Biometrika (2025), **112**, 3, asaf037 https://doi.org/10.1093/biomet/asaf037
Advance Access publication 26 May 2025

Consistent and scalable composite likelihood estimation of probit models with crossed random effects

By R. BELLIO

Department of Economics and Statistics, University of Udine,
Via Tomadini, 30/A, 33100 Udine, Italy
ruggero.bellio@uniud.it

S. GHOSH, A. B. OWEN

Department of Statistics, Stanford University,
Sequoia Hall, Stanford California 94305, U.S.A.
gswarnadip@gmail.com owen@stanford.edu

and C. VARIN

Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University, Via Torino 155, 30172 Venice, Italy
cristiano.varin@unive.it

Summary

Estimation of crossed random effects models commonly incurs computational costs that grow faster than linearly in the sample size N, often as fast as $\Omega(N^{3/2})$, making them unsuitable for large datasets. For non-Gaussian responses, integrating out the random effects to obtain a marginal likelihood poses significant challenges, especially for high-dimensional integrals for which the Laplace approximation may not be accurate. In this article we develop a composite likelihood approach to probit models that replaces the crossed random effects model with some hierarchical models that require only one-dimensional integrals. We show how to consistently estimate the crossed effects model parameters from the hierarchical model fits. We find that the computation scales linearly in the sample size. The method is illustrated by applying it to approximately five million observations from Stitch Fix, where the crossed effects formulation would require an integral of dimension larger than 700 000.

Some key words: Adaptive Gauss–Hermite quadrature; Binary regression; E-commerce; High-dimensional data.

1. Introduction

In this article we develop a new composite likelihood approach to handling probit models with two crossed random effects. The initial motivation was to obtain point and interval parameter estimates at a computational cost that grows only linearly with the sample size N. Standard algorithms for crossed random effects typically have superlinear cost, commonly $\Omega(N^{3/2})$, making them unsuitable for modern large datasets. A second issue is that the marginal likelihood in a crossed random effects model is an integral over $\mathbb{R}^D$, where D is large enough to make the integration problem challenging. Our scalable estimation method replaces this D-dimensional integral by D integrals of dimension one.

The common notation for mixed effects models combines fixed and random effects through a formula such as $X\beta + Zb$ involving matrices X and Z of known predictors with unknown coefficient vectors β and b, where b happens to be random. This formulation is simple and elegant, but hides some extremely important practical differences. As discussed above, the crossed setting leads to one high-dimensional integral while the hierarchical one uses many low-dimensional integrals. For Gaussian responses we can use generalized least squares on the response vector, without explicitly solving an integral. Even in that case, the crossed setting is harder. A hierarchical model has a block-diagonal covariance matrix for the response vector, resulting in a linear cost. For unbalanced crossed random effects, generalized least squares typically has a superlinear cost.

With the size N of datasets growing rapidly, it is not possible to use estimation methods with a cost of $\Omega(N^{3/2})$; ideally the cost should be $O(N)$. The present work is motivated by electronic commerce problems with large datasets. Consider a company that has customers $i = 1, \ldots, R$, to which it sells items $j = 1, \ldots, C$. The company might be interested in modelling how a response Y_{ij} depends on some predictors $x_{ij} \in \mathbb{R}^p$. If the model does not account for the fact that Y_{ij} and Y_{is} are correlated because of a common customer i or that Y_{ij} and Y_{rj} are correlated because of a common item j, an inefficient estimate will result. This flaw is less serious when N is large. What is very concerning is that the company will get unreliable standard errors for their estimates. In a setting where accounting for random effects is computationally impossible, it is expected that many users will simply ignore them, thus obtaining very naïve variance estimates and finding too many things to be significant.

A typical feature of data in our motivating applications is very sparse and imbalanced sampling. Only $N \ll RC$ of the possible (x_{ij}, Y_{ij}) values are observed. There is generally no simple structure in the pattern of which (i,j) pairs are observed. It is common for the data to have very unequal sampling frequencies in each of the row and column variables.

The scaling problem is easiest to describe for generalized least squares solutions to linear mixed models, based on results of Gao & Owen (2020). They noted that the algorithm for generalized least squares involves solving a system of $R+C$ equations in $R+C$ unknowns, which has a cost of $\Omega\{(R+C)^3\}$ in standard implementations. If $RC > N$, then $\max(R, C) > N^{1/2}$ and so $(R + C)^3 > N^{3/2}$. The average number of observations per level is $N/(RC)$. This ratio is well below 1 in our motivating applications, and as long as it is $o(N^{1/3})$, the cost of standard algebra will be superlinear.

Standard Bayesian solutions run into a similar difficulty. For an intercept plus crossed random effects model, Gao & Owen (2017) showed that the Gibbs sampler takes $\Omega(N^{1/2})$ iterations that each have a cost proportional to N, for an $\Omega(N^{3/2})$ cost overall. Several other Bayesian approaches they considered also encountered difficulties.

There has been recent progress on scalable algorithms for crossed random effects problems, improving upon both the frequentist and the Bayesian approaches. For regression problems, Ghosh et al. (2022a) replaced standard equation solving by a backfitting algorithm that has $O(N)$ cost per iteration and gave conditions under which the number of iterations to convergence is $O(1)$ as $N \to \infty$. See also Ghandwani et al. (2023) for regression with random slopes. Papaspiliopoulos et al. (2020) used a collapsed Gibbs sampler and gave conditions under which it has linear cost in N for the intercept-only crossed random effects regression model. Ghosh & Zhong (2021) did the same after weakening a stringent balance assumption.

Here we consider a binary response requiring a generalized linear mixed effects model with crossed random effects that encounters the high-dimensional integration problem mentioned above. Fewer scalable solutions are available for this problem. There is a frequentist approach due to Ghosh et al. (2022b) and a Bayesian approach developed by Papaspiliopoulos et al. (2023). The all-row-column probit presented here is simpler than those two approaches and uses much weaker sampling assumptions.

Ghosh et al. (2022b) developed a penalized quasilikelihood approach to logistic regression on fixed effects and two crossed random effects. Their method involves iterations costing $O(N)$ each, and empirically the number of iterations is $O(1)$. They used estimating equations from Breslow & Clayton (1993), which were based on work by Schall (1991) to maximize the marginal likelihood using a Laplace approximation. The quantity being estimated is not exactly the maximum likelihood estimate; it is a posterior mode corresponding to a not-very-informative prior, using some plugged-in weights, a quantity that goes back to Stiratelli et al. (1984). Penalized quasilikelihood has a bias that can prevent it from being consistent. Even with just the intercept and one random effect, it requires the number of levels of that effect and the number of observations at each of those levels to diverge to infinity in order to yield a consistent estimate. With sample sizes of $R = N^{\rho}$ and $C = N^{\kappa}$ for $\rho, \kappa \in (0, 1)$, penalized quasilikelihood requires $\max(\rho + 2\kappa, 2\rho + \kappa) < 2$, and the observation probability for (x_{ij}, Y_{ij}) can vary over only a narrow range.

Papaspiliopoulos et al. (2023) extended the collapsed Gibbs sampler to obtain scalable Bayesian inference in generalized linear mixed models with crossed random effects using a reparameterization called local centring. They included an intercept and $K \geqslant 2$ crossed random effects, while also discussing how fixed effects could be incorporated. Their approach requires a stringent balance assumption. For our data we would need $N_{i\bullet} = N/C$ for all rows $i = 1, \ldots, R$ and $N_{\bullet j} = N/R$ for all $j = 1, \ldots, C$. Under this condition, their cost per iteration is $O(N + R + C)$ in our notation. The total cost is this cost per iteration times a relaxation time. For $K = 2$ random effects and a discrete response like the one we study, they introduced an auxiliary relaxation time T_{aux} and showed that the cost is $O[\max(N, R + C) \min\{2N/(R + C), T_{\text{aux}}\}]$. Our problem has $\max(R, C) \ll N$, and then their cost is linear in N if and only if $T_{\text{aux}} = O(1)$. They get $T_{\text{aux}} = O(1)$ when the observation pattern is uniformly distributed over all patterns with $N_{i\bullet} = N/C$ and $N_{\bullet j} = N/R$. Both R and C must grow linearly with N. They then require $N_{i\bullet} = O(1)$ and $N_{\bullet j} = O(1)$.

Our balance criteria are minimal. The main results require $\max(\max_i N_{i\bullet}, \max_j N_{\bullet j})/N = o(1)$. Both R and C can grow with N at different rates. Our simulations, but not our theory, use a balance condition with nearly equal values among the $N_{i\bullet}$ and nearly equal values among the $N_{\bullet j}$.

We study the probit model because its Gaussian latent variable is a good match for Gaussian random effects. The probit and logit link functions are nearly proportional outside tail regions (Agresti, 2002, pp. 246–7), so they often give similar results.

