ESTIMATION AND INFERENCE FOR VERY LARGE LINEAR MIXED EFFECTS MODELS

Katelyn Gao and Art B. Owen

Intel Inc. and Stanford University

Abstract: Linear mixed models with large imbalanced crossed random effects structures pose severe computational problems for maximum likelihood estimation and for Bayesian analysis. The costs can grow as fast as $N^{3/2}$ when there are N observations. Such problems arise in any setting where the underlying factors satisfy a many to many relationship (instead of a nested one) and in electronic commerce applications, the N can be quite large. Methods that do not account for the correlation structure can greatly underestimate uncertainty. We propose a method of moments approach that takes account of the correlation structure and that can be computed at $O(N)$ cost. The method of moments can easily be parallelized and does not require parametric distributional assumptions, tuning parameters or convergence diagnostics. For the regression coefficients, we give conditions for consistency and asymptotic normality as well as a consistent variance estimate. We give conditions for consistent estimation of the variance components and we provide consistent estimates of a mildly conservative upper bound on the variance of the variance component estimates. All of these computations can be done in $O(N)$ work. We illustrate the algorithm with some data from Stitch Fix where the crossed random effects correspond to clients and items. There a naive analysis can overestimate the effective sample size by hundreds and yield unreliable conclusions about parameters.

Key words and phrases: Crossed random effects, Linear mixed models, Scalable inference

1. Introduction

The field of statistics is confronting two important challenges at present. The first is the arrival of ever larger data sets, sometimes described as 'big data'. See, for instance, [26] and [31]. The second is the reproducibility crisis, in which published findings cannot be replicated. This problem was clearly presented by [17] among others, and it has led to the American Statistical Association releasing a statement on p-values [32].

We might naively hope that the first problem would in fact be a blessing in disguise and would remove the second problem due to the decrease in uncertainty. However, this may not be true in reality, where difficulties remain when data are not independent. We consider here one such situation, when there is crossed random effects structure in the data. That structure introduces a dense tangle of correlations that can sharply reduce the effective sample size of the data at hand. If, as we suspect, most data scientists treat these large data sets as IID samples, then they will greatly underestimate the uncertainty in their fitted models. The usual methods for this problem, whether by maximum likelihood or restricted maximum likelihood (REML) or Bayes, have a cost that grows superlinearly in the sample size and simply cannot be run on the largest data sets. We present and study a method of moments approach with cost that scales linearly in the problem size, among other advantages.

The sort of data that motivates us arise in e-commerce applications. The factors are variables like cookies, customer IDs, query strings, IP addresses, product IDs (e.g., SKUs), URLs and so on. The most direct way to handle such variables is to treat them as categorical variables that simply happen to have a large number of levels, including many that have not yet appeared in the data. We think that a random effects model is more appropriate [22]. For instance, internet cookies are cleared regularly and hence any specific cookie is likely to disappear shortly. It is therefore more appropriate to consider the specific cookies in a data set as a sample from some distribution, that is, as a random effect. Similarly there is turnover in popular products and queries that motivates treating them as random effects too.

While the largest crossed random effect data sets we know of are in e-commerce and social media (for example, the Netflix data set [4]), we expect the problem to arise in other settings where data set sizes are growing. The crossed random effects structure is fundamental. Any setting with a many to many mapping of factor levels involves crossed effects that one might want to model as random. In agriculture and genomics there are gene by environment or gene by patient crosses. In education, neither schools nor neighborhoods are perfectly nested within the other [27] and in multiyear data sets there is a many to many relationship between teachers and students.

When our chosen model involves only one of these random effect entities then a hierarchical model, based on Bayes or empirical Bayes, can be quite effective [34, 12]. Things change considerably when we want to use two or more crossed random effects. In this paper, we consider the following model,

Model 1. *Two-factor linear mixed effects:*

$$\begin{aligned}&Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + b_j + e_{ij}, \quad x_{ij} \in \mathbb{R}^p, \quad i, j \in \mathbb{N} \quad \textit{where,}\\ &a_i \overset{\text{iid}}{\sim} (0, \sigma_A^2), \quad b_j \overset{\text{iid}}{\sim} (0, \sigma_B^2), \quad e_{ij} \overset{\text{iid}}{\sim} (0, \sigma_E^2) \quad \textit{(independently) and,}\\ &E(a_i^4) < \infty, \quad E(b_j^4) < \infty, \quad E(e_{ij}^4) < \infty.\end{aligned} \tag{1.1}$$

For instance, customer i might assign a score Y_{ij} to product j. Then x_{ij} contains features about the customer or product or some joint properties of them, β is of interest for the company choosing what product to recommend, b_j measures some general appeal of the product not captured by the features in x_{ij}, while a_i captures variation in which customers are harder or easier to please and e_{ij} is an error term. This is a mixed effects model because it contains both random effects a_i, b_j and fixed effects x_{ij}.

Model 1 describes any ij pair, but the given data set will only contain some finite number N of them. If the available data are laid out as rows i and columns j with R distinct rows and C distinct columns, then the cost of fitting a generalized least squares regression model for β scales as $O((R+C)^3)$ because it solves a $p \times p$ system of equations with $p \geqslant R+C$. See [30], [27] and [1]. Now because $RC \geqslant N$ we have $\max(R, C) \geqslant \sqrt{N}$ and $(R+C)^3 > N^{3/2}$.

Gao and Owen [11] consider an intercept-only version of Model 1 where $x_{ij}^{\mathsf{T}}\beta$ is simply a constant $\mu \in \mathbb{R}$ for all i and j. They find that Markov chain Monte Carlo (MCMC) does not solve the inference problem under the assumption that the random effects are normally distributed. All of the MCMC methods considered either failed to mix, or converged to the wrong answer, and this took place already at modest sample sizes. For the specific case of a Gibbs sampler and Gaussian a_i, b_j and e_{ij}, using methods from [28] they prove that it will take $O(N^{1/2})$ iterations costing $O(N)$ each to converge, for a total cost of $O(N^{3/2})$. Fox [9] presents a very general equivalence between the convergence rate of an iterative equation solver and the convergence rate of an associated MCMC scheme, so these identical rates may be a sign of a deeper connection. Consensus Bayes [29] splits the data into shards, one per processor. However the data given to each shard has to be independent and here data sets corresponding to a subset of rows will have correlations due to commonly sampled columns (and vice versa). As an alternative to MCMC, variational Bayesian methods can be used to approximate the posterior distribution of the parameters by minimizing the divergence of a chosen parametric density from the posterior [6]. However, it is not straightforward to choose the parametric density and little is known about the theoretical properties of such methods.

Likelihood-based approaches are commonly used to analyze Model 1 under a Gaussian model for the random effects. See for instance Jiang [19]. Maximum likelihood (ML) maximizes the log likelihood of the data with respect to the parameters [7, 14], while REML mitigates the bias of the ML estimates by estimating the variance components via ML on the residuals from fitting β using least squares. Various optimization algorithms have been applied to compute ML and REML estimates, including gradient-free ones such as BOBYQA [25] and Expectation-Maximization [8] and gradient-based ones such as Gauss-Newton [1]. The asymptotic variances of these estimates are readily obtained from the Fisher information matrix. The main disadvantage of likelihood-based approaches is that even computing the value of the likelihood at given values of the parameters requires $O(N^{3/2})$ time [1].

Thus, we find that existing Bayes and likelihood methods are not effective for this problem. Here we present an approach based on the method of moments. We seek estimates $\hat{\beta}$, $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\sigma}_E^2$ along with variance estimates for these quantities. We have three criteria:

1) the total computational cost *must* be $O(N)$ time and $O(R+C)$ space,
2) the variance estimates *should* be reliable or conservative, and
3) we *prefer* $\hat{\beta}$ to be statistically efficient.

We regard the first criterion as a constraint that must be met. For the second criterion, a mild over-estimate of $\mathrm{Var}(\hat{\beta})$ is acceptable in order to keep the costs in $O(N)$. The third criterion is to be met as well as we can, subject to constraints given by the first two. Computational efficiency is more important than statistical efficiency in this context. For very large N, requiring $O(N^{3/2})$ computation is like asking for an oracle: it is something no user could ever get.

With an apt choice of the estimating equations, the method of moments meets our $O(N)$ time and $O(R+C)$ space criteria, and we show that it can also yield reliable variance estimates. Further advantages of the method of moments are that it does not require parametric distributional assumptions (e.g., Gaussianity), there are no tuning parameters to choose, and most importantly for large N, it is very well suited to parallel computation. The method of moments is not without drawbacks. Sometimes it yields parameter estimates that are out of bounds, such as negative variance estimates, which are often set to zero [30]. The method of moments has been highly successful in the case of nested random effects. In [33], the authors propose estimators with linear computational complexity for not only the regression coefficients and the variance components, but higher moments of the random effects. See also [24].

An outline of this paper is as follows. Section 2 introduces most of the notation for Model 1, especially the pattern of missingness in the data, and gives some of the asymptotic assumptions. Section 3 presents our algorithm and shows that it takes $O(N)$ time and $O(R+C)$ space. We compute a generalized least squares (GLS) estimate for a model with either row or column variance components, but not both. We choose based on an efficiency criterion. Then we estimate $\mathrm{Var}(\hat{\beta})$ accounting for all three error terms including the one left out of the GLS estimate. Section 4 illustrates our algorithm with some ratings data from Stitch Fix. There Y_{ij} is a rating, from a 10 point scale, by customer i on item j, with features x_{ij}. Compared to ordinary least squares (OLS) estimates, the random effects model leads to standard errors on coefficients β_j that can be more than ten times higher. That may be interpreted as an effective sample size which is less than 1% of the nominal sample size. Section 5 gives conditions under which $\hat{\beta}$ and the variance components are consistent. There is also a central limit theorem for $\hat{\beta}$. Section 6 compares our method of moments estimator to a state of the art GLMM code [2] written in Julia [5]. That algorithm takes $O(N^{3/2})$ cost per iteration, with a number of iterations that, in our simulations, depends on N. On problems where the GLMM code gives an answer we find it more statistically efficient for β and σ_E^2 but not for σ_A^2 or σ_B^2. Section 7 discusses some future work and related literature.

