Supplementary material for "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 CA 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

S1. Proofs of theorems

We report here the proofs of Theorems 1 and 2 given in §3 of the manuscript and discuss the construction of an alternative composite likelihood estimator of the variance components using all of the data from a subset of rows with a large number observations. Recall that $Z_{ij} = 1$ if (x_{ij}, Y_{ij}) is observed (0 otherwise) with $N_{i\bullet} = \sum_j Z_{ij}$, and $N_{\bullet j} = \sum_i Z_{ij}$. We defined $\epsilon_R = \max_i N_{i\bullet}/N$ and $\epsilon_C = \max_j N_{\bullet j}/N$.

S1.1. *Weak consistency of $\hat{\gamma}$*

Theorem S1. *Let $Y_{ij} \in \{0, 1\}$ follow the crossed random effects probit model (manuscript Equation 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 $\max(\epsilon_R, \epsilon_C) \to 0$. Let $x_{ij} \in \mathbb{R}^p$ satisfy*

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

Let $\hat{\gamma} \in \mathbb{R}^p$ be any maximizer of the all likelihood (manuscript Equation 8). Then for any $\epsilon > 0$

$$\Pr(\|\hat{\gamma} - \gamma_0\| > \epsilon) \to 0$$

as $N \to \infty$.

Proof. The proof uses the arguments of Lumley & Mayer Hamblett (2003) to prove the consistency of the root of quasi-score equations in marginal generalized linear models with

sparse correlations. Write the all log likelihood as $\ell_{\mathrm{all}}(\gamma) = \sum_{ij} Z_{ij}\eta_{ij}$ for random variables

$$\eta_{ij}(\gamma) = Y_{ij}\log\Phi(x_{ij}^{\mathrm{T}}\gamma) + (1-Y_{ij})\log\Phi(-x_{ij}^{\mathrm{T}}\gamma).$$

Under our assumptions, the $\eta_{ij}(\gamma)$ are uniformly bounded in i, j and γ. Letting B be that bound

$$\begin{aligned}
\mathrm{var}\Big\{\frac{\ell_{\mathrm{all}}(\gamma)}{N}\Big\} &\le \frac{B^2}{N^2}\sum_{ij}\sum_{rs} Z_{ij}Z_{rs}\Big(1\{i=r\}+1\{j=s\}-1\{i=r\}1\{j=s\}\Big)\\
&= \frac{B^2}{N^2}\Big(\sum_{ijs} Z_{ij}Z_{is} + \sum_{ijr} Z_{ij}Z_{rj} - N\Big)\\
&\le \frac{B^2}{N^2}\Big(\sum_{i} N_{i\bullet}^2 + \sum_{j} N_{\bullet j}^2\Big)\\
&\le \frac{B^2}{N^2}\Big(\sum_{i} N_{i\bullet}N\epsilon_R + \sum_{j} N_{\bullet j}N\epsilon_C\Big)\\
&\le B^2(\epsilon_R+\epsilon_C)\\
&\to 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{\ell_{\mathrm{all}}(\gamma)}{N} \to \frac{\mathrm{E}\{\ell_{\mathrm{all}}(\gamma)\}}{N} &= \frac{1}{N}\sum_{ij} Z_{ij}\mathrm{E}\Big\{\Phi(x_{ij}^{\mathrm{T}}\gamma_0)\log\Phi(x_{ij}^{\mathrm{T}}\gamma) + \Phi(-x_{ij}^{\mathrm{T}}\gamma_0)\log\Phi(-x_{ij}^{\mathrm{T}}\gamma)\Big\}\\
&=: g(\gamma)
\end{aligned}$$

in probability as $N\to\infty$, for all $\gamma\in\mathbb{R}^p$. For all N, $\ell_{\mathrm{all}}(\gamma)/N$ is a concave function and $g(\cdot)$ is also a concave function. Therefore from Appendix II of Andersen & Gill (1982), $\hat\gamma$ is weakly consistent for the maximizer of g. Now $\partial g(\gamma_0)/\partial\gamma = 0$ so γ_0 is a maximizer of g. For our conclusion, we require a unique minimizer. For that, we write

$$\frac{\partial^2 g(\gamma_0)}{\partial\gamma\partial\gamma^{\mathrm{T}}} = -\frac{1}{N}\sum_{ij} Z_{ij}\mathrm{E}\Big\{\frac{\varphi(x_{ij}^{\mathrm{T}}\gamma_0)^2}{\Phi(x_{ij}^{\mathrm{T}}\gamma_0)\Phi(-x_{ij}^{\mathrm{T}}\gamma_0)}x_{ij}x_{ij}^{\mathrm{T}}\Big\} =: H.$$