Crossed models also differ from hierarchical models in that only recently has the maximum likelihood estimate for generalized linear mixed models been shown to be consistent for crossed random effects. This was accomplished by the subset argument of Jiang (2013), showing that the score equations have a root that is consistent for the parameter. One of the main open questions in generalized linear mixed models, such as the ones we consider here, is at what rate the estimators of the model parameters converge. Intuitively, one might expect $O(N^{-1/2})$ or $O\{\min(R, C)^{-1/2}\}$, in accordance with the results for nested designs

(Jiang et al., 2022). In some settings estimators for different parameters converge at different rates. See Jiang (2013) for some discussion and Lyu et al. (2024) for recent work assuming balanced sampling.

2. THE ALL-ROW-COLUMN METHOD

2.1. *Preliminaries*

We consider two crossed random effects. There is a vector $a \in \mathbb{R}^R$ with elements a_i and another vector $b \in \mathbb{R}^C$ with elements b_j. Conditionally on a and b, the Y_{ij} are independent with

$$\text{pr}(Y_{ij} = 1 \mid a, b) = \Phi(x_{ij}^{\mathrm{T}}\beta + a_i + b_j), \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. In these models, random effects are typically assumed to be uncorrelated, $a \sim \mathcal{N}(0, \sigma_A^2 I_R)$ independently of $b \sim \mathcal{N}(0, \sigma_B^2 I_C)$, where I_n is the $n \times n$ identity matrix; see, for example, McCullagh & Nelder (1989, p. 444). The probit model has a representation in terms of latent variables $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$ as

$$Y_{ij} = 1\{x_{ij}^{\mathrm{T}}\beta + a_i + b_j + \varepsilon_{ij} > 0\}, \tag{2}$$

where $1\{E\}$ is the indicator function of the event E. The probability (1) and the likelihoods derived from it are all conditional on the values of x_{ij}.

We write $\mathcal{S} \subset \{1, \ldots, R\} \times \{1, \ldots, C\}$ for the set of (i, j) pairs where (x_{ij}, Y_{ij}) was observed. We also work conditionally on $\mathcal{S}$. In our motivating applications, the pattern of observation or missingness could be informative. Addressing that issue would necessarily require information from outside the data. Furthermore, the scaling problem is still the subject of ongoing research even in the noninformative missingness setting. Therefore, we consider estimation strategies without taking account of missingness.

The likelihood for $\theta = (\beta^{\mathrm{T}}, \sigma_A^2, \sigma_B^2)^{\mathrm{T}}$ is a cumbersome integral of size $R + C$,

$$L(\theta) = \sigma_A^{-R}\sigma_B^{-C} \int_{\mathbb{R}^{R+C}} L(\beta \mid a, b) \prod_{i=1}^{R} \varphi\left(\frac{a_i}{\sigma_A}\right) \prod_{j=1}^{C} \varphi\left(\frac{b_j}{\sigma_B}\right) \mathrm{d}a\, \mathrm{d}b, \tag{3}$$

where $\varphi(\cdot)$ is the standard normal probability density function and $L(\beta \mid a, b)$ is the conditional likelihood of β, given the random effects. The conditional likelihood we need is

$$L(\beta \mid a, b) = \prod_{(i,j) \in \mathcal{S}} \Phi(x_{ij}^{\mathrm{T}}\beta + a_i + b_j)^{y_{ij}} \Phi(-x_{ij}^{\mathrm{T}}\beta - a_i - b_j)^{1-y_{ij}},$$

and $L(\theta)$ is commonly called the marginal likelihood.

The first-order Laplace approximation is a standard approach to approximating the integral in the marginal likelihood. The Laplace algorithm maximizes the logarithm of the integrand in the marginal likelihood (3) over $a \in \mathbb{R}^R$ and $b \in \mathbb{R}^C$ for fixed θ. It then multiplies the maximum value of the integrand by $\det\{H^{-1/2}(\theta)\}$, where $H(\theta) \in \mathbb{R}^{(R+C)\times(R+C)}$ is the Hessian of the log integrand with respect to a and b for fixed θ. The result is an approximate marginal likelihood $\tilde{L}(\theta)$ that is optimized to obtain $\hat{\theta}$; see, for example, Ogden (2021) or Shun & McCullagh (1995). If the square root of the inverse Hessian is computed by standard methods, then that alone has a cost of $\Omega(N^{3/2})$ by the argument for generalized linear models discussed in § 1. Similarly, if the inner optimization is done using Newton steps, that will have a cost of $\Omega(N^{3/2})$ per iteration. We return to this issue in § 3.2, where a Laplace approximation is shown to have a superlinear cost that is $o(N^{3/2})$.

Even if the Laplace approximation were computable for large N, it would not provide asymptotically valid results in our context, because the size of the likelihood integral corresponds to the number of random effects $\Omega(N^{1/2})$, and therefore grows too fast with the sample size to ensure consistent results. See Shun & McCullagh (1995), Ogden (2021) and Tang & Reid (2025) for discussions of the conditions that must be satisfied to make the Laplace approximation reliable when the size of the integral grows with the sample size.

2.2. *Scalable composite likelihood inference*

Our approach to obtaining a consistent and scalable estimator in high-dimensional probit models with crossed random effects combines estimates for three misspecified probit models. Each of them is constructed through the omission of some random effects. By combining (1) and (2), we find that marginally

$$\mathrm{pr}(Y_{ij}=1)=\Phi(x_{ij}^{\mathrm{T}}\gamma) \tag{4}$$

for $\gamma=\beta/(1+\sigma_A^2+\sigma_B^2)^{1/2}$. The proposed method begins with estimation of γ from the naïve model (4) that omits both of the random effects through maximization of the likelihood

$$L_{\mathrm{all}}(\gamma)=\prod_{(i,j)\in\mathcal{S}}\Phi(x_{ij}^{\mathrm{T}}\gamma)^{y_{ij}}\Phi(-x_{ij}^{\mathrm{T}}\gamma)^{1-y_{ij}}. \tag{5}$$

Maximization of this likelihood requires neither high-dimensional integrals nor expensive algebra, and it is observed to take $O(N)$ computation in our examples. We then need estimates of σ_A^2 and σ_B^2 to get the scale right. Our analysis of (4) will also account for within-row and within-column correlations among the Y_{ij}. While $\mathrm{sign}(\gamma_k)=\mathrm{sign}(\beta_k)$ $(k=1,\ldots,p)$, confidence intervals for γ_k based on model (4) would be naïve if they did not account for the dependence among the responses.

Consider the reparameterization $\psi=(\gamma^{\mathrm{T}},\tau_A^2,\tau_B^2)^{\mathrm{T}}$, where

$$\gamma=\frac{\beta}{(1+\sigma_A^2+\sigma_B^2)^{1/2}},\quad \tau_A^2=\frac{\sigma_A^2}{1+\sigma_B^2},\quad \tau_B^2=\frac{\sigma_B^2}{1+\sigma_A^2}. \tag{6}$$

After dividing $x_{ij}^{\mathrm{T}}\beta+a_i+b_j+\varepsilon_{ij}$ by $(1+\sigma_A^2)^{1/2}$ or by $(1+\sigma_B^2)^{1/2}$, model (2) implies that

$$\mathrm{pr}(Y_{ij}=1\mid a)=\Phi(x_{ij}^{\mathrm{T}}\gamma_A+u_i), \tag{7}$$

$$\mathrm{pr}(Y_{ij}=1\mid b)=\Phi(x_{ij}^{\mathrm{T}}\gamma_B+v_j), \tag{8}$$

where $u\sim\mathcal{N}(0,\tau_A^2I_R)$ and $v\sim\mathcal{N}(0,\tau_B^2I_C)$ for $\gamma_A=\gamma(1+\tau_A^2)^{1/2}$ and $\gamma_B=\gamma(1+\tau_B^2)^{1/2}$. Given $\hat{\gamma}$, the maximizer of (5), we proceed with estimation of τ_A^2 and τ_B^2 from the two probit models (7) and (8) that each omit one of the random effects. Fitting models (7) and (8) involves simpler integrals than (1) since the latent variable representations of these models,

$$Y_{ij} = 1\{x_{ij}^{\mathrm{T}}\gamma_A + u_i + \varepsilon_{ij} > 0\}, \quad Y_{ij} = 1\{x_{ij}^{\mathrm{T}}\gamma_B + v_j + \varepsilon_{ij} > 0\},$$

have hierarchical (not crossed) error structures. This is where we are able to replace the $(R+C)$-dimensional integral (3) by $R+C$ univariate ones. Model (7) is fitted by maximizing the row-wise likelihood

$$L_{\mathrm{row}}(\tau_A^2) = \tau_A^{-R} \prod_{i=1}^{R} \int_{\mathbb{R}} L_{i\bullet}(\hat{\gamma}_A \mid u_i)\varphi\left(\frac{u_i}{\tau_A}\right) \mathrm{d}u_i, \tag{9}$$

where $L_{i\bullet}(\hat{\gamma}_A \mid u_i)$ is the conditional likelihood of $\hat{\gamma}_A = \hat{\gamma}\,(1+\tau_A^2)^{1/2}$, given u_i,

$$L_{i\bullet}(\hat{\gamma}_A \mid u_i) = \prod_{j \mid i} \Phi(x_{ij}^{\mathrm{T}}\hat{\gamma}_A + u_i)^{y_{ij}}\Phi(-x_{ij}^{\mathrm{T}}\hat{\gamma}_A - u_i)^{1-y_{ij}},$$

where $j \mid i = \{j : (i,j) \in \mathcal{S}\}$ is the set of indices j such that (x_{ij}, Y_{ij}) is observed. The row-wise likelihood (9) is a product of R one-dimensional integrals and is a function of τ_A^2 only, because we fix γ at the estimate $\hat{\gamma}$ obtained from the maximization of the all likelihood in the previous step. Rows with a single observation do not contribute to estimation of τ_A^2 because $\mathrm{pr}(Y_{ij} = 1) = \Phi(x_{ij}^{\mathrm{T}}\hat{\gamma})$, which does not depend on τ_A^2. Reversing the roles of the rows and columns, we get a column-wise likelihood L_{col}, which we maximize to obtain an estimate $\hat{\tau}_B^2$ of τ_B^2.