Our method of moments approach is similar to Henderson's classical methods [15] for Gaussian data, as presented in [30]. For an intercept-only model, our previous paper [11] uses U-statistics to find a counterpart to the Henderson I estimator that can be computed in $O(N)$ time and $O(R+C)$ space. We also get a variance estimator for the variance components, without assuming a Gaussian distribution. The variance estimator can be computed in $O(N)$ time. It targets a mildly conservative upper bound on the variance as the variance itself, like the one for Henderson's estimates, takes more than $O(N)$ computation. In this paper we incorporate fixed effects along with the random effects, just as Henderson II does in generalizing Henderson I, by transforming the original model to one with only random effects. Like Henderson II, our algorithm alternates between estimating the regression coefficients and variance components. However, our estimators are different, having linear computational complexity instead of superlinear. Henderson III uses estimating equations based on the residual sum of squares when treating subsets of the random effects as fixed effects and fitting OLS models. In addition, Henderson III also allows for interactions between fixed and random effects. We believe such interactions are very reasonable in our motivating applications, but incorporating them is outside the scope of this article.

Our analysis is for a fixed dimension p. This is reasonable for our motivating data from Stitch Fix, where $p \ll N$. It remains to develop methods for cases where $p \to \infty$ with N.

Another issue that we do not address in this article is selection bias in the available observations. Sometimes ratings are biased towards the high end because customers seek products that they expect to like and companies endeavor to recommend such products. In other data sets, such as restaurant reviews, customers may be more likely to make a rating when they are either very unhappy or very happy. For such data, the ratings will be biased towards both extremes and away from the middle. Accounting for selection bias requires assumptions or information from outside the given data. Propensity weighting [16, Chapter 13] could possibly fit into our framework, but we leave that out of this paper, as the basic problem without selection bias is already a challenge.

2. Notation and asymptotic conditions

Here we give a fuller presentation of our notation. Equation (1.1) describes the distribution of observed and future data. We call the first index of Y_{ij} the 'row' and the second the 'column'. We use integers i, i', r, r' to index rows and j, j', s, s' for columns, but the actual indices may be URLs, customer IDs, or query strings. The index sets are countably infinite to always leave room for unseen levels in the future.

The variable Z_{ij} takes the value 1 if (x_{ij}, Y_{ij}) is observed and 0 otherwise. We assume that there is at most one observation in position (i, j). For customer rating data, we suppose that if i has rated j multiple times, then only the most recent rating is retained. In many other settings, only a negligible fraction of ij pairs will have been duplicated.

The sample size is $N = \sum_{ij} Z_{ij} < \infty$. The number of observations in row i is $N_{i\bullet} = \sum_j Z_{ij}$ and the number in column j is $N_{\bullet j} = \sum_i Z_{ij}$. The number of distinct rows is $R = \sum_i 1_{N_{i\bullet}>0}$ and there are $C = \sum_j 1_{N_{\bullet j}>0}$ distinct columns. In the following, summing over rows i means summing over just the R rows i with $N_{i\bullet} > 0$, and similarly for sums over columns. This convention corresponds to what happens when one makes a pass through the whole data set.

Let Z be the matrix containing Z_{ij}. Then $(ZZ^\mathsf{T})_{ii'} = \sum_j Z_{ij} Z_{i'j}$ is the number of columns for which we have data in both rows i and i'. Similarly, $(Z^\mathsf{T} Z)_{jj'}$ is the number of rows in which both columns j and j' are observed.

Note that $(ZZ^{\mathsf{T}})_{ii'} \leqslant N_{i\bullet}$ and $(Z^{\mathsf{T}}Z)_{jj'} \leqslant N_{\bullet j}$. We will use the following identities:

$$\sum_{ir}(ZZ^{\mathsf{T}})_{ir} = \sum_j N_{\bullet j}^2, \quad \text{and} \quad \sum_{js}(Z^{\mathsf{T}}Z)_{js} = \sum_i N_{i\bullet}^2.$$

This notation allows for an arbitrary pattern of observations. We mention three special cases. A balanced crossed design has $Z_{ij} = 1_{i\leqslant R}1_{j\leqslant C}$. If $\max_i N_{i\bullet} = 1$ but $\max_j N_{\bullet j} > 1$ then the data have a hierarchical structure with rows nested in columns. If $\max_i N_{i\bullet} = \max_j N_{\bullet j} = 1$, then the observed Y_{ij} have IID errors. Some of these patterns cause problems for parameter estimation. For example, if the errors are IID, then the variance components are not identifiable. Our assumptions rule these out to focus on large genuinely crossed data sets.

The following vectors are useful for subsequent analyses. Let $v_{1,i}$ be the length-N vector with ones in entries $\sum_{r=1}^{i-1} N_{r\bullet} + 1$ to $\sum_{r=1}^{i} N_{r\bullet}$ and zeros elsewhere. Similarly, let $v_{2,j}$ be the length-N vector with ones in entries $\sum_{s=1}^{j-1} N_{\bullet s} + 1$ to $\sum_{s=1}^{j} N_{\bullet s}$ and zeros elsewhere.

Next, we describe our asymptotic assumptions. First

$$\epsilon_R = \max_i N_{i\bullet}/N \to 0, \quad \text{and} \quad \epsilon_C = \max_j N_{\bullet j}/N \to 0, \tag{2.2}$$

so no single row or column dominates. The average row size can be measured by N/R or by $\sum_i N_{i\bullet}^2/N$; the latter is $E(N_{i\bullet})$ when choosing one of the N data points (i, j, x_{ij}, Y_{ij}) at random (uniformly). Similar formulas hold for the average column size. These average row and column sizes are $o(N)$, because

$$\frac{1}{N^2}\sum_i N_{i\bullet}^2 \leqslant \epsilon_R \to 0, \quad \text{and} \quad \frac{1}{N^2}\sum_j N_{\bullet j}^2 \leqslant \epsilon_C \to 0.$$

We often expect the average row and column sizes, while growing slower than N, should diverge:

$$\begin{aligned}&\min(N/R, N/C) \to \infty, \quad \text{and}\\ &\min\Bigl(\frac{1}{N}\sum_i N_{i\bullet}^2, \frac{1}{N}\sum_j N_{\bullet j}^2\Bigr) \to \infty.\end{aligned}$$

We do not however impose these conditions.

Even for large average row and columns sizes, there can still be numerous new or rare entities with $N_{i\bullet} = 1$ or $N_{\bullet j} = 1$. Our analysis can include such small rows and columns without requiring that they be deleted. When there are covariates x_{ij} we need to rule out degenerate settings where the sample variance of x_{ij} does not grow with N or where it is dominated by a handful of observations. We add some such conditions when we discuss central limit theorems in Section 5.2.

The finite fourth moments $E(a_i^4)$, $E(b_j^4)$ and $E(e_{ij}^4)$ are conveniently described through finite kurtoses κ_A, κ_B and κ_E, respectively. Some of the variance expressions in [11] are dominated by terms proportional to $\kappa+2$ for one of these kurtoses. Following [11] we assume that $\min(\kappa_A, \kappa_B, \kappa_E) > -2$. This lower bound rules out some symmetric binary distributions for a_i, b_j and e_{ij}. Such cases seem unrealistic for our motivating applications.

The randomness in Y_{ij} comes from a_i, b_j and e_{ij}. In some places we combine them into $\eta_{ij} \equiv a_i + b_j + e_{ij}$.

We will make use of certain method of moment estimators $\hat\sigma_A^2$, $\hat\sigma_B^2$ and $\hat\sigma_E^2$ for σ_A^2, σ_B^2 and σ_E^2, from [11]. That paper gives exact finite sample formulas for the variance of those estimators. Then it gives asymptotic variance expressions letting ϵ_R, ϵ_C, R/N and C/N approach zero. The Stitch Fix data that we consider in Section 4 does not have a very small value for R/N. Here we develop non-asymptotic magnitude bounds for bias and variance that do not require R/N and C/N to be close to zero. They need only be bounded away from one.

Theorem 1. *Suppose that* $\max(R/N, C/N) \leqslant \theta$ *for some* $\theta < 1$ *and let* $\epsilon = \max(\epsilon_R, \epsilon_C)$. *Then the moment based estimators from [11] satisfy*

$$\begin{aligned}
E(\hat\sigma_A^2) &= (\sigma_A^2 + \Upsilon)(1 + O(\epsilon)),\\
E(\hat\sigma_B^2) &= (\sigma_B^2 + \Upsilon)(1 + O(\epsilon)), \quad \textit{and}\\
E(\sigma_E^2) &= (\sigma_E^2 + \Upsilon)(1 + O(\epsilon)),
\end{aligned}$$

where

$$\Upsilon \equiv \sigma_A^2 \frac{\sum_i N_{i\bullet}^2}{N^2} + \sigma_B^2 \frac{\sum_j N_{\bullet j}^2}{N^2} + \frac{\sigma_E^2}{N} = O(\epsilon).$$

Furthermore

$$\max\Bigl(\mathrm{Var}(\hat\sigma_A^2), \mathrm{Var}(\hat\sigma_B^2), \mathrm{Var}(\hat\sigma_E^2)\Bigr) = O\Bigl(\frac{\sum_i N_{i\bullet}^2}{N^2} + \frac{\sum_j N_{\bullet j}^2}{N^2}\Bigr) = O(\epsilon).$$

Proof. See Section S1 in the supplement. □

Theorem 1 has the same variance rate for all variance components. In our computed examples $\mathrm{Var}(\sigma_E^2) \ll \min(\mathrm{Var}(\sigma_A^2), \mathrm{Var}(\sigma_B^2))$ because $N \gg \max(R, C)$, a condition not imposed in Theorem 1. Both bias and variance are $O(\epsilon)$ and so a (conservative) effective sample size is then $O(1/\epsilon)$. The quantity Υ appearing in Theorem 1 is $\mathrm{Var}(\bar{Y}_{\bullet\bullet})$ where $\bar{Y}_{\bullet\bullet} = (1/N)\sum_{ij} Z_{ij}Y_{ij}$. The variances of the variance components contain similar quantities to Υ although kurtoses and other quantities appear in their implied constants.

3. An Alternating Algorithm

Our estimation procedure for Model 1 is given in Algorithm 1. We alternate twice between finding $\hat{\beta}$ and the variance component estimates $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, and $\hat{\sigma}_E^2$. One can continue iterating if desired, but our theory shows that two iterations suffice, and in our experience further iterations did not change the estimates much. Further details of these steps, including the way we choose generalized least squares (GLS) estimator to use in step 3, are given in the next two subsections.

The data are a collection of (i, j, x_{ij}, Y_{ij}) tuples. A pass over the data proceeds via iteration over all tuples in the data set. Such a pass may generate $O(R+C)$ intermediate values to be retained for future computations.

