sigma_B^2\Big(N^2 - \sum_j N_{\bullet j}^2\Big) + \sigma_E^2(N^2 - N).$$

By matching moments, we can estimate θ by solving the 3×3 linear system

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

for θ.

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

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

(1) Compute $\hat\beta_{\mathrm{OLS}}$ via (2).
(2) Get residuals $\hat\eta_{ij} = Y_{ij} - x_{ij}^\mathsf{T}\hat\beta_{\mathrm{OLS}}$.
(3) Compute $\hat\sigma_A^2$, $\hat\sigma_B^2$ and $\hat\sigma_E^2$ by the method of moments on $\hat\eta_{ij}$.
(4) Compute $\tilde{\mathcal{X}} = (I_N - \tilde{\mathcal{S}}_G)\mathcal{X}$ using doubly centered backfitting $M^{(3)}$.
(5) Compute $\hat\beta_{\mathrm{GLS}}$ by (28).
(6) If we want BLUPs $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ backfit $\mathcal{Y} - \mathcal{X}\hat\beta_{\mathrm{GLS}}$ to get them.
(7) Compute $\widehat{\mathrm{cov}}(\hat\beta_{\mathrm{GLS}})$ by plugging $\hat\sigma_A^2$, $\hat\sigma_B^2$ and $\hat\sigma_E^2$ into $\mathcal{V}$ at (28).

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

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

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

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

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

We can quantify the naivete of OLS via the ratios $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat\beta_{\mathrm{OLS},j})/\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat\beta_{\mathrm{OLS},j})$ for $j = 1, \ldots, p$. Figure 6 plots these values. They range from 1.75 to 345.28 and can be interpreted as factors by which OLS naively overestimates its sample size. The largest and second largest

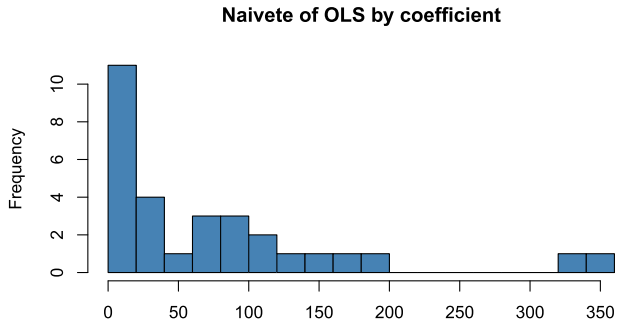


FIG. 6. *OLS naivete* $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})/\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})$ *for coefficients* β_j *in the Stitch Fix data.*

ratios are for material indicators corresponding to "Modal" and "Tencel," respectively. These appear to be two names for the same product with Tencel being a trademarked name for Modal fibers (made from wood). We can also identify the linear combination of $\hat{\beta}_{\mathrm{OLS}}$ for which OLS is most naive. We maximize the ratio $x^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})x/x^{\mathsf{T}}\widehat{\mathrm{cov}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})x$ over $x \neq 0$. The resulting maximal ratio is the largest eigenvalue of

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

and it is about 361 for the Stitch Fix data.

We can quantify the inefficiency of OLS, coefficient by coefficient, via the ratio $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})/\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS},j})$. Figure 7 plots these values. They range from just over 1 to 50.6 and can be interpreted as factors by which using OLS reduces the effective sample size. There is a clear outlier: the coefficient of the match variable is very inefficiently estimated by OLS. The second largest inefficiency factor is for the intercept term. The most inefficient linear combination of $\hat{\beta}$ reaches a variance ratio of 52.6, only slightly more inefficient than the match coefficient alone.

The variables for which OLS is more naive tend to also be the variables for which it is most inefficient. Figure 8 plots these quantities against each other for the 30 coefficients in our model.

6.5. *Convergence speed of backfitting.* The Stitch Fix data have row and column sample sizes that are much more uneven than our sampling model for Z allows. Accordingly, we cannot rely on Theorem 4.3 to show that backfitting must converge rapidly for it.