Finally, we invert the equations in (6) to get

$$\hat{\beta} = \hat{\gamma}\,(1+\hat{\sigma}_A^2+\hat{\sigma}_B^2)^{1/2}, \quad \hat{\sigma}_A^2 = \frac{\hat{\tau}_A^2(1+\hat{\tau}_B^2)}{1-\hat{\tau}_A^2\hat{\tau}_B^2}, \quad \hat{\sigma}_B^2 = \frac{\hat{\tau}_B^2(1+\hat{\tau}_A^2)}{1-\hat{\tau}_A^2\hat{\tau}_B^2}.$$

By the definitions in (6), $\tau_A^2\tau_B^2 < 1$. In our computations, we have never encountered a setting where $\hat{\tau}_A^2\hat{\tau}_B^2 \geqslant 1$, so our estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ have never been negative.

We call our method the *all-row-column method*. The name comes from model (4), which uses all the data at once, model (7), which combines likelihood contributions from within each row, and model (8), which combines likelihood contributions from within each column. Figure 1 illustrates these models for a dataset of $N = 39$ observations in $R = 10$ rows and $C = 10$ columns. In each of three misspecified probit models, points in different boxes are assumed to be independent. Figure 1(a) shows the *all* model with independent data. Panels (b) and (c) illustrate the *row* and *column* hierarchical models, respectively. We combine fits from these three misspecified models to get consistent scalable formulas despite the dependencies involved. This approach can be viewed as a new form of composite likelihood (Lindsay, 1988; Varin et al., 2011). An earlier version of composite likelihood that was applied to crossed random effects is discussed by Bellio & Varin (2005). They considered a standard pairwise likelihood which is, however, not scalable, and thus inappropriate for our problem. Instead, our all-row-column method has $O(N)$ cost per iteration and achieves $O(N)$ cost empirically.

Computation of the row- and column-wise likelihoods requires the approximation of up to R and C univariate integrals of a standard hierarchical probit model. Since R and

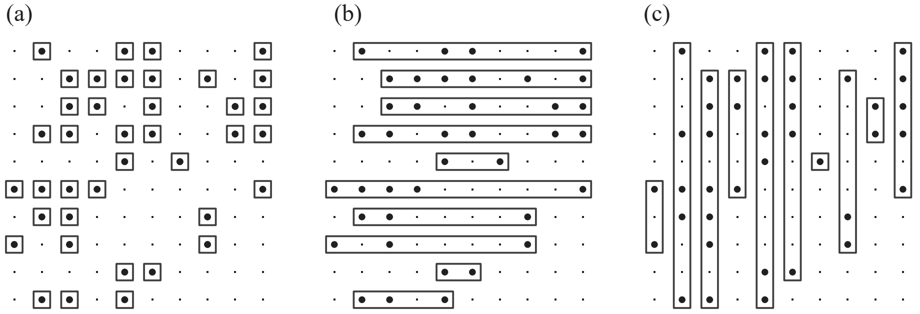

Fig. 1. The all-row-column method for $N = 39$ observations (marked •) in $R = 10$ rows and $C = 10$ columns: (a) the all model; (b) the row hierarchical model; (c) the column hierarchical model.

C are large in the applications that motivate this work, accurate approximation of the one-dimensional integrals is crucial. Otherwise, the accumulation of approximation errors could induce serious biases in the estimation of τ_A^2 and τ_B^2. Our intensive numerical studies have indicated that accurate approximation of the integrals can be achieved by the well-established adaptive Gauss–Hermite quadrature (e.g., Liu & Pierce, 1994) with a suitable choice of the number of quadrature nodes.

The next theorem establishes the weak consistency of the maximizer $\hat{\gamma}$ of the all likelihood (5) with respect to the value γ in (6). First we introduce some notation. Let $Z_{ij} = 1$ if (x_{ij}, Y_{ij}) is observed and $Z_{ij} = 0$ otherwise. The number of observations in row i is $N_{i\bullet} = \sum_{j=1}^{C} Z_{ij}$, and similarly column j has $N_{\bullet j} = \sum_{i=1}^{R} Z_{ij}$ observations in it. Let $\epsilon_R = \max_i N_{i\bullet}/N$ and $\epsilon_C = \max_j N_{\bullet j}/N$. We assume that $\max(\epsilon_R, \epsilon_C) \to 0$ as $N \to \infty$. We take Z_{ij} to be deterministic with at most one observation for any (i,j) pair. In our motivating applications one would seldom, if ever, have multiple observations for any (i,j) pair. Even then, one might only use the most recent of those observations. The x_{ij} are not dependent on the Z_{ij}. Finally, the Y_{ij} are sampled from their probit distribution conditionally on x_{ij}.

Theorem 1. *Let $Y_{ij} \in \{0, 1\}$ follow the crossed random effects probit model (1) with true value γ_0 for the parameter $\gamma = \beta/(1 + \sigma_A^2 + \sigma_B^2)^{1/2}$. Let the number of observations $N \to \infty$ while $\max(\epsilon_R, \epsilon_C) \to 0$. Suppose that $x_{ij} \in \mathbb{R}^p$ satisfy the following conditions:*

(i) $\|x_{ij}\| \leqslant B < \infty$;
(ii) $N^{-1} \sum_{ij} Z_{ij} x_{ij} x_{ij}^{\mathrm{T}} \to V \in \mathbb{R}^{p\times p}$, *where V is positive definite;*
(iii) *there is no nonzero vector $v \in \mathbb{R}^p$ such that $v^{\mathrm{T}} x_{ij} \geqslant 0$ for all (i,j) with $Z_{ij} = 1$ and $y_{ij} = 1$ and $v^{\mathrm{T}} x_{ij} \leqslant 0$ for all (i,j) with $Z_{ij} = 1$ and $y_{ij} = 0$.*

Let $\hat{\gamma} \in \mathbb{R}^p$ be any maximizer of (5). Then for any $\epsilon > 0$,

$$\mathrm{pr}(\|\hat{\gamma} - \gamma_0\| > \epsilon) \to 0 \quad \text{as } N \to \infty.$$

The proof is given in the Supplementary Material and was obtained by adapting the proof strategy of Lumley & Mayer Hamblett (2003) to our setting. The balance condition that no single row or column have a nonvanishing fraction of data is much weaker than the norm in the crossed random effects literature.

Now we consider consistent estimation of τ_A^2 and τ_B^2 from the row and column likelihoods, respectively. These provide consistent estimates of σ_A^2 and σ_B^2, with which one can then adjust the consistent estimate of γ to get a consistent estimate of β.

Theorem 2. *Under the assumptions of Theorem 1, there is a root of the row likelihood equation that is a consistent estimator for τ_A^2, and there is a root of the column likelihood equation that is a consistent estimator for τ_B^2.*

The proof of this theorem is also presented in the Supplementary Material. It uses the subset argument of Jiang (2013) to show Cramer consistency of the maximum row likelihood estimator of τ_A^2 and, equivalently, of the maximum column likelihood estimator of τ_B^2.