Algorithm 1: Alternating Algorithm

1. Estimate β via ordinary least squares (OLS): $\hat{\beta} = \hat{\beta}_{\mathrm{OLS}}$.
2. Let $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, and $\hat{\sigma}_E^2$ be the method of moments estimates from [11] defined on the data $(i, j, \hat{\eta}_{ij})$, where $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}_{\mathrm{OLS}}$.
3. Compute a more efficient $\hat{\beta}$ using $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, and $\hat{\sigma}_E^2$. If $\hat{\sigma}_A^2 \max_i N_{i\bullet} \geqslant \hat{\sigma}_B^2 \max_j N_{\bullet j}$, estimate β via GLS accounting for row correlations: $\hat{\beta} = \hat{\beta}_{\mathrm{RLS}}$. Otherwise, estimate it via GLS accounting for column correlations: $\hat{\beta} = \hat{\beta}_{\mathrm{CLS}}$.
4. Repeat step 2 using $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}$ with $\hat{\beta}$ from step 3.
5. Compute an estimate $\widehat{\mathrm{Var}}(\hat{\beta})$ for $\hat{\beta}$ from step 3 using $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\sigma}_E^2$ from step 4.

3.1 Step by step details for Algorithm 1

Step 1

The first step of Algorithm 1 is to compute the OLS estimate of β. Let $X \in \mathbb{R}^{N\times p}$ have rows x_{ij} in some order and let $Y \in \mathbb{R}^N$ be elements Y_{ij} in the same order. Then,

$$\hat\beta_{\mathrm{OLS}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y = \Bigl(\sum_{ij} Z_{ij}x_{ij}x_{ij}^\mathsf{T}\Bigr)^{-1}\sum_{ij} Z_{ij}x_{ij}Y_{ij}. \tag{3.3}$$

In one pass over the data, we can compute $X^\mathsf{T}X$ and $X^\mathsf{T}Y$ and solve for $\hat\beta$. Solving the normal equations this way is easy to parallelize but more prone to roundoff error than the usual alternative based on computing the SVD of X. The numerical conditioning of the SVD computation is like doubling the number of floating point bits available compared to solving normal equations. One can compensate by solving normal equations in extended precision. It costs $O(p^3)$ to compute $\hat\beta_{\mathrm{OLS}}$ and so the cost of step 1 is $O(Np^2+p^3)$. The space cost is $O(p^2)$.

Step 2

Step 2 uses the algorithm from [11] to compute variance component estimates $\hat\sigma^2_A$, $\hat\sigma^2_B$ and $\hat\sigma^2_E$ in $O(N)$ time and $O(R+C)$ space. A more detailed account is in Section 3.2. This takes $O(Np)$ time to recompute $\hat\eta_{ij}$.

Step 3

GLS estimators: First we define and compare GLS estimators of β accounting for row correlations, or column correlations, or both. These estimators are most easily presented through a reordering of the data. Our algorithm does not have to actually sort the data which would be a major inconvenience in our motivating applications. We work with one row ordering of the data, in which ij precedes $i'j'$ whenever $i<i'$ and with one column ordering of the data. Let P be the $N\times N$ permutation matrix corresponding to the transformation of the column ordering to the row ordering. Let $A_R \in \mathbb{N}^{N\times N}$ be the block diagonal matrix with i'th block $1_{N_{i\bullet}}1_{N_{i\bullet}}^\mathsf{T}$ and $B_C \in \mathbb{N}^{N\times N}$ the block diagonal matrix with j'th block $1_{N_{\bullet j}}1_{N_{\bullet j}}^\mathsf{T}$.

If Y is given in the row ordering, then

$$\mathrm{Cov}(Y) = V_R \equiv \sigma^2_E I_N + \sigma^2_A A_R + \sigma^2_B B_R, \quad \text{for} \quad B_R = PB_CP^\mathsf{T}. \tag{3.4}$$

For Y in the column ordering,

$$\mathrm{Cov}(Y) = V_C \equiv \sigma_E^2 I_N + \sigma_A^2 A_C + \sigma_B^2 B_C, \quad \text{for} \quad A_C = P^\mathsf{T} A_R P. \tag{3.5}$$

GLS algorithms based on (3.4) or (3.5) have computational complexity $O(N^{3/2})$. This is better than the $O(N^3)$ that we might face had V_R or V_C been arbitrary dense matrices, instead of being comprised of the identity and some low rank block diagonal matrices, but it is still too slow for large scale applications.

In a hierarchical model where only row correlations were present we could take $\sigma_B^2 = 0$ and define

$$\hat\beta_{\mathrm{RLS}} = (X^\mathsf{T} \hat V_A^{-1} X)^{-1} X^\mathsf{T} \hat V_A^{-1} Y, \quad \text{for} \quad \hat V_A = \hat\sigma_E^2 I_N + \hat\sigma_A^2 A_R, \tag{3.6}$$

using sample estimates $\hat\sigma_A^2$ and $\hat\sigma_E^2$ of σ_A^2 and σ_E^2. This GLS estimator of β accounts for the intra-row correlations in the data. Similarly, the GLS estimator of β accounting for the intra-column correlations is

$$\hat\beta_{\mathrm{CLS}} = (X^\mathsf{T} \hat V_B^{-1} X)^{-1} X^\mathsf{T} \hat V_B^{-1} Y, \quad \text{for} \quad \hat V_B = \hat\sigma_E^2 I_N + \hat\sigma_B^2 B_C. \tag{3.7}$$

We show next that $\hat\beta_{\mathrm{RLS}}$ and $\hat\beta_{\mathrm{CLS}}$ can be computed in $O(N)$ time.

GLS Computations in $O(N)$ cost: From the Woodbury formula [13] and defining $Z_a \in \{0,1\}^{N\times R}$ as the matrix with ith column $v_{1,i}$ (from Section 2), we have

$$\begin{aligned}
&X^\mathsf{T} \hat V_A^{-1} X \\
&= X^\mathsf{T} (\hat\sigma_E^2 I_N + \hat\sigma_A^2 Z_a {Z_a}^\mathsf{T})^{-1} X \\
&= \frac{X^\mathsf{T} X}{\hat\sigma_E^2} - \frac{\hat\sigma_A^2}{\hat\sigma_E^2} X^\mathsf{T} Z_a \operatorname{diag}\Bigl(\frac{1}{\hat\sigma_E^2 + \hat\sigma_A^2 N_{i\bullet}}\Bigr) Z_a^\mathsf{T} X \\
&= \frac{1}{\hat\sigma_E^2} \sum_{ij} Z_{ij} x_{ij} x_{ij}^\mathsf{T} - \frac{\hat\sigma_A^2}{\hat\sigma_E^2} \sum_i \frac{1}{\hat\sigma_E^2 + \hat\sigma_A^2 N_{i\bullet}} \Bigl(\sum_j Z_{ij} x_{ij}\Bigr) \Bigl(\sum_j Z_{ij} x_{ij}\Bigr)^\mathsf{T}.
\end{aligned}$$

Likewise, $X^\mathsf{T} \hat V_A^{-1} Y$ equals

$$\frac{1}{\hat\sigma_E^2} \sum_{ij} Z_{ij} x_{ij} Y_{ij} - \frac{\hat\sigma_A^2}{\hat\sigma_E^2} \sum_i \frac{1}{\hat\sigma_E^2 + \hat\sigma_A^2 N_{i\bullet}} \Bigl(\sum_j Z_{ij} x_{ij}\Bigr) \Bigl(\sum_j Z_{ij} Y_{ij}\Bigr).$$

One pass over the data allows us to compute $\sum_{ij} Z_{ij} x_{ij} x_{ij}^\mathsf{T}$ and $\sum_{ij} Z_{ij} x_{ij} Y_{ij}$, as well as $N_{i\bullet}$, and the row sums $\sum_j Z_{ij} x_{ij}$ and $\sum_j Z_{ij} Y_{ij}$ for

$i = 1, \dots, R$. The cost is $O(Np^2)$ time and $O(Rp)$ space. None of these quantities require us to sort the data. We then compute $X^\mathsf{T}\hat{V}_A^{-1}X$ and $X^\mathsf{T}\hat{V}_A^{-1}Y$ in time $O(Rp^2)$. Then, $\hat{\beta}_{\mathrm{RLS}}$ is computed in $O(p^3)$. Hence, $\hat{\beta}_{\mathrm{RLS}}$ can be found within $O(Rp)$ space and $O(Np^2+p^3) = O(Np^2)$ time. Clearly $\hat{\beta}_{\mathrm{CLS}}$ costs $O(Cp)$ space and $O(Np^2)$ time.

Efficiencies: We can compute either $\hat{\beta}_{\mathrm{RLS}}$ or $\hat{\beta}_{\mathrm{CLS}}$ in our computational budget. We will choose RLS if the variance component associated with rows is dominant and CLS otherwise. The choice could be made dependent on X but in many applications one considers numerous different X matrices and we prefer to have a single choice for all regressions. Accordingly, we find a lower bound on the efficiency of RLS when X is a single nonzero vector $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$. We choose RLS if that lower bound is higher than the corresponding bound for CLS, in this $p = 1$ setting.

The full GLS estimator is $\hat{\beta}_{\mathrm{GLS}} = (X^\mathsf{T}V_R^{-1}X)^{-1}X^\mathsf{T}V_R^{-1}Y$ when the data are ordered by rows and $(X^\mathsf{T}V_C^{-1}X)^{-1}X^\mathsf{T}V_C^{-1}Y$ when the data are ordered by columns. For data ordered by rows, the efficiency of $\hat{\beta}_{\mathrm{RLS}}$ is

$$\mathrm{eff}_{\mathrm{RLS}} = \frac{\mathrm{Var}(\hat{\beta}_{\mathrm{GLS}})}{\mathrm{Var}(\hat{\beta}_{\mathrm{RLS}})} = \frac{(\boldsymbol{x}^\mathsf{T}V_A^{-1}\boldsymbol{x})^2}{(\boldsymbol{x}^\mathsf{T}V_A^{-1}V_RV_A^{-1}\boldsymbol{x})(\boldsymbol{x}^\mathsf{T}V_R^{-1}\boldsymbol{x})}. \tag{3.8}$$

For data ordered by columns, the corresponding efficiency of $\hat{\beta}_{\mathrm{CLS}}$ is

$$\mathrm{eff}_{\mathrm{CLS}} = \frac{\mathrm{Var}(\hat{\beta}_{\mathrm{GLS}})}{\mathrm{Var}(\hat{\beta}_{\mathrm{CLS}})} = \frac{(\boldsymbol{x}^\mathsf{T}V_B^{-1}\boldsymbol{x})^2}{(\boldsymbol{x}^\mathsf{T}V_B^{-1}V_CV_B^{-1}\boldsymbol{x})(\boldsymbol{x}^\mathsf{T}V_C^{-1}\boldsymbol{x})}. \tag{3.9}$$

The next two theorems establish lower bounds on these efficiencies.