The sufficient conditions in that theorem may not be necessary and we can compute our norms and the spectral radius on the update matrices for the Stitch Fix data using some sparse

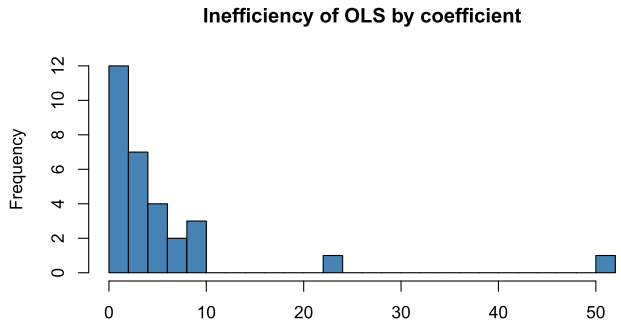


FIG. 7. *OLS inefficiency* $\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})/\widehat{\mathrm{cov}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS},j})$ *for coefficients* β_j *in the Stitch Fix data.*

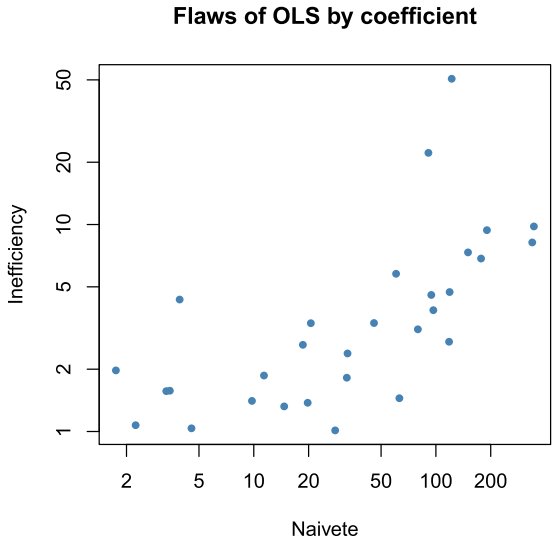


FIG. 8. *Inefficiency vs. naivete for OLS coefficients in the Stitch Fix data.*

matrix computations. Here, $Z \in \{0,1\}^{762{,}752\times 6318}$, so $M^{(k)} \in \mathbb{R}^{6318\times 6318}$ for $k \in \{0,1,2,3\}$. The results are

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

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

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

Theorem 4 of Papaspiliopoulos, Roberts and Zanella (2020) has the rate of convergence for their collapsed Gibbs sampler for balanced data. It involves an auxiliary convergence rate ρ_{aux} defined as follows. Consider the Gibbs sampler on (i,j) pairs where given i a random

j is chosen with probability $Z_{ij}/N_{i\bullet}$ and given j a random i is chosen with probability $Z_{ij}/N_{\bullet j}$. That Markov chain has invariant distribution Z_{ij}/N on (i, j) pairs and ρ_{aux} is the rate at which the chain converges. In our notation,

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

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

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

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

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

Whether one prefers a GLS estimate or a Bayesian one depends on context and goals. We believe that there is a strong computational advantage to GLS for large data sets. The cost of one backfitting iteration is comparable to the cost to generate one more sample in the MCMC. We may well find that only a dozen or so iterations are required for convergence of the GLS. A Bayesian analysis requires a much larger number of draws from the posterior distribution than that. For instance, Gelman and Shirley (2011) recommend an effective sample size of about 100 posterior draws, with autocorrelations requiring a larger actual sample size. Vats, Flegal and Jones (2019) advocate even greater effective sample sizes.

It is usually reasonable to assume that there is a selection bias underlying which data points are observed. Accounting for any such selection bias must necessarily involve using infor-

mation or assumptions from outside the data set at hand. We expect that any approach to take proper account of informative missingness must also make use of solutions to GLS perhaps after reweighting the observations. Before one develops any such methods, it is necessary to first be able to solve GLS without regard to missingness.

Many of the problems in electronic commerce involve categorical outcomes, especially binary ones, such as whether an item was purchased or not. Generalized linear mixed models are then appropriate ways to handle crossed random effects, and we expect that the progress made here will be useful for those problems.

APPENDIX

Table 1 shows results of OLS and GLS regression for the Stitch Fix data in Section 6. OLS is estimated to be naive when $\widehat{\mathrm{SE}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}}) < \widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})$ and inefficient when $\widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}}) > \widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS}})$. Estimates that are more than double their corresponding standard error get an asterisk.