2.3. *Robust sandwich variance*

After a customary Taylor approximation, the variance of $\hat\theta$ is $\text{var}(\hat\theta) \doteq D\,\text{var}(\hat\psi)D^{\mathrm{T}}$, where D is the Jacobian matrix of the reparameterization from θ to ψ. Letting ψ_0 denote the true value of ψ and $u_{\text{arc}}(\psi)$ the score vector of the all-row-column estimator constructed by stacking the scores of the three misspecified likelihoods, the asymptotic variance of $\hat\psi$ is

$$\text{avar}(\hat\psi) = \mathcal{I}_{\text{arc}}^{-1}(\psi_0)V_{\text{arc}}(\psi_0)\mathcal{I}_{\text{arc}}^{-1}(\psi_0),$$

where $\mathcal{I}_{\text{arc}}(\psi) = -E\{\partial u_{\text{arc}}(\psi)/\partial\psi\}$ and $V_{\text{arc}}(\psi) = \text{var}\{u_{\text{arc}}(\psi)\}$ are the expected information and score variance for the all-row-column estimator. These two matrices are not equal because the second Bartlett identity does not hold for the misspecified likelihoods that constitute the all-row-column method. While estimation of the *bread* matrix $\mathcal{I}_{\text{arc}}$ of the sandwich is not problematic, direct computation of the *filling* matrix V_{arc} is not feasible in our large-scale set-up because it requires the approximation of a large number of multi-dimensional integrals with a cost that does not meet our $O(N)$ constraint.

Since all-row-column estimates require $O(N)$ computations, we can estimate the variance of $\hat\theta$ with a parametric bootstrap. However, it is preferable to evaluate the estimation uncertainty without assuming the correctness of the fitted model, for example by using the nonparametric pigeonhole bootstrap described in Owen (2007). In that approach, the rows in the dataset are resampled independently of the columns. So if a row is included twice and a column is included three times, the corresponding element is included six times. The resulting bootstrap variance for a mean (such as one in a score equation) overestimates the random effects variance by an asymptotically negligible amount. It does not require homoscedasticity of either the random effects or the errors.

We now describe a convenient approach that we have developed to estimate the variance of $\hat\beta$ in $O(N)$ operations without the need of resampling and repeated fitting as in the bootstraps mentioned above. The partitioned expected all-row-column information matrix is

$$\mathcal{I}_{\text{arc}}(\psi) = -\begin{pmatrix} E\{\partial^2\ell_{\text{all}}/(\partial\gamma\,\partial\gamma^{\mathrm{T}})\} & 0 & 0 \\ E\{\partial^2\ell_{\text{row}}/(\partial\gamma\,\partial\tau_A^2)\} & E\{\partial^2\ell_{\text{row}}/(\partial\tau_A^2\partial\tau_A^2)\} & 0 \\ E\{\partial^2\ell_{\text{col}}/(\partial\gamma\,\partial\tau_B^2)\} & 0 & E\{\partial^2\ell_{\text{col}}/(\partial\tau_B^2\partial\tau_B^2)\} \end{pmatrix},$$

where $\ell_{\text{all}} = \log L_{\text{all}}(\gamma)$ from (5), $\ell_{\text{row}} = \log L_{\text{row}}(\tau_A^2)$ from (9) and $\ell_{\text{col}} = \log L_{\text{col}}(\tau_B^2)$. Since $\mathcal{I}_{\text{arc}}(\psi)$ is triangular, the asymptotic variance for $\hat{\gamma}$ is

$$\text{avar}(\hat{\gamma}) = \mathcal{I}_{\text{all}}^{-1}(\gamma_0) V_{\text{all}}(\gamma_0) \mathcal{I}_{\text{all}}^{-1}(\gamma_0), \tag{10}$$

where $\mathcal{I}_{\text{all}}(\gamma) = -E\{\partial^2 \ell_{\text{all}}/(\partial\gamma\, \partial\gamma^{\text{T}})\}$ and $V_{\text{all}}(\gamma) = \text{var}(\partial \ell_{\text{all}}/\partial\gamma)$ are the Fisher expected information and the score variance of the all likelihood. The asymptotic variance (10) for $\hat{\gamma}$ is thus the same as that of the estimator that maximizes the all likelihood when the nuisance parameters τ_A^2 and τ_B^2 are known. The robust sandwich estimator of the variance of $\hat{\gamma}$ is obtained by replacing $\mathcal{I}_{\text{all}}(\gamma_0)$ and $V_{\text{all}}(\gamma_0)$ with some estimators that are consistent and robust to misspecification. The expected information is naturally estimated with the observed information,

$$J_{\text{all}}(\hat{\gamma}) = \sum_{ij} Z_{ij}\varphi(\hat{\eta}_{ij}) \left\{ y_{ij} \frac{\varphi(\hat{\eta}_{ij}) + \hat{\eta}_{ij}\, \Phi(\hat{\eta}_{ij})}{\Phi(\hat{\eta}_{ij})^2} + (1 - y_{ij}) \frac{\varphi(\hat{\eta}_{ij}) - \hat{\eta}_{ij}\, \Phi(-\hat{\eta}_{ij})}{\Phi(-\hat{\eta}_{ij})^2} \right\} x_{ij} x_{ij}^{\text{T}},$$

where $\hat{\eta}_{ij} = x_{ij}^{\text{T}}\hat{\gamma}$. Estimation of $V_{\text{all}}(\gamma_0)$ is more involved. This matrix can be decomposed into the sum of three terms,

$$\begin{aligned} V_{\text{all}}(\gamma_0) &= \sum_{ij}\sum_{rs} Z_{ij} Z_{rs} E\left\{u_{ij}(\gamma_0) u_{rs}^{\text{T}}(\gamma_0)\right\} \\ &= \sum_{ijs} Z_{ij} Z_{is} E\left\{u_{ij}(\gamma_0) u_{is}^{\text{T}}(\gamma_0)\right\} + \sum_{ijr} Z_{ij} Z_{rj} E\left\{u_{ij}(\gamma_0) u_{rj}^{\text{T}}(\gamma_0)\right\} \\ &\quad - \sum_{ij} Z_{ij} E\left\{u_{ij}(\gamma_0) u_{ij}^{\text{T}}(\gamma_0)\right\}, \end{aligned}$$

where $u_{ij}(\gamma)$ is the score for a single observation Y_{ij},

$$u_{ij}(\gamma) = \frac{\varphi(x_{ij}^{\text{T}}\gamma)\{y_{ij} - \Phi(x_{ij}^{\text{T}}\gamma)\} x_{ij}}{\Phi(x_{ij}^{\text{T}}\gamma)\Phi(-x_{ij}^{\text{T}}\gamma)}.$$

The corresponding estimator of $V_{\text{all}}(\gamma_0)$ is

$$\hat{V}_{\text{all}}(\hat{\gamma}) = \hat{V}_A(\hat{\gamma}) + \hat{V}_B(\hat{\gamma}) - \hat{V}_{A\cap B}(\hat{\gamma}),$$

whose components are computed by grouping the individual scores with respect to each random effect and their interaction,

$$\hat{V}_A(\hat{\gamma}) = \sum_i u_{i\bullet}(\hat{\gamma}) u_{i\bullet}^{\text{T}}(\hat{\gamma}), \quad \hat{V}_B(\hat{\gamma}) = \sum_j u_{\bullet j}(\hat{\gamma}) u_{\bullet j}^{\text{T}}(\hat{\gamma}), \quad \hat{V}_{A\cap B}(\hat{\gamma}) = \sum_{ij} Z_{ij} u_{ij}(\hat{\gamma}) u_{ij}^{\text{T}}(\hat{\gamma}),$$

where $u_{i\bullet}(\gamma) = \sum_j Z_{ij} u_{ij}(\gamma)$ and $u_{\bullet j}(\gamma) = \sum_i Z_{ij} u_{ij}(\gamma)$. Estimators of the form $\hat{V}_{\text{all}}(\hat{\gamma})$ are used in statistical modelling of data clustered within multiple levels in medical applications (Miglioretti & Heagerty, 2004) and in economics (Cameron et al., 2011), where they are known as two-way cluster-robust sandwich estimators.

Finally, we approximate the variance of $\hat{\beta}$ by plugging in the estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$,

$$\widehat{\text{var}}(\hat{\beta}) = (1 + \hat{\sigma}_A^2 + \hat{\sigma}_B^2)\,\widehat{\text{var}}(\hat{\gamma}) = (1 + \hat{\sigma}_A^2 + \hat{\sigma}_B^2)\,J_{\text{all}}^{-1}(\hat{\gamma})\,\hat{V}_{\text{all}}(\hat{\gamma})J_{\text{all}}^{-1}(\hat{\gamma}). \tag{11}$$

A limitation of this approach is that it neglects the uncertainty in the estimation of the variance components: although we do not expect a substantial impact in high-dimensional applications, it could be possible to adjust (11) for the variability of the variance components through bootstrapping the row-wise and column-wise estimates $\hat{\tau}_A^2$ and $\hat{\tau}_B^2$.

3. SIMULATIONS

3.1. *Simulation settings*

We simulate from the probit model with crossed random effects (1) and compare the performance of the all-row-column estimator with that of the traditional estimator obtained by maximizing the first-order Laplace approximation of the likelihood. Another method considered is an infeasible oracle estimator that uses the unknown true values of σ_A^2 and σ_B^2 to estimate the regression parameters as $\hat{\beta}_{\text{oracle}} = \hat{\gamma}\,(1 + \sigma_A^2 + \sigma_B^2)^{1/2}$. The all-row-column method instead corrects $\hat{\gamma}$ using estimates of the variance components. All methods were implemented in the R (R Development Core Team, 2025) language. The package TMB (Kristensen et al., 2016) was used for the Laplace approximation, with the nlminb optimization function employed for its maximization. We did not use the popular glmer function from the R package lme4 (Bates et al., 2015) because the current version of TMB is substantially more computationally efficient, thus allowing us to compare our method with the first-order Laplace approximation at larger dimensions than would otherwise be possible with glmer. The row- and column-wise likelihoods were coded in C++ and integrated in R with Rcpp (Eddelbuettel, 2013) and were optimized by Brent's method as implemented in the optimise function.