Theorem 2. *Let A be a positive definite Hermitian matrix and $\boldsymbol{u}$ be a unit vector. If the eigenvalues of A are bounded below by $m > 0$ and above by $M < \infty$, then*

$$(\boldsymbol{u}^\mathsf{T}A\boldsymbol{u})(\boldsymbol{u}^\mathsf{T}A^{-1}\boldsymbol{u}) \leqslant \frac{(m+M)^2}{4mM}.$$

Equality may hold, for example when $\boldsymbol{u}^\mathsf{T}A\boldsymbol{u} = (M+m)/2$ and the only roots of A are m and M.

Proof. This is Kantorovich's inequality [21]. □

By two applications of Theorem 2 on (3.8) and (3.9) we prove:

Theorem 3. *For $p = 1$ and $\sigma_E^2 > 0$, let* $\mathrm{eff}_{\mathrm{RLS}}$ *and* $\mathrm{eff}_{\mathrm{CLS}}$ *be defined as in* (3.8) *and* (3.9). *Then*

$$\mathrm{eff}_{\mathrm{RLS}} \geqslant \frac{4\sigma_E^2(\sigma_E^2 + \sigma_B^2 \max_j N_{\bullet j})}{(2\sigma_E^2 + \sigma_B^2 \max_j N_{\bullet j})^2} \quad \textit{and}$$
$$\mathrm{eff}_{\mathrm{CLS}} \geqslant \frac{4\sigma_E^2(\sigma_E^2 + \sigma_A^2 \max_i N_{i\bullet})}{(2\sigma_E^2 + \sigma_A^2 \max_i N_{i\bullet})^2}.$$

Both inequalties are tight.

Proof. See Section S2.1 in the supplement. □

After some algebra, we see that the worst case efficiency of $\hat{\beta}_{\mathrm{RLS}}$ is higher than that of $\hat{\beta}_{\mathrm{CLS}}$ when $\sigma_A^2 \max_i N_{i\bullet} > \sigma_B^2 \max_j N_{\bullet j}$. We set $\hat{\beta}$ to be $\hat{\beta}_{\mathrm{RLS}}$ when $\hat{\sigma}_A^2 \max_i N_{i\bullet} \geqslant \hat{\sigma}_B^2 \max_j N_{\bullet j}$, and $\hat{\beta}_{\mathrm{CLS}}$ otherwise.

Optimizing a lower bound does not necessarily optimize the quantity of interest, and so we expect that our choice here is not the only reasonable one. The efficiency of $\hat{\beta}_{\mathrm{RLS}}$ depends only on the ratio $\hat{\sigma}_A^2/\hat{\sigma}_E^2$ in use. We investigated GLS estimators of β based on $\hat{V}_A = \hat{\sigma}_A^2 A_R + (\hat{\sigma}_E^2 + \lambda\hat{\sigma}_B^2)I_N$ for λ chosen by the Kantorovich inequality. It did not appear to bring an improved accuracy over our default choice in some simulations. In practice, one can also compute both $\hat{\beta}_{\mathrm{RLS}}$ and $\hat{\beta}_{\mathrm{CLS}}$ and compare $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{RLS}})$ and $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{CLS}})$.

Steps 4 and 5

Step 4 is just like step 2 and it costs $O(Np)$ time. Step 5 is described in Section 5.3 where we derive $\mathrm{Var}(\hat{\beta}_{\mathrm{RLS}})$ and $\mathrm{Var}(\hat{\beta}_{\mathrm{CLS}})$.

3.2 Method of Moments (Steps 2 and 4)

In this subsection, we discuss steps 2 and 4 of Algorithm 1 in more detail. The errors $Y_{ij} - x_{ij}^{\mathsf{T}}\beta$ follow a two-factor crossed random effects model [11]. If $\hat{\beta}$ is a good estimate of β, then the residuals $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}$ approximately follow a two-factor crossed random effects model with $\mu = 0$ and variance components σ_A^2, σ_B^2, and σ_E^2.

We estimate σ_A^2, σ_B^2, and σ_E^2, with the algorithm from [11] with data $(i, j, \hat{\eta}_{ij})$. That algorithm gives unbiased estimates of the variance components in a two-factor crossed random effects model.

The algorithm of [11] applies the method of moments to three statistics; a weighted sum of within-row sample variances, a weighted sum of within-column sample variances, and a multiple of the full sample variance. For Algorithm 1, these are:

$$\begin{aligned}
U_a(\hat{\beta}) &= \sum_i S_{i\bullet}, \quad S_{i\bullet} = \sum_j Z_{ij}(\hat{\eta}_{ij} - \hat{\eta}_{i\bullet})^2 \\
U_b(\hat{\beta}) &= \sum_j S_{\bullet j}, \quad S_{\bullet j} = \sum_i Z_{ij}(\hat{\eta}_{ij} - \hat{\eta}_{\bullet j})^2, \quad \text{and} \\
U_e(\hat{\beta}) &= \sum_{ij} Z_{ij}(\hat{\eta}_{ij} - \hat{\eta}_{\bullet\bullet})^2,
\end{aligned} \tag{3.10}$$

where subscripts replaced by $\bullet$ are averaged over. The variance component estimates are obtained by solving the system

$$M \begin{pmatrix} \hat{\sigma}^2_A \\ \hat{\sigma}^2_B \\ \hat{\sigma}^2_E \end{pmatrix} = \begin{pmatrix} U_a(\hat{\beta}) \\ U_b(\hat{\beta}) \\ U_e(\hat{\beta}) \end{pmatrix}, \ M = \begin{pmatrix} 0 & N-R & N-R \\ N-C & 0 & N-C \\ N^2 - \sum_i N_{i\bullet}^2 & N^2 - \sum_j N_{\bullet j}^2 & N^2 - N \end{pmatrix}. \tag{3.11}$$

The matrix M is nonsingular under very weak conditions. It suffices to have $R \geqslant 2$, $C \geqslant 2$, $\epsilon_R \leqslant 1/2$ and $\epsilon_C \leqslant 1/2$ [11, Section 4.1].

Gao and Owen [11] compute the U-statistics in one pass over the data taking $O(N)$ time and $O(R+C)$ space. Solving (3.11) takes constant time. Thus, steps 2 and 4 each have computational complexity $O(N)$ and space complexity $O(R+C)$.

4. Stitch Fix rating data

Stitch Fix sells clothing, mostly women's clothing. They mail their clients a sample of clothing items. A client keeps and purchases some items and returns the others. It is important to predict which items a client will like. In the context of our model, client i might get item j and then rate that item with a score Y_{ij}.

Stitch Fix has provided us some of their client ratings data. This data is fully anonymized and void of any personally identifying information. The data provided by Stitch Fix is a sample of their data, and consequently does not reflect their actual numbers of clients, items or their ratios, for example. Nonetheless this is an interesting data set with which to illustrate a linear mixed effects model.

We received data on clients' ratings of items they received, as well as the following information about the clients and items. For client i and item j, the response is a composite rating Y_{ij} on a scale from 1 to 10. There was a categorical variable giving the item's material, with 23 categories. We also received a binary variable indicating whether the item style is considered to be 'edgy', and another one on whether the client likes edgy styles. Similarly, there was another pair of binary variables indicating whether items were labeled 'boho' (Bohemian) and whether the client likes boho items. Finally, there was a match score. That is an estimate of the probability that the client keeps the item, predicted before it is actually sent. The match score is a prediction from a baseline model and is not representative of all algorithms used at Stitch Fix.

The observation pattern in the data is as follows. We received $N = 5{,}000{,}000$ ratings on $C = 6{,}318$ items by $R = 762{,}752$ clients. Thus $C/N \doteq 0.00126$ and $R/N \doteq 0.153$. The latter ratio indicates that only a relatively small number of ratings from each client are included in the data (their full shipment history is not included in the sampled data). The data are not dominated by a single row or column because $\epsilon_R \doteq 9\times 10^{-6}$ and $\epsilon_C \doteq 0.0143$. Similarly

$$\frac{N}{\sum_i N_{i\bullet}^2} \doteq 0.103, \qquad \frac{\sum_i N_{i\bullet}^2}{N^2} \doteq 1.95 \times 10^{-6},$$
$$\frac{N}{\sum_j N_{\bullet j}^2} \doteq 1.22 \times 10^{-4}, \quad \text{and} \qquad \frac{\sum_j N_{\bullet j}^2}{N^2} \doteq 0.00164.$$

Our two-factor linear mixed effects model for this data is:

Model 2. *For client i and item j,*

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

Here material_{ij} *is a categorical variable that is implemented via indicator variables for each type of material. We chose 'Polyester', the most common material, to be the baseline.*

In a regression analysis, Model 2 would be only one of many models one might consider. There would be numerous ways to encode the variables, and

Table 1: Stitch Fix Regression Results (omitting material type).

	$\hat\beta_{\mathrm{OLS}}$	$\widehat{\mathrm{se}}_{\mathrm{OLS}}(\hat\beta_{\mathrm{OLS}})$	$\widehat{\mathrm{se}}(\hat\beta_{\mathrm{OLS}})$	$\hat\beta$	$\widehat{\mathrm{se}}(\hat\beta)$
Intercept	4.635^*	0.005397	0.05808	5.110^*	0.01250
Match	5.048^*	0.01174	0.1464	3.529^*	0.02153
$\mathbb{I}\{$client edgy$\}$	0.001020	0.002443	0.004593	0.001860	0.003831
$\mathbb{I}\{$item edgy$\}$	-0.3358^*	0.004253	0.03730	-0.3328^*	0.01542
$\mathbb{I}\{$both edgy$\}$	0.3925^*	0.006229	0.01352	0.3864^*	0.006432
$\mathbb{I}\{$client boho$\}$	0.1386^*	0.002264	0.004354	0.1334^*	0.003622
$\mathbb{I}\{$item boho$\}$	-0.5499^*	0.005981	0.03049	-0.6261^*	0.01661
$\mathbb{I}\{$both boho$\}$	0.3822^*	0.007566	0.01057	0.3837^*	0.007697

the coefficients in any one model would depend on which other variables were included. The odds of settling on exactly this model are low. Our focus is on the estimated standard errors due to variance components and so we will work with a naive face-value interpretation of the coefficients β_j in Model 2. If the emphasis is on prediction, then one can use $x_{ij}^{\mathsf{T}}\hat\beta$ perhaps adding shrunken row and/or column means of the residuals. See [11] for a discussion of how estimates of σ_A^2, σ_B^2, and σ_E^2 can be used to shrink row and/or column means in the intercept-only setting. Even in prediction, underestimating the uncertainty in $\hat\beta$ could be costly.