TABLE 1
Stitch Fix regression results

	$\hat{\beta}_{\mathrm{OLS}}$	$\widehat{\mathrm{SE}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})$	$\widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS}})$	$\hat{\beta}_{\mathrm{GLS}}$	$\widehat{\mathrm{SE}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS}})$
Intercept	4.635*	0.005397	0.05148	5.103*	0.01092
Match	5.048*	0.01174	0.1297	3.442*	0.01823
$\mathbb{I}${client edgy}	0.001020	0.002443	0.004444	0.003041	0.003550
$\mathbb{I}${item edgy}	−0.3358*	0.004253	0.03307	−0.3515*	0.01375
$\mathbb{I}${client edgy}					
$*\mathbb{I}${item edgy}	0.3925*	0.006229	0.01233	0.3793*	0.005916
$\mathbb{I}${client boho}	0.1386*	0.002264	0.004211	0.1296*	0.003356
$\mathbb{I}${item boho}	−0.5499*	0.005981	0.02713	−0.6266*	0.01485
$\mathbb{I}${client boho}					
$*\mathbb{I}${item boho}	0.3822*	0.007566	0.01001	0.3763*	0.007123
Acrylic	−0.06482*	0.003778	0.03371	−0.005360	0.01909
Angora	−0.01262	0.007848	0.08530	0.07486	0.05177
Bamboo	−0.04593	0.06215	0.2096	0.03251	0.1535
Cashmere	−0.1955*	0.02484	0.1414	0.008930	0.1048
Cotton	0.1752*	0.003172	0.04220	0.1033*	0.01612
Cupro	0.5979*	0.3016	0.4519	0.2089	0.4363
Faux Fur	0.2759*	0.02008	0.07694	0.2749*	0.06691
Fur	−0.2021*	0.03121	0.1388	−0.07924	0.1182
Leather	0.2677*	0.02482	0.07759	0.1674*	0.06545
Linen	−0.3844*	0.05632	0.2429	−0.08658	0.1499
Modal	0.002587	0.009775	0.1816	0.1388*	0.05804
Nylon	0.03349*	0.01552	0.08878	0.08174	0.05751
Patent Leather	−0.2359	0.1800	0.3838	−0.3764	0.3771
Pleather	0.4163*	0.008916	0.08774	0.3292*	0.04468
PU	0.4160*	0.008225	0.07989	0.4579*	0.03737
PVC	0.6574*	0.06545	0.3462	0.9688*	0.3441
Rayon	−0.01109*	0.002951	0.04074	0.05155*	0.01329
Silk	−0.1422*	0.01317	0.08907	−0.1828*	0.04871
Spandex	−0.3916*	0.00931	0.1373	0.4140*	0.1141
Tencel	0.4966*	0.01729	0.1712	0.1234*	0.05982
Viscose	0.04066*	0.006953	0.08519	−0.02259	0.03145
Wool	−0.06021*	0.006611	0.07211	−0.05883	0.03319

Acknowledgments. We are grateful to Brad Klingenberg and Stitch Fix for sharing some test data with us and James Johndrow for some helpful discussions. We thank the reviewers for remarks that have helped us improve the paper.

Funding. This work was supported by the U.S. National Science Foundation under Grant IIS-1837931.

REFERENCES

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48. https://doi.org/10.18637/jss.v067.i01

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249 https://doi.org/10.1214/aos/1176347115

CAMERON, A. C., GELBACH, J. B. and MILLER, D. L. (2011). Robust inference with multiway clustering. *J. Bus. Econom. Statist.* **29** 238–249. MR2807878 https://doi.org/10.1198/jbes.2010.07136

GAO, K. (2017). Scalable estimation and inference for massive linear mixed models with crossed random effects. Ph.D. thesis, Stanford Univ. MR4257222

GAO, K. and OWEN, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electron. J. Stat.* **11** 1235–1296. MR3635913 https://doi.org/10.1214/17-EJS1236

GAO, K. and OWEN, A. B. (2020). Estimation and inference for very large linear mixed effects models. *Statist. Sinica* **30** 1741–1771. MR4260743 https://doi.org/10.5705/ss.202018.0029

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, Cambridge.

GELMAN, A. and SHIRLEY, K. (2011). Inference from simulations and monitoring convergence. In *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 163–174. CRC Press, Boca Raton, FL. MR2858448 https://doi.org/10.1201/b10905

GOODMAN, J. and SOKAL, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D, Part. Fields* **40** 2035–2071. https://doi.org/10.1103/physrevd.40.2035

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

OWEN, A. B. (2007). The pigeonhole bootstrap. *Ann. Appl. Stat.* **1** 386–411. MR2415741 https://doi.org/10.1214/07-AOAS122

PAPASPILIOPOULOS, O., ROBERTS, G. O. and ZANELLA, G. (2020). Scalable inference for crossed random effects models. *Biometrika* **107** 25–40. MR4064138 https://doi.org/10.1093/biomet/asz058

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

ROBERTS, G. O. and SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **59** 291–317. MR1440584 https://doi.org/10.1111/1467-9868.00070

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–51. MR1108815

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR1190470 https://doi.org/10.1002/9780470316856

STRASSEN, V. (1969). Gaussian elimination is not optimal. *Numer. Math.* **13** 354–356. MR0248973 https://doi.org/10.1007/BF02165411

VATS, D., FLEGAL, J. M. and JONES, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106** 321–337. MR3949306 https://doi.org/10.1093/biomet/asz002