We consider eight different settings defined by combining three binary factors. The first factor is whether the simulation is balanced (i.e., equal numbers of rows and columns) or imbalanced (with very unequal numbers of rows and columns) like we typically see in applications. The second factor is whether the regression model is null apart from a nonzero intercept or has nonzero regression coefficients. The third factor is whether the random effect variances are set at a high level or at a low level. Given R and C, the set $\mathcal{S}$ is obtained by independent and identically distributed Bernoulli sampling with probability $1.27 \times N/(RC)$. The value 1.27 is the largest for which Ghosh et al. (2022a) could prove that backfitting takes $O(1)$ iterations. This sampling makes the attained value of N random, but with a very small coefficient of variation.

We denote by $R = N^{\rho}$ and $C = N^{\kappa}$ the numbers of rows and columns in the data expressed as powers of the total sample size N. The two levels of the balance factor are termed *balanced*, with $\rho = \kappa = 0.56$, and *imbalanced*, with $\rho = 0.88$ and $\kappa = 0.53$. The balanced case was used in Ghosh et al. (2022b), and the imbalanced case is similar to the Stitch Fix data in § 4. Because $\rho + \kappa > 1$, the fraction of possible observations in the data is $N/(RC) = N^{1-\rho-\kappa} \to 0$ as $N \to \infty$, providing asymptotic sparsity in both cases. While the first choice has R/C being constant, the second choice has $R/C \to \infty$ with N. We believe that this asymptotic behaviour provides a better description of our motivating problems than either a setting with C fixed as $R \to \infty$ or the setting that is common in random matrix theory (Edelman & Rao, 2005), where R and C diverge with R/C approaching a constant

Table 1. *Summary of the eight simulation settings*

Setting	Sparsity		Predictors	Variances	
	ρ	κ		σ_A	σ_B
Bal-Nul-Hi	0.56	0.56	All zero	1.0	1.0
Imb-Nul-Hi	0.88	0.53	All zero	1.0	1.0
Bal-Lin-Hi	0.56	0.56	Not all zero	1.0	1.0
Imb-Lin-Hi	0.88	0.53	Not all zero	1.0	1.0
Bal-Nul-Lo	0.56	0.56	All zero	0.5	0.2
Imb-Nul-Lo	0.88	0.53	All zero	0.5	0.2
Bal-Lin-Lo	0.56	0.56	Not all zero	0.5	0.2
Imb-Lin-Lo	0.88	0.53	Not all zero	0.5	0.2

value. When $\rho = 0.88$ and $\kappa = 0.53$, the condition $\max(\rho + 2\kappa, 2\rho + \kappa) < 2$ invoked for the estimator in Ghosh et al. (2022b) does not hold.

We consider seven predictors generated from a multivariate zero-mean normal distribution with covariance matrix Σ corresponding to an autocorrelation process of order 1, so that the (k, l) entry of Σ is $\phi^{|k-l|}$. We set $\phi = 0.5$ in all the simulations. We always use the intercept $\beta_0 = -1.2$ because in our applications $\text{pr}(Y = 1) < 1/2$ is typical. For the predictor coefficients we consider two choices, referred to as *null*, with $\beta_\ell = 0$, and *linear*, with $\beta_\ell = -1.2 + 0.3\ell$ $(\ell = 1, \ldots, 7)$. The first choice is a null setting where x is not predictive at all, while the second has modestly important nonzero predictors whose values are in linear progression.

The two choices for the variance component parameters are termed *high variance*, with $\sigma_A = 1$ and $\sigma_B = 1$, and *low variance*, with $\sigma_A = 0.5$ and $\sigma_B = 0.2$. We choose the first setting to include variances higher than those typically observed in applications. The second setting is closer to what is seen in data such as in § 4. We represent the eight settings with mnemonics as shown in Table 1. For example, Imb-Nul-Hi means row-column imbalance ($\rho = 0.88$ and $\kappa = 0.53$) with all predictor coefficients being zero and the main effect variances being large ($\sigma_A = 1$ and $\sigma_B = 1$).

For each of these eight settings, we consider 13 increasing sample sizes N in the interval from 10^3 to 10^6 obtained by taking 13 equispaced values on the $\log_{10}$ scale. As will be seen next, the Laplace method has a cost that grows superlinearly, so to keep costs reasonable we only use sample sizes up to 10^5 for that method. For each of the 13 sample sizes and each of the eight settings, we simulate 1000 datasets.

As suggested by a referee, we experimented with different values for the number of quadrature nodes k to approximate the univariate integrals of the row- and column-wise likelihoods; see also the work of Bilodeau et al. (2024). The value $k = 1$, which corresponds to the Laplace approximation, produced estimates of σ_A and σ_B affected by substantial downward bias even at large N. The bias disappeared upon increasing k. With $k = 5$, we obtained results that were not affected by bias and essentially indistinguishable from those with $k = 25$. Therefore, the results discussed in the rest of this section are those obtained with $k = 5$ nodes.

Graphs comparing the computational costs, the statistical properties and the scalability of the three estimation methods for all eight settings are reported in the Supplementary Material. To save space, here we present graphs for only one of the settings, Imb-Nul-Hi, and merely summarize the other settings. This chosen setting is a challenging one. It is not surprising that imbalance and large variances are challenging. The binary regression setting

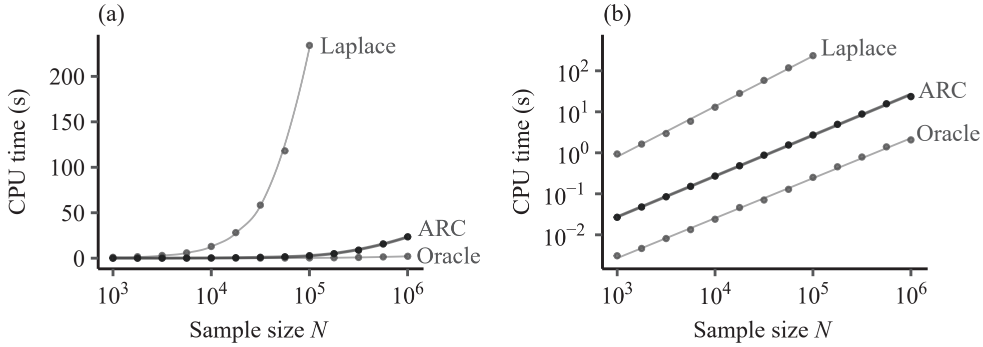

Fig. 2. (a) Computation time in seconds versus sample size N for the Imb-Nul-Hi setting for the Laplace, all-row-column and oracle methods. (b) The same plot with times displayed on the $\log_{10}$ scale.

is different from linear modelling, where estimation difficulty is unrelated to predictor coefficient values. The main reason to highlight this setting is that it illustrates an especially bad outcome for the Laplace method's estimate of σ_A. Similar, but less extreme, difficulties for the Laplace method's estimates of σ_A arise in the Imb-Nul-Lo setting.

3.2. *Computational cost*

Figure 2 shows the average computation times, in seconds, for the three methods, obtained using a 16-core 3.5 GHz AMD processor equipped with 128 GB of RAM. It also shows regression lines of log cost versus $\log N$, marked with the regression slopes. The all-row-column and oracle methods both have slopes that are nearly 1, as expected. The Laplace method's slope is clearly larger than 1, and as mentioned above, we curtailed the sample sizes used for that case to keep costs reasonable. If we extrapolate the Laplace cost to $N = 10^8$, comparable to the Netflix data (Bennett & Lanning, 2007), then the cost grows to more than 12.9 days, while the all-row-column cost grows only to about 45 minutes. The computational cost of the all-row-column method can be further reduced with parallel computing by distributing the approximations of the $R + C$ one-dimensional integrals across multiple cores.

Similar estimated computational costs were obtained for the all-row-column method in the other seven settings, as summarized in Table 2, which shows estimated computational cost rates very close to 1 for all eight cases. The oracle method's slope is consistently close to 1 in imbalanced settings and somewhat greater than 1 in balanced settings. The Laplace method's slope is consistently greater than 1; it tends to be higher for balanced settings, although one of the imbalanced cases also has a large slope.