Suppose that one ignored client and item random effects and simply ran OLS. Table 1 shows the results for all coefficients except the material type indicator variables. Section S4 of the Supplement has the complete table. The next column has $\widehat{\mathrm{se}}_{\mathrm{OLS}}(\hat\beta_{\mathrm{OLS}})$, the standard error that OLS produces for the OLS coefficient estimate. Then $\widehat{\mathrm{se}}_{\mathrm{Mom}}(\hat\beta_{\mathrm{OLS}})$ reports a moment-based standard error for $\hat\beta_{\mathrm{OLS}}$ using the estimated variance components. The next two columns are the method of moments estimator $\hat\beta_{\mathrm{Mom}}$ and its own standard error $\widehat{\mathrm{se}}_{\mathrm{Mom}}(\hat\beta_{\mathrm{Mom}})$ based on the variance components.

Figure 1 has a graphical presentation of these results. The leftmost panel shows that ignoring the random effects greatly underestimates the uncertainty in the regression coefficients. The underestimation can be tens or even hundreds-fold when interpreted via effective sample sizes. Even if N is 'big data', $N/100$ might not be. The right panel shows that properly accounting for uncertainty makes a great many material indicator variables change from significant to not significant if one were using a threshold of $|t| \geqslant 2$. In other words the difference in effective sample size could have left the user of this model with stronger conclusions and different decisions

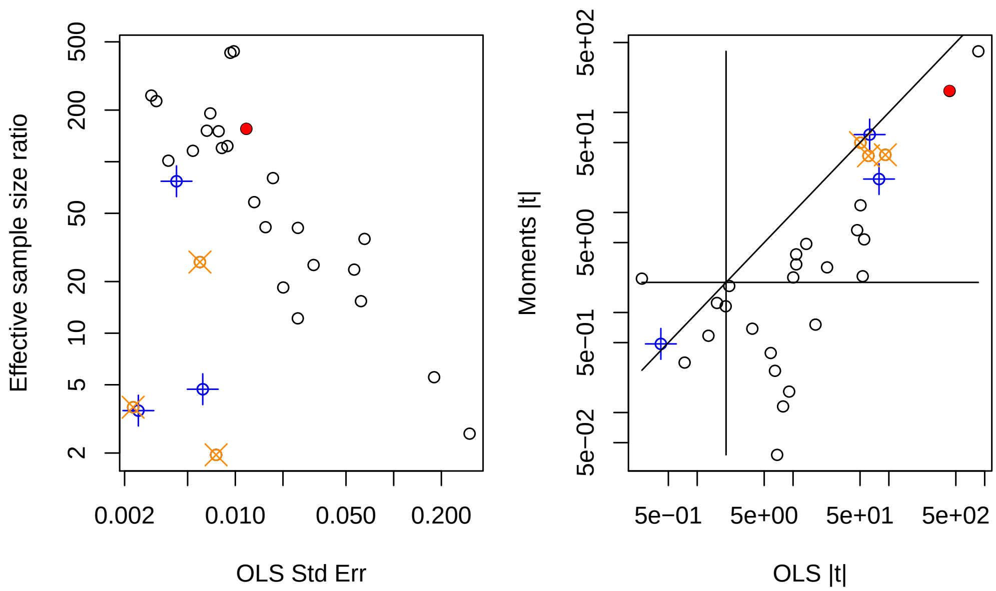

Figure 1: The left panel shows how OLS underestimates variance. The horizontal axis is $\widehat{\mathrm{se}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})$. The vertical axis is $[\widehat{\mathrm{se}}_{\mathrm{Mom}}(\hat{\beta}_{\mathrm{OLS}})/\widehat{\mathrm{se}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}})]^2$, which we interpret as the extent to which OLS overestimates the effective sample size. The right panel plots absolute t statistics, $|\hat{\beta}_{\mathrm{Mom}}|/\widehat{\mathrm{se}}(\beta_{\mathrm{Mom}})$ versus $|\hat{\beta}_{\mathrm{OLS}}|/\widehat{\mathrm{se}}(\beta_{\mathrm{OLS}})$. There are references lines at $|t| = 2$ and at 45 degrees. The Match variable is plotted as a solid circle. The edgy variables include a + and the boho variables include a $\times$. Material types have open circles.

than they should have had. It is likely that industry uses more elaborate models than our simple regression, but a lower than anticipated effective sample size will remain an issue.

The estimated variance components are $\hat\sigma_A^2 = 1.133$, $\hat\sigma_B^2 = 0.1463$, and $\hat\sigma_E^2 = 4.474$. Their standard errors are approximately 0.0046, 0.00089, and 0.0050 respectively, so these components are well determined. The error variance component is largest, and the client effect dominates the item effect by almost a factor of eight.

The 'Match' variable is significantly positively associated with rating, indicating that the baseline prediction provided by Stitch Fix is a useful predictor in this data set. However the random effects model reduces its coefficient from about 5 to about 3.5, a change that is quite a large number of estimated standard errors. We have seen that some clients tend to give higher ratings on average than others. That is, client indicator variables take away some of the explanatory power of the match variable.

Shipping an edgy item to a client who does not like edgy styles is associated with a rating decrease of about 0.33 points, but shipping such an item to a client who does like edgy styles is associated with a small increase in rating.

The boho indicator variable also has a negative overall estimated coefficient $\hat\beta_6 < 0$. The modeled impact of a boho item sent to a boho client is $\hat\beta_5+\hat\beta_6+\hat\beta_7 < 0$, unlike the positive result we saw for sending and edgy item to an edgy client. This suggests that it is more difficult to make matches for boho items. Perhaps there is an interaction where 'boho to boho' has a positive impact for a sufficiently high value of the match variable. For large data sets, such an interaction can be conveniently handled by filtering the data to cases with $\mathrm{Match}_{ij} \geqslant t$ and refitting. We did so but did not find a threshold that yielded $\hat\beta_5 + \hat\beta_6 + \hat\beta_7 > 0$.

Of the materials, 'Cotton', 'Faux Fur', 'Leather', 'Modal', 'Pleather', 'PU', 'PVC', 'Silk', 'Spandex', and 'Tencel' are significantly different from the baseline, 'Polyester' in our crossed random effects model. 'PU' and 'PVC' are associated with an increase in rating of at least half a point. Those materials are often used to make shoes and specialty clothing, which may be related to their association with high ratings.

The computations in this section were done in Python; code is available at `https://github.com/kxgao/scalable-crossed-mixed-effects`.

5. Asymptotic behavior

Here we give sufficient conditions to ensure that the parameter estimates $\hat\beta$, $\hat\sigma_A^2$, $\hat\sigma_B^2$, and $\hat\sigma_E^2$ obtained from Algorithm 1 are consistent. We also give a central limit theorem for $\hat{\beta}$. We use the sample size growth conditions from Section 2 and some additional conditions on x_{ij}. Our results are conditional on the observed predictors x_{ij} for which $Z_{ij} = 1$.

As in ordinary IID error regression problems, our central limit theorem requires the information in the observed x_{ij} to grow quickly in every projection while also imposing a limit on the largest x_{ij}. For each i with $N_{i\bullet} > 0$, let $\bar{x}_{i\bullet}$ be the average of those x_{ij} with $Z_{ij} = 1$ and similarly define column averages $\bar{x}_{\bullet j}$.

For a symmetric positive semi-definite matrix V, let $\mathcal{I}(V)$ be the smallest eigenvalue of V. We will need lower bounds on $\mathcal{I}(V)$ for various V to rule out singular or nearly singular designs. Some of those V involve centered variables. In most applications x_{ij} will include an intercept term, and so we assume that the first component of every x_{ij} equals 1. That term raises some technical difficulties as centering that component always yields zero. We will treat that term specially in some of our proofs. For a symmetric matrix $V \in \mathbb{R}^{p\times p}$, we let

$$\mathcal{I}_0(V) = \mathcal{I}((V_{ij})_{2\leqslant i,j\leqslant p})$$

be the smallest eigenvalue of the lower $(p-1)\times(p-1)$ submatrix of V.

In our motivating applications, it is reasonable to assume that $\|x_{ij}\|$ are uniformly bounded. We let

$$M_N \equiv \max_{ij} Z_{ij}\|x_{ij}\|^2 \tag{5.12}$$

quantify the largest x_{ij} in the data so far. Some of our results would still hold if we were to let M_N grow slowly with N. To focus on the essential ideas, we simply take $M_N \leqslant M_\infty < \infty$ for all N.

5.1 Consistency

First, we give conditions under which $\hat{\beta}_{\mathrm{OLS}}$ from step 1 is consistent.

Theorem 4. *Let* $\max(\epsilon_R, \epsilon_C) \to 0$ *and* $\mathcal{I}(X^\mathsf{T}X) \geqslant cN$ *for some* $c > 0$, *as* $N \to \infty$. *Then* $E(\|\hat{\beta}_{\mathrm{OLS}} - \beta\|^2) = O((\epsilon_R + \epsilon_C)/\mathcal{I}(X^\mathsf{T}X/N)) \to 0$ *and* $\hat{\beta}_{\mathrm{OLS}} \xrightarrow{p} \beta$.

Proof. See Section S3.1 in the supplement. □

Second, we show that the variance component estimates computed in step 2 are consistent. Recall that we compute the U-statistics (3.10) on data $(i, j, \hat{\eta}_{ij} = Y_{ij} - x_{ij}^\mathsf{T}\hat{\beta})$ and use them to obtain estimates $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$, and $\hat{\sigma}^2_E$ via (3.11).

Theorem 5. *Suppose that as* $N \to \infty$ *that* $\max(\epsilon_R, \epsilon_C) \to 0$, $\max(R, C)/N \leqslant \theta \in (0,1)$, $\hat\beta \xrightarrow{p} \beta$, *and that* M_N *is bounded. Then* $\hat\sigma_A^2 \xrightarrow{p} \sigma_A^2$, $\hat\sigma_B^2 \xrightarrow{p} \sigma_B^2$, *and* $\hat\sigma_E^2 \xrightarrow{p} \sigma_E^2$.

Proof. See Section S3.2 in the supplement. □

From Theorem 4, the estimate of β obtained in step 1 of Algorithm 1 is consistent. Therefore, from Theorem 5, the variance component estimates obtained in step 2 are consistent, given the combined assumptions of those two theorems. The proof of Theorem 4 shows that the estimated variance components differ by $O(\|\hat\beta - \beta\|^2 + \epsilon\|\hat\beta - \beta\|)$ from what we would get replacing $\hat\beta$ by an oracle value β and computing variance components of $Y_{ij} - x_{ij}^{\mathsf{T}}\beta$. Such an estimate would have mean squared error $O(\epsilon)$ by Theorem 1. As a result the mean squared error for all parameters of interest is $O(\epsilon)$.

Our third result shows that the estimate of β obtained in step 3 is consistent. We do so by showing that estimators $\hat\beta_{\mathrm{RLS}}$ and $\hat\beta_{\mathrm{CLS}}$ are consistent when constructed using consistent variance component estimates. We give the version for $\hat\beta_{\mathrm{RLS}}$.

Theorem 6. *Let* $\hat\beta_{\mathrm{RLS}}$ *be computed with* $\hat\sigma_A^2 \xrightarrow{p} \sigma_A^2$ *and* $\hat\sigma_E^2 \xrightarrow{p} \sigma_E^2$ *as* $N \to \infty$, *where* $\sigma_E^2 > 0$. *If* $\max(\epsilon_R, \epsilon_C) \to 0$ *and,*

$$\mathcal{I}_0\Bigl(\sum_{ij} Z_{ij}(x_{ij} - \bar x_{i\bullet})(x_{ij} - \bar x_{i\bullet})^{\mathsf{T}}/N\Bigr) \geqslant c > 0 \tag{5.13}$$

and

$$\frac{1}{R^2}\sum_{ir}(ZZ^{\mathsf{T}})_{ir} N_{i\bullet}^{-1} N_{r\bullet}^{-1} \to 0, \tag{5.14}$$

then $\hat\beta_{\mathrm{RLS}} \xrightarrow{p} \beta$.

Proof. See Section S3.3 in the supplement. □

The most complicated part of the proof of Theorem 6 involves handling the contribution of b_j to $\hat\beta_{\mathrm{RLS}}$. In row weighted GLS it is quite standard to have random errors a_i and e_{ij} but here we must also contend with errors b_j that do not appear in the model for which $\hat\beta_{\mathrm{RLS}}$ is the MLE. Condition (5.14) is used to control the variance contribution of the column random effects to the intercept in $\hat\beta_{\mathrm{RLS}}$. For balanced data it reduces to $1/C \to 0$ and so it has an effective number of columns interpretation. Recalling that $(ZZ^{\mathsf{T}})_{ir}$ is the number of columns sampled in both rows i and r, we have $(ZZ^\mathsf{T})_{ir} \leqslant N_{r\bullet}$ and so a sufficient condition for (5.14) is that $(1/R)\sum_i N_{i\bullet}^{-1} \to 0$. For sparsely observed data we expect $(ZZ^\mathsf{T})_{ir} \ll \max(N_{i\bullet}, N_{r\bullet})$ to be typical, and then these bounds are conservative.

Any realistic setting will have $\sigma_E^2 > 0$ and we need $\sigma_E^2 > 0$ for $\hat\beta_{\mathrm{RLS}}$ to be well defined. So that condition in Theorem 6 is not restrictive.

It remains to show that the variance component estimates from step 4 are consistent. We can just apply Theorem 5 again. Therefore the final estimates returned by Algorithm 1 are consistent given only weak conditions on the behavior of Z_{ij} and x_{ij}.

5.2 Asymptotic Normality of $\hat\beta$

Here we show that the estimator $\hat\beta_{\mathrm{RLS}}$ constructed using consistent estimates of σ_A^2, σ_B^2, and σ_E^2 is asymptotically Gaussian, which implies that $\hat\beta_{\mathrm{CLS}}$ and thus β are as well by symmetry. We need stronger conditions than we needed for consistency.

These conditions are expressed in terms of some weighted means of the predictors. First, let

$$\tilde x_{\bullet j} = \frac{1}{N_{\bullet j}}\sum_i Z_{ij}\frac{\sigma_A^2}{\sigma_A^2+\sigma_E^2/N_{i\bullet}}\bar x_{i\bullet}. \tag{5.15}$$

This is a ‘second order’ average of x for column j: it is the average over rows i that intersect j, of averages $\bar x_{i\bullet}$ shrunken towards zero. For a balanced design with $Z_{ij} = 1_{i\leqslant R}1_{j\leqslant C}$ we would have $\tilde x_{\bullet j} = \bar x_{\bullet\bullet}\sigma_A^2/(\sigma_A^2+\sigma_E^2/C)$, so then the second order means would all be very close to $\bar x_{\bullet\bullet}$ for large C. Apart from the shrinkage, we can think of $\tilde x_{\bullet j}$ as a local version of $\bar x_{\bullet\bullet}$ appropriate to column j. Next let

$$k = \sum_j N_{\bullet j}^2(\bar x_{\bullet j} - \tilde x_{\bullet j}) \Bigm/ \sum_j N_{\bullet j}^2 \in \mathbb{R}^p. \tag{5.16}$$

This is a weighted sum of adjusted columns means, weighted by the squared column size. The intercept component of this k will not be used.

Theorem 7. *Let $\hat\beta_{\mathrm{RLS}}$ be computed with $\hat\sigma_A^2 \xrightarrow{p} \sigma_A^2$, $\hat\sigma_B^2 \xrightarrow{p} \sigma_B^2$, and $\hat\sigma_E^2 \xrightarrow{p} \sigma_E^2 > 0$ as $N\to\infty$. Suppose also that*

$$\mathcal{I}\biggl(\sum_i \bar x_{i\bullet}\bar x_{i\bullet}^\mathsf{T}\biggr),\quad \mathcal{I}_0\biggl(\sum_{ij} Z_{ij}(x_{ij}-\bar x_{i\bullet})(x_{ij}-\bar x_{i\bullet})^\mathsf{T}\biggr),\quad \textit{and}$$

$$\mathcal{I}_0\biggl(\sum_j N_{\bullet j}^2(\bar x_{\bullet j}-\tilde x_{\bullet j}-k)(\bar x_{\bullet j}-\tilde x_{\bullet j}-k)^\mathsf{T}\biggr) \Bigm/ \max_j N_{\bullet j}^2$$

all tend to infinity, where $\tilde{x}_{\bullet j}$ *is given by* (5.15) *and* k *is given by* (5.16). *Next for* $c_j = \sum_i Z_{ij}\sigma_E^2/(\sigma_E^2 + \sigma_A^2 N_{i\bullet})$ *and* $c_{ij} = \sigma_E^2/(\sigma_E^2 + \sigma_A^2 N_{i\bullet})$ *assume that both* $\max_j c_j^2/\sum_j c_j^2$ *and* $\max_{ij} c_{ij}^2/\sum_{ij} c_{ij}^2$ *converge to zero. Then* $\hat{\beta}_{\mathrm{RLS}}$ *is asymptotically distributed as*

$$\mathcal{N}(\beta, (X^\mathsf{T} V_A^{-1} X)^{-1} X^\mathsf{T} V_A^{-1} V_R V_A^{-1} X (X^\mathsf{T} V_A^{-1} X)^{-1}). \tag{5.17}$$

Proof. See Section S3.4 in the supplement. □

The statement that $\hat{\beta}_{\mathrm{RLS}}$ has asymptotic distribution $\mathcal{N}(\beta, V)$ is shorthand for $V^{-1/2}(\hat{\beta} - \beta) \xrightarrow{p} \mathcal{N}(0, I_p)$.

Theorem 7 imposes three information criteria. First, the R rows i with $N_{i\bullet} > 0$ must have sample average $\bar{x}_{i\bullet}$ vectors with information tending to infinity. It would be reasonable to expect that information to be proportional to R and also reasonable to require $R \to \infty$ for a CLT. Next, the sum of within row sums of squares and cross products of row-centered x_{ij} must have growing information, apart from the intercept term. Finally, thinking of $\bar{x}_{\bullet j} - \tilde{x}_{\bullet j}$ as the locally centered mean for column j, those quantities centered on the vector k must have a weighted sum of squares that is not dominated by any single column when weights proportional to $N_{\bullet j}^2$ are applied.

The conditions on c_j and c_{ij} are used to show that the CLT will apply to the intercept in the regression. The condition on $\max_j c_j^2/\sum_j c_j^2$ will fail if for example column $j = 1$ has half of the N observations, all in rows of size $N_{i\bullet} = 1$. In the case of an $R \times C$ grid $\max_j c_j^2/\sum_j c_j^2 = 1/C$ and so we can interpret this condition as requiring a large enough effective number $\sum_j c_j^2/\max_j c_j^2$ of columns in the data.

The condition on $\max_{ij} c_{ij}^2/\sum_{ij} c_{ij}^2$ will fail if for example the data contain a full $R \times C$ grid of values plus a single observation with $i = R+1$ and $j = C+1$. The problem is that in a row based regression, a single small row can get outsized leverage. It can be controlled by dropping relatively small rows. This pruning of rows is only used for the CLT to apply to the intercept term. It is not needed for other components of β nor is it needed for consistency. We do not know if it is necessary for the CLT.

5.3 Computing $\mathrm{Var}(\hat\beta_{\mathrm{RLS}})$

Here we show how to compute the estimate of the asymptotic variance of $\hat\beta_{\mathrm{RLS}}$ from Theorem 7. First,

$$\begin{aligned}
&(X^\mathsf{T}V_A^{-1}X)^{-1}X^\mathsf{T}V_A^{-1}V_RV_A^{-1}X(X^\mathsf{T}V_A^{-1}X)^{-1}\\
&= (X^\mathsf{T}V_A^{-1}X)^{-1}X^\mathsf{T}V_A^{-1}(V_A+\sigma_B^2B_R)V_A^{-1}X(X^\mathsf{T}V_A^{-1}X)^{-1}\\
&= (X^\mathsf{T}V_A^{-1}X)^{-1} + (X^\mathsf{T}V_A^{-1}X)^{-1}X^\mathsf{T}V_A^{-1}\sigma_B^2B_RV_A^{-1}X(X^\mathsf{T}V_A^{-1}X)^{-1}\\
&= (X^\mathsf{T}V_A^{-1}X)^{-1} + (X^\mathsf{T}V_A^{-1}X)^{-1}\mathrm{Var}(X^\mathsf{T}V_A^{-1}b)(X^\mathsf{T}V_A^{-1}X)^{-1}, \qquad (5.18)
\end{aligned}$$

where b is the length-N vector of column random effects for each observation. That is b_j appears $N_{\bullet j}$ times in b.