We have investigated the data behind the oracle method slopes, but cannot yet explain the mild superlinearity that is sometimes seen. The number of Fisher-scoring iterations used by the oracle method varies at small sample sizes, but is consistently near 7 at larger sample sizes for which the superlinearity is more prominent. We have seen some outliers in the computation times at small sample sizes, but not at large sample sizes. Replacing the means at different N by medians does not materially change the slopes. We consider the amount of unexplained nonlinearity to be small, but not negligible. For example, when going from $N = 10^3$ to $N = 10^6$, a rate such as $N^{1.06}$ yields a roughly 1500-fold cost increase instead of the expected 1000-fold increase. The anomaly is concentrated in the balanced simulations, but knowing this has not been enough to identify the cause.

Table 2. *Computational cost for all eight settings: linear regression slopes for* $\log(CPU\ time\ in\ seconds)$ *versus* $\log(N)$

Setting	Oracle	All-row-column	Laplace
Bal-Nul-Hi	1.06	1.01	1.30
Bal-Nul-Lo	1.05	1.01	1.24
Bal-Lin-Hi	1.06	1.00	1.29
Bal-Lin-Lo	1.07	0.99	1.26
Imb-Nul-Hi	0.98	1.00	1.23
Imb-Nul-Lo	0.98	1.00	1.13
Imb-Lin-Hi	1.02	0.99	1.15
Imb-Lin-Lo	0.98	1.00	1.15

3.3. *Regression coefficient estimation*

Next, we turn to estimation of the regression coefficients, treating the intercept differently from the other coefficients. The intercept poses a challenge because it is somewhat confounded with the random effects. For instance, if we replace β_0 by $\beta_0 + \lambda$ while replacing a_i by $a_i - \lambda$, then the Y_{ij} are unchanged. Large λ would change $\bar{a} = (1/R)\sum_{i=1}^{R} a_i$ by an implausible amount that should be statistically detectable, given that $a_i \sim \mathcal{N}(0, \sigma_A^2)$. On the other hand, $|\lambda| = O(\sigma_A R^{-1/2})$ would be hard to detect statistically. The other regression parameters are not similarly confounded with main effects in our settings. Ghosh et al. (2022b) observed that a categorical predictor that is a function of just the row index i or just the column index j leads to a similar confounding.

Because of the confounding described above, we anticipate that the true mean square error rate for the intercept cannot be better than $O\{\min(R, C)\}$, which is $O(N^{-0.53})$ in our imbalanced settings and $O(N^{-0.56})$ in our balanced settings. For the other coefficients, $O(N^{-1})$ is not ruled out by this argument. In the Supplementary Material we report the mean square errors for the intercept and the coefficient of the first predictor estimated at different sample sizes, for the three estimators under study in the Imb-Nul-Hi setting. We present the plot for only the first predictor, because the mean square errors of the estimates of the seven regression coefficients were essentially equivalent. All three estimators show a mean square error very close to $O(N^{-1})$ for β_1. Where we anticipated a mean square error no better than $O(N^{-0.53})$ (imbalanced) or $O(N^{-0.56})$ (balanced) for the intercept, we saw slightly better mean square errors with slopes between -0.57 and -0.61, confirming our expectation that the intercept would be harder to estimate than the regression coefficients.

3.4. *Variance component estimation*

In this subsection we look at the estimation errors in the variance component parameters σ_A^2 and σ_B^2. The oracle method is given the true values of these parameters and so the comparison is only between the all-row-column and Laplace methods. For the variance parameter σ_A^2, the data only have R levels $a_1, \ldots, a_R$. If these were observed directly, then we could estimate σ_A^2 by $(1/R)\sum_{i=1}^{R} a_i^2$ and have a mean square error of $O(R^{-1})$. In practice, the a_i are obscured by the presence of the signal, the noise ε_{ij} and the other random effects b_j. Accordingly, the best rate we could expect for σ_A^2 is $O(R^{-1}) = O(N^{-\rho})$, and the best we could expect for σ_B^2 is $O(C^{-1}) = O(N^{-\kappa})$.

Figure 3 plots the mean square errors for estimation of σ_A and σ_B by the all-row-column and Laplace methods in the Imb-Nul-Hi setting. Because of the imbalance, our anticipated

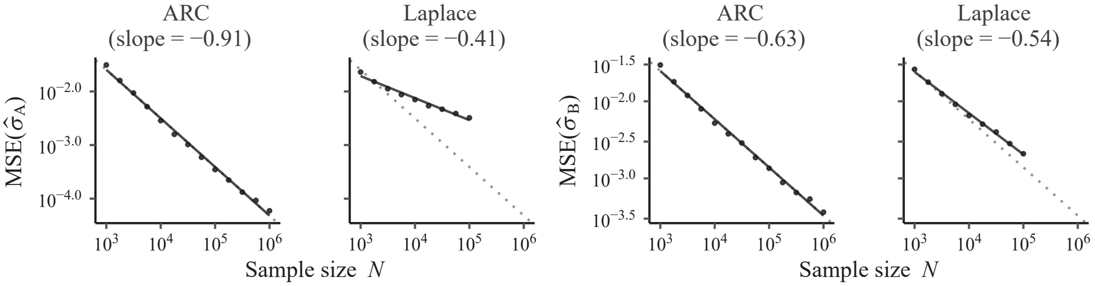

Fig. 3. Mean square errors and estimated convergence rates for σ_A and σ_B for the all-row-column (ARC) and Laplace methods in setting Imb-Nul-Hi. The dotted line in the plots for the Laplace method represents the convergence rate of the all-row-column method.

rates are $O(N^{-0.88})$ for σ_A and $O(N^{-0.53})$ for σ_B. The all-row-column method does slightly better than these rates. The Laplace method almost attains the predicted rate for σ_B, but does much worse for σ_A. We can understand both of these discrepancies in terms of biases, described next.

Figure 4 shows boxplots for the parameter estimates of σ_A and σ_B with the all-row-column and Laplace methods. We can compare the centres of these boxplots to the reference line at the true parameter values; this shows that the all-row-column estimates have a bias decreasing at a faster rate than the width of the boxes, which explains the slightly better-than-predicted rates seen for the all-row-column method. In contrast, the Laplace method has a substantial bias that only decreases very slowly as N increases, giving the Laplace method a worse-than-expected rate, especially for σ_A.

3.5. *Other settings*

The simulation results for all eight settings are reported in full in the Supplementary Material. We have already discussed the computational cost for the eight settings, referring to Table 2. Here we make some brief accuracy comparisons. We compare the proposed all-row-column method with the oracle method, which is infeasible because it requires knowledge of σ_A and σ_B, and with the Laplace method, which becomes infeasible for large N because it does not scale as $O(N)$. In most settings and for most parameters, the oracle method was slightly more accurate than the all-row-column method, but not always: the all-row-column method was slightly more accurate than the oracle method for β_1 in three of the eight settings, namely Bal-Lin-Hi (Figure S3), Bal-Lin-Lo (Figure S4) and Imb-Lin-Hi (Figure S7).

The figures in the Supplementary Material show some cases where the all-row-column method has a slight advantage over the Laplace method and some cases where it has a slight disadvantage. There are a small number of situations where the Laplace method has outliers at $N = 10^3$ that cause the linear regression slope to be questionable. These are present in the estimates of σ_A and σ_B for setting Imb-Lin-Hi (Figure S23). However, the attained mean square errors at $N = 10^5$ do not differ much between the all-row-column and Laplace methods in that setting (Figure S15). Our conclusion is that, compared to the Laplace method, the all-row-column method is scalable and robust.

In the Supplementary Material we also compare the all-row-column method with the maximum pairwise likelihood estimator of Bellio & Varin (2005). Their pairwise likelihood involves all the pairs of correlated observations, i.e., those pairs that share the row-