Using the Woodbury formula we find that $\mathrm{Var}(X^\mathsf{T}V_A^{-1}b)$ equals

$$\frac{\sigma_B^2}{\sigma_E^4}\sum_j\Bigl(X_{\bullet j}-\sigma_A^2\sum_i Z_{ij}\frac{X_{i\bullet}}{\sigma_E^2+\sigma_A^2N_{i\bullet}}\Bigr)\Bigl(X_{\bullet j}-\sigma_A^2\sum_i Z_{ij}\frac{X_{i\bullet}}{\sigma_E^2+\sigma_A^2N_{i\bullet}}\Bigr)^\mathsf{T}. \qquad (5.19)$$

Recall that $X_{i\bullet}$ and $X_{\bullet j}$ are row and column totals, not means.

In practice, we plug consistent estimates $\hat\sigma_A^2$, $\hat\sigma_B^2$, and $\hat\sigma_E^2$ in for σ_A^2, σ_B^2, and σ_E^2 in (5.18) and (5.19). We already have $(X^\mathsf{T}\hat V_A^{-1}X)^{-1}$ as well as $N_{i\bullet}$ and $X_{i\bullet}$ for $i=1,\dots,R$ available from computing $\hat\beta_{\mathrm{RLS}}$. In a new pass over the data, we compute $X_{\bullet j}$ and $\sum_i Z_{ij}X_{i\bullet}$ for $j=1,\dots,C$, incurring $O(Np)$ computational and $O(Cp)$ storage cost. Then, (5.19) can be found in $O(Cp^2)$ time; a final step finds (5.18) in $O(p^3)$ time. Overall, estimating the variance of $\hat\beta_{\mathrm{RLS}}$ requires $O(Np+Cp^2+p^3)$ additional computation time and $O(Cp)$ additional space.

6. Comparisons to the MLE

Here we compare Algorithm 1 to maximum likelihood for a linear mixed effects model, looking at both computational efficiency and statistical efficiency. We use a state of the art code for linear mixed models called MixedModels [2]. This is written in Julia [5] and is much faster than other linear mixed model code we have tried.

Our examples have $R=C=2\sqrt{N}$ for various N. We create an $R\times C$ matrix of Z_{ij} and randomly choose exactly $RC/4$ components to be 1. We have an intercept and p other x's with $x_{ij,t}\stackrel{\mathrm{iid}}{\sim}\mathcal{N}(0,1)$, for $2\leqslant t\leqslant p+1$. We use all $p\in\{1,5,10,20\}$. We take $\sigma_A^2=2$, $\sigma_B^2=1/2$, $\sigma_E^2=1$ and all $\beta_j=1$. Our simulated random effects and our noise are all Gaussian because we are comparing to code that computes a Gaussian MLE.

6.1 Computational cost

The Julia package MixedModels package uses a derivative-free optimization method from the BOBYQA package [25]. At each iteration it evaluates the log likelihood at a set of points, fits a quadratic function to those points and minimizes the quadratic. The number of likelihood evaluations per iteration is fixed, but we are unable to model the number of iterations required. We consider the cost per likelihood evaluation next.

The log likelihood is

$$
\begin{aligned}
&(Y - X\beta)^{\mathsf{T}}(\sigma_A^2 A_R + \sigma_B^2 B_R + \sigma_E^2 I_N)^{-1}(Y - X\beta)\\
&+ \ln|\sigma_A^2 A_R + \sigma_B^2 B_R + \sigma_E^2 I_N|.
\end{aligned}
$$

In an analysis using the Woodbury formula we find that the log likelihood can be computed in $O(R^3 + \sum_i N_{i\bullet}^2)$ time. Because $1 \leqslant R \leqslant N$ we can write $R = N^\alpha$ for some $0 \leqslant \alpha \leqslant 1$. Then

$$
R^3 + \sum_i N_{i\bullet}^2 = R^3 + R\Bigl(\frac{1}{R}\sum_i N_{i\bullet}\Bigr)^2 \geqslant R^3 + N^2R^{-1} = N^{3\alpha} + N^{2-\alpha}.
$$

Now $N^{3\alpha}+N^{2-\alpha} > \max(N^{3\alpha}, N^{2-\alpha})$ and $\alpha = 1/2$ minimizes $\max(3\alpha, 2-\alpha)$. Therefore $R^3 + \sum_i N_{i\bullet}^2 > N^{3/2}$.

This is the same estimate one gets by considering the cost of solving a system of $R + C$ equations in $R + C$ unknowns. There are faster ways to solve the equations in special cases like nested models, and there is the possibility that sparsity patterns in the data can be exploited for speed, as is done in MixedModels [3]. However, we are interested in arbitrarily complicated sampling plans where these special cases cannot be assumed.

Figure 2 shows computed cost per iteration for 10 replicates at each of 11 different sample sizes $R^2/4$ given by $R = 10, 20, 40, \dots, 2^9 \times 10$, with $p = 5$. The cost per iteration is flat for small N presumably due to some overhead. It grows slowly until about $N = 10^4$ and then it appears to increase parallel to a reference line with slope 3/2.

Figure 3 shows total cost versus N in a setting with $p = 5$ averaged over 100 data sets. The cost curve for the MLE computation looks different from Figure 2. It does not start out flat for small N. We found that the number of iterations required to find the MLE generally rose over the range $64 \leqslant N \leqslant 6400$ and then declined gently thereafter, making it harder to discern the $O(N^{3/2})$ rate. Since the number of iterations cannot be below 1 we can be sure that the MLE cost is at least a multiple of $N^{3/2}$.

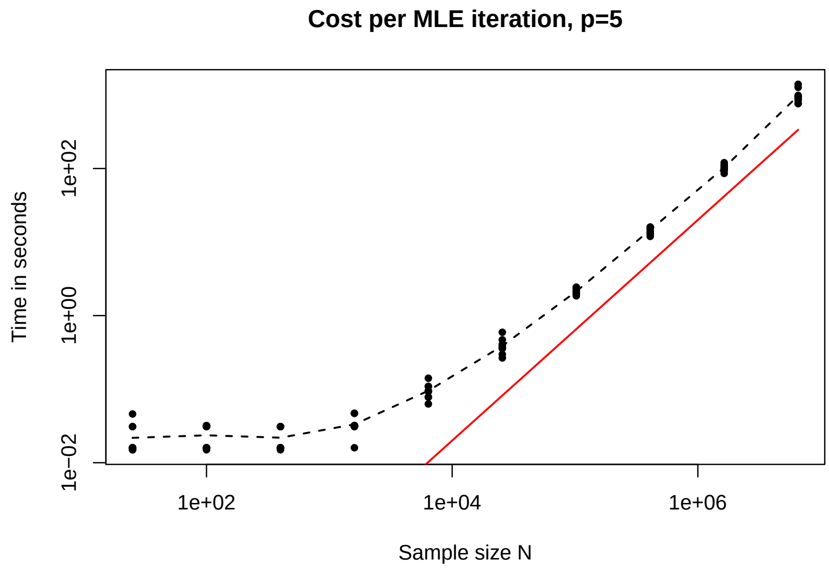

Figure 2: For $p = 5$, cost per iteration of MLE. The dashed curve is the average of 10 replicates. The solid reference line is parallel to $N^{3/2}$.

From the analysis and empirical results, we find that a cost per iteration of $O(N^{3/2})$ is a realistic lower bound for the MLE code. The method of moments cost is $O(N)$ theoretically and appears to be proportional to N empirically.

Our computations were done with data generated in memory. In commercial applications, there could be a much larger time cost proportional to N involved in reading the data from external storage. However, the $N^{3/2}$ cost component would be considerably larger at commercial scale, where N is much larger than in our examples. For the method of moments it is straightforward to read and use the data in parallel even for large N.

A second computational issue arises with the linear mixed effects MLE. The code crashes on large enough data sets because the algorithm requires $O((R+C)^2)$ memory. For $p = 5$ we were unable to take the next step past $N = 5120^2/4 \doteq 6.5 \times 10^6$. The program runs out of memory on our cluster. For $p = 1$, it crashes for N near 18 million observations. The method of moments in Algorithm 1 has linear cost both theoretically and empirically and can be implemented in $O(R+C)$ memory. The difference is minor for our CPU time simulations that also keep all N observations in memory, but it will be critical in large commercial applications.

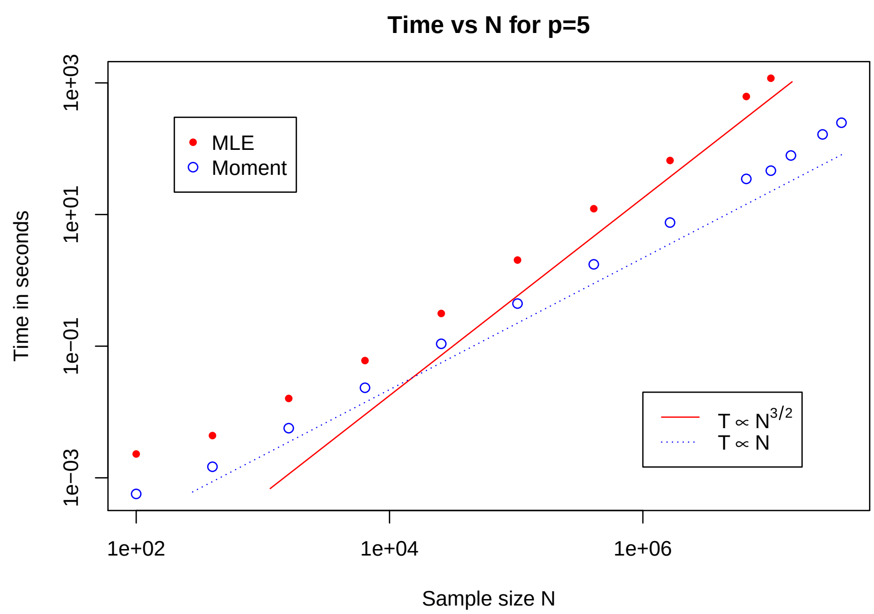

Figure 3: Computational cost for MLE and moments versus sample size N. There are reference lines parallel to $N^{3/2}$ and N^1.