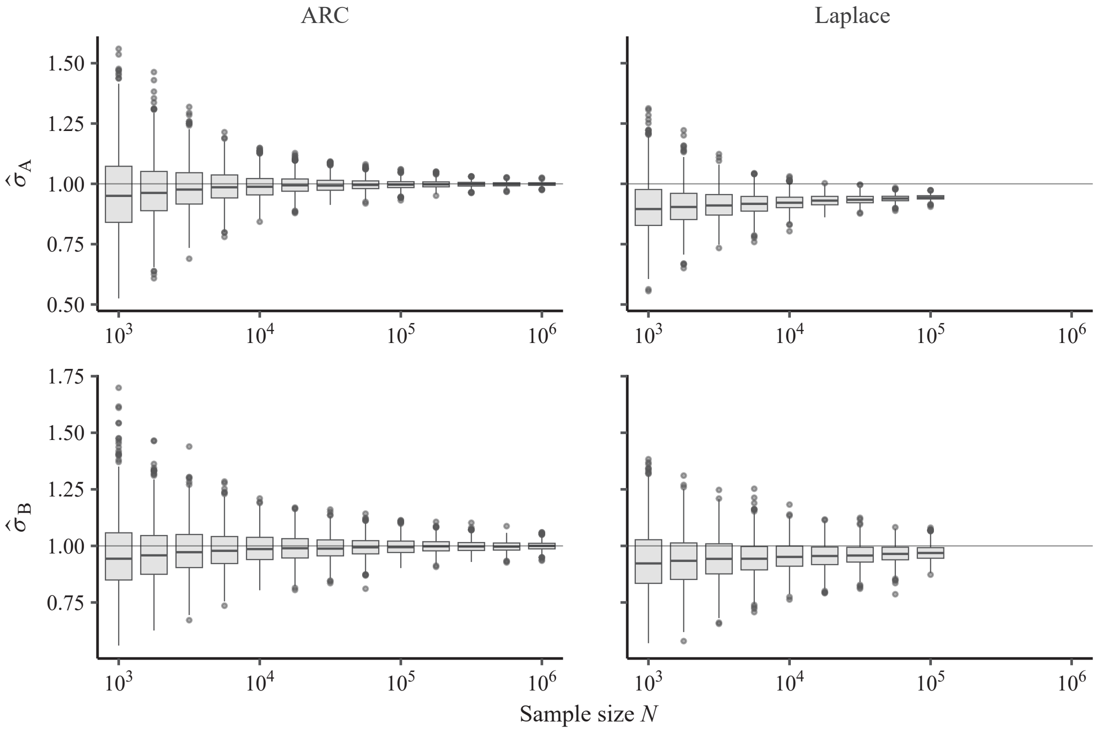

Fig. 4. Boxplots of σ_A and σ_B estimates for the all-row-column (ARC) and Laplace methods in the setting Imb-Nul-Hi. Horizontal reference lines are drawn at the true parameter values.

or the column-random effect. The pairwise likelihood is not computationally attractive. The number of pairs of data points far exceeds the $O(N)$ constraint. In a setting with R rows, the average row has N/R elements in it. If all the rows had that many elements, then the number of pairs would be $\Omega\{R(N/R)^2\} = \Omega(N^2/R)$. Unequal numbers of observations per row can only increase this count. As a result, the cost must be $\Omega\{\max(N^2/R, N^2/C)\} = \Omega\{\max(N^{2-\rho}, N^{2-\kappa})\}$, so it cannot be $O(N)$.

The comparison is made for the Imb-Nul-Hi setting that we have been focusing on. For that imbalanced setting, the cost of the pairwise likelihood is $\Omega(N^{2-0.53}) = \Omega(N^{1.47})$. We see in § S3 of the Supplementary Material that the empirical cost of the pairwise composite likelihood grows as $N^{1.46}$, close to the predicted rate. The pairwise method attains very similar parameter estimation accuracy to the all-row-column method. The most important difference in this example is that the all-row-column method costs $O(N)$, while pairwise likelihood is far more expensive and does not scale to large datasets.

4. Application to Stitch Fix data

In this section we illustrate the all-row-column method on a dataset from Stitch Fix. As described in Ghosh et al. (2022b):

> Stitch Fix is an online personal styling service. One of their business models involves sending customers a sample of clothing items. The customer may keep and purchase any of those items and return the others. They have provided us with some of their client ratings data. That data was anonymized, void of personally identifying information,

Table 3. *Predictors available in the Stitch Fix data*

Variable	Description	Levels
Client fit	Client fit profile	fitted loose or oversize straight or tight
Edgy	Edgy style?	yes/no
Boho	Bohemian style?	yes/no
Chest	Chest size	numeric
Size	Dress size	numeric
Material	Primary material of item	artificial fibre leather or animal fibre regenerated fibre vegetable fibre
Item fit	Fit of clothing item	fitted loose or oversized straight or tight

> 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 Stitch Fix data consist of $N = 5\,000\,000$ ratings from $R = 744\,482$ clients on $C = 3547$ items. The data also include client- and item-specific covariates. In the data, the binary response Y_{ij} of interest was whether customer i thought that item j was a top-rated fit or not, with $Y_{ij} = 1$ for an answer of 'yes'. The predictor variables we used are listed in Table 3. There is one block of client predictors followed by a block of item predictors. Some of the categorical variables in the data had only a small number of levels. The table shows how we aggregated them.

Some of the observations with $Z_{ij} = 1$ were nonetheless incomplete with a few missing entries. Deleting them left us with $N = 4\,965\,960$ ratings from 741 221 clients on 3523 items. The data are not dominated by a single row or column. The customer with the most records accounts for $N\epsilon_R$ records, where $\epsilon_R \doteq 1.25 \times 10^{-5}$. The item with the most records accounts for $N\epsilon_C$ records, with $\epsilon_C \doteq 2.77 \times 10^{-2}$. The data are sparse because $N/(RC) \doteq 1.9 \times 10^{-3}$.

In a business setting, one would fit and compare a wide variety of different binary regression models in order to understand the data. Our purpose here is to study large-scale probit models including crossed random effects, so we choose just one model for illustration, possibly the first model of many that one could consider. We consider the probit model with crossed random effects whose fixed effects are specified according to the symbolic model formula

$$\text{Top} \sim \text{Client fit} + \text{Edgy} + \text{Boho} + \text{Chest} + \text{Size} + \text{Material} + \text{Item fit},$$

where Top is the binary response variable for 'top-rated' described earlier. The model has $p = 12$ parameters for fixed effects, including the intercept. The first level of each categorical predictor in alphabetical order (Table 3) is used as the reference level in fitting the model.

Table 4 displays (i) the maximum likelihood estimates of the regression parameters under a naïve probit model that ignores the customer and item heterogeneity, and (ii) the all-row-column estimates for the probit model with two crossed random effects for the customers and items. The random effects probit parameter estimate $\hat{\beta}$ equals the naïve probit estimate $\hat{\gamma}$ multiplied by $(1 + \hat{\sigma}_A^2 + \hat{\sigma}_B^2)^{1/2}$. Using adaptive Gauss–Hermite quadrature we obtained

Table 4. *Stitch Fix binary regression results; all estimated predictor parameters are multiplied by 100; the first level of each categorical predictor in alphabetical order is used as the reference level*

		Naïve probit			Random effects probit		
Variable		Estimate	z-value	p-value	Estimate	z-value	p-value
Intercept		43.1	31.64	<0.001	50.9	10.52	<0.001
Client fit	loose or oversize	8.7	61.04	<0.001	10.3	10.81	<0.001
	straight or tight	5.1	34.30	<0.001	6.0	10.67	<0.001
Edgy	yes	−3.0	−25.08	<0.001	−3.5	−7.49	<0.001
Boho	yes	8.9	76.25	<0.001	10.5	25.81	<0.001
Chest		−0.5	−12.30	<0.001	−0.6	−7.32	<0.001
Size		0.2	10.94	<0.001	0.3	2.99	0.003
Material	leather or animal	−12.9	−12.99	<0.001	−15.2	−1.57	0.116
	regenerated	2.5	20.06	<0.001	3.0	0.65	0.516
	vegetable	−12.2	−58.39	<0.001	−14.5	−3.13	0.002
Item fit	loose or oversized	9.7	36.15	<0.001	11.4	1.78	0.075
	straight or tight	−2.1	−9.55	<0.001	−2.5	−0.67	0.500

estimates $\hat{\sigma}_A \doteq 0.53$ and $\hat{\sigma}_B \doteq 0.34$. Following what we learned from the simulation studies, the estimates on all rows and columns were calculated with $k = 5$ quadrature nodes. An earlier conservative calculation used $k = 28$, but $k = 5$ gave indistinguishable results.

The z-values reported in the table were computed with the observed information for the naïve probit model and with the two-way cluster-robust sandwich estimator described in § 2.3. As expected, ignoring the customer and item heterogeneity leads to large underestimation of the uncertainty in the parameter estimates, and hence in the naïve probit all the predictors are strongly significant, given the very large sample size. Conversely, the crossed random effects model takes into account the sources of heterogeneity and reveals that the item fit is not a significant predictor of top rank and that items made from vegetable fibres are less likely to be ranked as top than clothes made with artificial fibres.

In the Supplementary Material, Figure S29 compares the two-way cluster-robust sandwich standard errors with (i) the standard errors from the naïve probit fit multiplied by $(1 + \hat{\sigma}_A^2 + \hat{\sigma}_B^2)^{1/2} \doteq 1.18$, to report them in the fixed effects scale of the probit model with crossed random effects, and (ii) the pigeonhole nonparametric bootstrap standard errors of Owen (2007), which does not assume correct model specification, as mentioned in § 2.3. Figure S29 shows how closely the sandwich and pigeonhole standard errors agree. The naïve standard errors, which ignore dependence between items and customers, correspond to variances underestimated by factors ranging from 3 to 954, depending on the parameter and only slightly on whether we use sandwich or pigeonhole estimates of the coefficient variances. Thus, ignoring the dependencies from correlated data makes an enormous difference here. These standard errors are reported in Table S1 of the Supplementary Material. We also computed parametric bootstrap standard errors (not shown); these were somewhat lower than the nonparametric standard errors, as expected.