6.2 Statistical efficiency

For statistical efficiency we considered $p = 1$, 5, 10 and 20. Sample sizes $N = 100 \times 4^j$ for $j = 0, 1, \ldots, 8$ were replicated 100 times each. A few larger values of N were replicated 10 times each, though the MLE code would not run on all of the largest sample sizes we tried. The pattern in the results was the same for all of those p. We display results for $p = 5$ in Figure 4. The MSEs for β decay proportionally to $1/N$. The reference curves for variance components in Figure 4 are what we would expect from IID sampling of a_i, b_j and e_{ij}, respectively: $2\sigma_A^4/R$, $2\sigma_B^4/C$ and $2\sigma_E^4/N$ where $R = C = 2\sqrt{N}$.

The parameter of greatest interest will ordinarily be β. The MLE has greater accuracy for β, as it must by the Gauss-Markov theorem. In this instance the MLE has about half the MSE that the method of moments does. For the variance components, the method of moments attains essentially the same MSE as the MLE does for σ_A^2 and σ_B^2. The MLE has greater efficiency for σ_E^2. In ordinary use we would want to know ratios of variance components and the uncertainty in such ratios is dominated by that in σ_A^2 and σ_B^2, where the two methods have comparable accuracy.

In this example, we saw a modest loss in statistical efficiency of $\hat{\beta}$ and

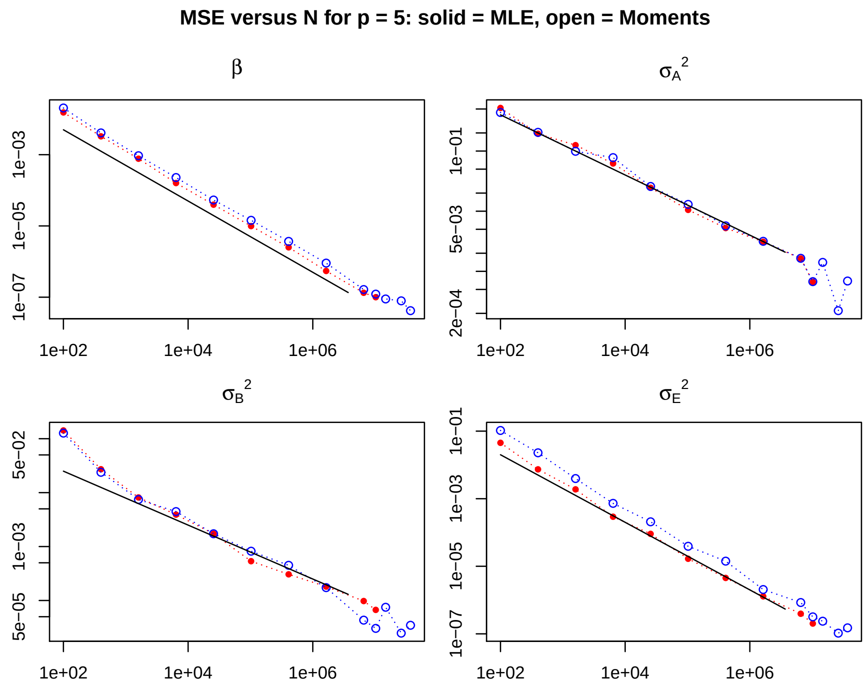

Figure 4: Mean squared errors for β, σ^2_A, σ^2_B and σ^2_E versus N. Reference lines for β and σ^2_E are parallel to N^{-1}. Reference lines for σ^2_A and σ^2_B are parallel to $N^{-1/2}$.

$\hat\sigma^2_E$ and comparable accuracy for $\hat\sigma^2_A$ and $\hat\sigma^2_B$. These comparisons were run on data simulated from the Gaussian model that the MLE assumes. The method of moments does not require that assumption. Likelihood based variance estimates for variance components, such as $\widehat{\mathrm{Var}}(\hat\sigma^2_A)$, can fail to be even asymptotically correct when the Gaussian model does not hold.

7. Conclusions, future work, and some references

We have proposed an algorithm for the two-factor linear mixed effects model with crossed covariance structure that provides consistent and asymptotically normal parameter estimates. It alternates twice between estimating the regression coefficients and estimating the variance components via the method of moments.

Unlike the available methods based on Bayes theorem or maximum likelihood, the moment estimates cost $O(N)$ time and $O(R+C)$ space. The variance estimate for $\hat\beta$ is obtained by substituting consistent estimates of

σ_A^2, σ_B^2, and σ_E^2 into exact finite sample formulas for that variance matrix. The variance estimates for $\hat\sigma_A^2$, $\hat\sigma_B^2$, and $\hat\sigma_E^2$ are obtained by such a substitution in mildly conservative formulas from [11]. Here the usual root n consistency from IID settings is replaced by a $1/\sqrt{\epsilon}$ consistency for $\epsilon = \max(\epsilon_R, \epsilon_C)$. Interpreting $1/\sqrt{\epsilon}$ as an effective sample size might be somewhat conservative because in theorems such as Theorem 1 the value of ϵ appears in upper bounds.

We exchange higher MSEs for an algorithm with cost only linear in the number of observations. We do not know how bad the efficiency loss might be in general, but we expect that when the pure error term σ_E^2 is meaningfully large that the loss will not be extreme. Also, if one of σ_A^2 and σ_B^2 very much dominates the other one, we can get a GLS estimate that accounts for the dominant source of correlation.

Gao [10] proves a martingale central limit theorem for the variance component estimates $\hat\sigma_A^2$, $\hat\sigma_B^2$, and $\hat\sigma_E^2$. We do not anticipate those variance components to be uncorrelated with $\hat\beta$ because the random variables a_i, b_j, and e_{ij} might not have symmetric distributions.

This paper is a second step in developing big data versions of mixed model procedures such as the Henderson estimators. One followup step is to incorporate interactions between fixed and random steps, as the Henderson III model allows. Another is to incorporate interactions among latent variables. At present both kinds of interactions would serve to inflate σ_E^2. A third step is to adapt to binary responses, for instance by replacing the identity link in Model 1, with a logit or probit link. This third step is of value because many responses in e-commerce are categorical, e.g., for Stitch Fix, whether the client keeps the item of clothing.

Computation for GLMMs is daunting, especially for large ones. Referring to penalized quasi-likelihood, [22, page 341] write

> In the "derivation" of the PQL equations quite a few approximations of undetermined accuracy are bandied about and the development has an air of ad hocery. How well do these methods work in practice? Unfortunately, not very.

The latest version of the `lme4` R package [1] does not include their previous `mcmcsamp` method because it was deemed to be unreliable. Jiang [18] proposes a method of simulated moments estimator for generalized linear mixed models in general by deriving sufficient statistics, but they have superlinear computational complexity. Even the theory of GLMMs is difficult. The consistency of the maximum likelihood estimate of μ in a balanced data set for a binary response Y_{ij} with $\text{logit}(\Pr(Y_{ij} = 1 \mid a_i, b_j) = \mu + a_i + b_j$ was only established in 2013 [20].

Finally, recently [23] has shown that for sparse observation patterns, the convergence time of a collapsed Gibbs sampler for an intercept-only version of Model 1, alternating between sampling μ and the a_i's and μ and the b_j's, would not grow with the size of the data set. This suggests that a reparameterization could enable MCMC to have good performance on our model as well. They do however require a strong balance assumption in which the rows are equally commonly represented in the data and similarly for the columns. That assumption is very unrealistic for e-commerce.

Supplementary Materials

The proofs of our results are in an online supplement:
`http://statweb.stanford.edu/~owen/reports/vllmemsupp.pdf`
For Statistica Sinica reviewers there is an accompanying PDF to this document that has the proper section numbers.

Acknowledgments

This work was supported by the US NSF under grants DMS-1407397 and DMS-1521145. KG was supported by US NSF Graduate Research Fellowship under grant DGE-114747. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank Stitch Fix and in particular Brad Klingenberg for providing us with the data used in our real-world experiment and motivation and encouragement during the project.

References

[1] Douglas Bates. Computational methods for mixed models. Technical report, Department of Statistics, University of Wisconsin–Madison, November 2014. `https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf`.

[2] Douglas Bates. Linear mixed-effects models in Julia. `https://github.com/dmbates/MixedModels.jl`, 2016.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[4] James Bennett and Stan Lanning. The Netflix prize. In *Proceedings of KDD Cup and Workshop 2007*, 2007.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[7] Eugene Demidenko. *Mixed models: theory and applications with R*. John Wiley & Sons, 2013.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[9] C. Fox. Polynomial accelerated MCMC and other sampling algorithms inspired by computational optimization. In J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 349–366. Springer, Berlin, 2013.

[10] Katelyn Gao. *Scalable Estimation and Inference for Massive Linear Mixed Models with Crossed Random Effects*. PhD thesis, Stanford University, 2017.

[11] Katelyn Gao and Art B. Owen. Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics*, 11(1):1235–1296, 2017.

[12] Andrew Gelman, David A. Van Dyk, Zaiying Huang, and John W. Boscardin. Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1), 2012.

[13] William W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, 1989.

[14] Herman O. Hartley and Jon N. K. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108, 1967.

[15] Charles R. Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2):226–252, 1953.

[16] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, Cambridge, 2015.

[17] J. P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[18] Jiming Jiang. Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, 93(442):720–729, 1998.

[19] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007.

[20] Jiming Jiang. The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *The Annals of Statistics*, 41(1):177–195, 2013.

[21] Albert W. Marshall and Ingram Olkin. Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40(1):89–93, 1990.

[22] Charles E McCulloch, Shayle R Searle, and John M Neuhaus. *Generalized, linear, and mixed models*. John Wiley & Sons, Hoboken, NJ, 2008.

[23] Omiros Papaspiliopoulos, Gareth O Roberts, and Giacomo Zanella. Scalable inference for crossed random effects models. *arXiv:1803.09460*, 2018.

[24] Patrick O Perry. Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society, Series B*, 79(1):267–291, 2017.

[25] Michael J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report NA2009/06, University of Cambridge, 2009.

[26] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013.

[27] Stephen W. Raudenbush. A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, 18(4):321–349, 1993.

[28] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, pages 291–317, 1997.

[29] S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

[30] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, New York, 1992.

[31] H. R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

[32] R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: context, process, and purpose. *Americal Statistician*, 70(2):129–133, 2016.

[33] Ping Wu, Winfried Stute, and Li-Xing Zhu. Efficient estimation of moments in linear mixed models. *Bernoulli*, 18(1):206–228, 2012.

[34] Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question – an ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

Katelyn Gao, Intel Inc.

E-mail: (katelyng3@gmail.com)

Art B. Owen, Stanford University

E-mail: (owen@stanford.edu)