In an application such as the Stitch Fix data analysis, the typical goal is to make inference about the probability of an item being ranked top for a specific customer and a specific item. Such evaluations also require estimating the customer (a_i) and item (b_j) random effects. The estimates of those random effects are a byproduct of the adaptive Gauss–Hermite quadrature used to approximate the row- and column-wise likelihoods. Figure S30 in the Supplementary Material shows the distribution of the estimated customer and item random effects.

5. DISCUSSION

In Theorem 2 we proved that there exists a root of the row (or column) likelihood equation that is a consistent estimator of τ_A^2 (or τ_B^2). This form of consistency is commonly called Cramer consistency. It is the same notion of consistency that Jiang (2013) established for the maximum likelihood estimate. If one does not find Cramer consistency sufficient, then it is possible to construct estimators $\hat{\tau}_A^2$ and $\hat{\tau}_B^2$ that converge in probability to τ_A^2 and τ_B^2 as $N \to \infty$. A consistent estimator of τ_A^2 can be obtained from one or more rows i for which $N_{i\bullet} \to \infty$, and a similar approach works for τ_B^2. In § S1.3 of the Supplementary Material we show how to construct such a consistent estimator, and we give conditions under which the number of large rows will diverge to infinity as $N \to \infty$; see Theorem 3 there. We prefer our all-row-column estimator to an approach using just large rows, because it would be awkward to have to decide in practice which rows to use, and also because we believe that there is valuable information in the other, smaller, rows.

We have used a probit model instead of a logistic one because a Gaussian latent variable is a very natural counterpart to the Gaussian random effects that are the default in random effects models. This connection simplified our modelling and computation. Gibbons & Hedeker (1997) made the following remark: 'As in the case of fixed effect models, selection of probit versus logistic response functions appears to have more to do with custom or practice within a particular discipline than differences in statistical properties.' An extension to logistic regression is outside the scope of this paper.

We are confident that our approach of combining multiple misspecified models can be extended to other settings with Gaussian random effects and latent variables. The code we use already handles ordinal regression. Extensions to more than two effects or multivariate responses may well work similarly, but are beyond the scope of this article.

We conclude with some additional references about recent work on inference for data with a crossed design. Goplerud et al. (2025) developed a variational approximation for scalable Bayesian estimation using an appropriate relaxing of the mean-field assumption to avoid underestimation of posterior uncertainty in high dimensions.

Xu et al. (2023) combined variational approximations and composite likelihoods that consider row-column decomposition in a similar way to ours. Their approach is particularly convenient for Poisson and gamma regression models, because in these cases analytical calculations allow the approximation of one-dimensional integrals that appear in the composite likelihood to be avoided. In the binary case considered in the present article, the approach of Xu et al. (2023) requires numerical integration and its consistency has not yet been established.

Hall et al. (2020) considered message-passing algorithms for generalized linear mixed models, and their § 6 includes a crossed effects binary regression. However, it has $R = 10$ and $C = 6$ in our notation, with three replicates at each (i, j) pair, and it does not address scalability. Ruli et al. (2016) proposed an improved version of the Laplace approximation to overcome the potential failure of the usual Laplace approximation and also illustrated it in the case of generalized linear models with crossed random effects in their § 3.5. However, their proposal is not scalable, as illustrated by the numerical results in Ruli et al. (2016).

Bartolucci et al. (2017) considered another composite likelihood that combines a row likelihood with a column likelihood to estimate a hidden Markov model for two-way data arrays. This approach shares the philosophy of our method, but differs from ours in terms of the model (latent discrete Markov variables versus crossed random effects), fitting procedure (EM algorithm versus direct maximization), data structure (balanced versus unbalanced and sparse) and motivating application (genomics versus e-commerce).

Acknowledgement

We are grateful to Silvia Bianconcini, Blair Bilodeau, Nancy Reid, Alex Stringer, Yanbo Tang and Matt Wand for helpful discussions about generalized linear mixed models, Laplace approximation and adaptive Gauss–Hermite quadrature. We thank Bradley Klingenberg and Stitch Fix for making the dataset available. This work was supported in part by grants IIS-1837931 and DMS-2152780 from the U.S. National Science Foundation, IRIDE from Ca' Foscari University of Venice, and the research project *Latent Variable Models for Complex Data* founded by the European Union – NextGenerationEU (MUR DM funds 737/2021).

Supplementary material

The Supplementary Material contains proofs of the theorems in § 2, additional simulation results discussed in § 3, further references to the literature, and a table and two plots about the Stitch Fix application mentioned in § 4. `R` code for replicating our results is available from the public repository `https://github.com/rugbel/arcProbit`.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Bartolucci, F., Chiaromonte, F., Don, P. K. & Lindsay, B. G. (2017). Composite likelihood inference in a discrete latent variable model for two-way ‘clustering-by-segmentation' problems. *J. Comp. Graph. Statist.* **26**, 388–402.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Software* **67**, 1–48.

Bellio, R. & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statist. Mod.* **5**, 217–27.

Bennett, J. & Lanning, S. (2007). The Netflix prize. In Proc. KDD Cup and Workshop 2007 (San Jose, California, August 12).

Bilodeau, B., Stringer, A. & Tang, Y. (2024). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *J. Am. Statist. Assoc.* **119**, 690–700.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Cameron, C. A., Gelbach, J. B. & Miller, D. L. (2011). Robust inference with multiway clustering. *J. Bus. Econ. Statist.* **29**, 238–49.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York: Springer.

Edelman, A. & Rao, N. R. (2005). Random matrix theory. *Acta Numer.* **14**, 233–97.

Gao, K. & Owen, A. B. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electron. J. Statist.* **11**, 1235–96.

Gao, K. & Owen, A. B. (2020). Estimation and inference for very large linear mixed effects models. *Statist. Sinica* **30**, 1741–71.

Ghandwani, D., Ghosh, S., Hastie, T. & Owen, A. B. (2023). Scalable solution to crossed random effects model with random slopes. *arXiv:* 2307.12378.

Ghosh, S., Hastie, T. & Owen, A. B. (2022a). Backfitting for large scale crossed random effects regressions. *Ann. Statist.* **50**, 560–83.

Ghosh, S., Hastie, T. & Owen, A. B. (2022b). Scalable logistic regression with crossed random effects. *Electron. J. Statist.* **16**, 4604–35.

Ghosh, S. & Zhong, C. (2021). Convergence rate of a collapsed Gibbs sampler for crossed random effects models. *arXiv:* 2109.02849.

Gibbons, R. D. & Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics* **53**, 1527–37.

Goplerud, M., Papaspiliopoulos, O. & Zanella, G. (2025). Partially factorized variational inference for high-dimensional mixed models. *Biometrika* **112**, asae067.

Hall, P., Johnstone, I. M., Ormerod, J. T., Wand, M. P. & Yu, J. C. F. (2020). Fast and accurate binary response mixed model analysis via expectation propagation. *J. Am. Statist. Assoc.* **115**, 1902–16.

Jiang, J. (2013). The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *Ann. Statist.* **41**, 177–95.

Jiang, J., Wand, M. P. & Bhaskaran, A. (2022). Usable and precise asymptotics for generalized linear mixed model analysis and design. *J. R. Statist. Soc. B* **84**, 55–82.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *J. Statist. Software* **70**, 1–21.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221–39.

Liu, Q. & Pierce, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81**, 624–9.

Lumley, T. & Mayer Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. UW Biostatistics Working Paper Series no. 207, University of Washington.

Lyu, Z., Sission, S. A. & Welsh, A. H. (2024). Increasing dimension asymptotics for two-way crossed mixed effect models. *Ann. Statist.* **52**, 2956–78.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Boca Raton, Florida: Chapman & Hall/CRC.

Miglioretti, D. L. & Heagerty, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5**, 381–98.

Ogden, H. (2021). On the error in Laplace approximations of high-dimensional integrals. *Stat* **10**, e380.

Owen, A. B. (2007). The pigeonhole bootstrap. *Ann. Appl. Statist.* **1**, 386–411.

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

Papaspiliopoulos, O., Stumpf-Fétizon, T. & Zanella, G. (2023). Scalable Bayesian computation for crossed and nested hierarchical models. *Electron. J. Statist.* **17**, 3575–612.

R Development Core Team (2025). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Ruli, E., Sartori, N. & Ventura, L. (2016). Improved Laplace approximation for marginal likelihoods. *Electron. J. Statist.* **10**, 3986–4009.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–27.

Shun, Z. & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *J. R. Statist. Soc. B* **57**, 749–60.

Stiratelli, R., Laird, N. & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–71.

Tang, Y. & Reid, N. (2025). Laplace and saddlepoint approximations in high dimensions. *Bernoulli* **31**, 1759–88.

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

Xu, L., Reid, N. & Kong, D. (2023). Gaussian variational approximation with composite likelihood for crossed random effect models. *arXiv:* 2310.12485.

[*Received on* 7 *August* 2024. *Editorial decision on* 28 *April* 2025]