Because every $x_{ij}^{\mathrm{T}}\gamma$ are uniformly bounded it follows that the Hessian above satisfies $aV \preceq -H \preceq AV$ for some constants $0<a<A<\infty$. As a result, g is strictly concave and $\hat\gamma$ converges weakly to γ_0. □

We know that we need $\min(R,C)\to\infty$ in order to have a consistent estimate of an intercept term, should the model contain one. In the above theorem the assumption that $\epsilon_R\to 0$ while $N\to\infty$ already implies $R\to\infty$ and similarly the condition on ϵ_C implies that $C\to\infty$.

S1.2. *Consistency of $\hat\sigma_A^2$ and $\hat\sigma_B^2$*

Theorem S2. *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 a root of the column likelihood equation that is a consistent estimator for τ_B^2.*

Proof. For a subset $\mathcal{S}_1\subseteq\mathcal{S}$ of observations, define

$$\widetilde{N}_{i\bullet} = \sum_{j=1}^{C} Z_{ij}1\{(i,j)\in\mathcal{S}_1\},\quad \widetilde{N}_{\bullet j} = \sum_{i=1}^{R} Z_{ij}1\{(i,j)\in\mathcal{S}_1\}. \tag{S1}$$

We seek such a set $\mathcal{S}_1$ where $\widetilde{N}_{\bullet j} \in \{0, 1\}$ for all $j = 1, \ldots, C$ and $\widetilde{N}_{i\bullet} \neq 1$ for any $i = 1, \ldots, N$. We let R_1 be the number of rows i with $\widetilde{N}_{i\bullet} > 0$. Then all such rows have $\widetilde{N}_{i\bullet} \geq 2$ and no column in the data set contributes to the data in $\mathcal{S}_1$ for two or more of those rows. We then have $\widetilde{N}_A = \sum_i \widetilde{N}_{i\bullet}$ independent observations from data following the row likelihood given in Equation (9) of the manuscript. If $\widetilde{N}_A \to \infty$, then the maximizer $\hat{\tau}_A^2$ of the row likelihood based only on $(i, j) \in \mathcal{S}_1$ corresponds to the MLE based on a sample of $\widetilde{N}_A$ independent observations from an equicorrelated probit model and thus provides consistent estimation of τ_A^2. See Section 5 of Ochi & Prentice (1984) for the asymptotic properties of the MLE in the equicorrelated probit model.

Next we use the subset argument of Jiang (2013) to show Cramer consistency of the row likelihood MLE. It amounts to showing that the full likelihood is not worse than the likelihood on $\mathcal{S}_1$ only. Following Jiang (2013) we let $y_{[1]}$ be the Y_{ij} values for $(i, j) \in \mathcal{S}_1$ and $y_{[2]}$ be the Y_{ij} values for $(i, j) \in \mathcal{S} \setminus \mathcal{S}_1$. Let $P_{\tau_A^2}(\cdot)$ denote the joint probability of a collection of Y_{ij} values computed at the true parameter τ_A^2. In our context this is a conditional probability given the corresponding x_{ij} values as shown in Equation (9) of the manuscript.

Now pick $\epsilon > 0$ and let $Y_{[2]}$ be the (finite) set of all possible values of $y_{[2]}$. Then

$$\begin{aligned}
\Pr\{P_{\tau_A^2}(y_{[1]}, y_{[2]}) \leq P_{\tau_A^2+\epsilon}(y_{[1]}, y_{[2]}) \mid y_{[1]}\} &= \Pr\left\{ \frac{P_{\tau_A^2+\epsilon}(y_{[1]}, y_{[2]})}{P_{\tau_A^2}(y_{[1]}, y_{[2]})} \geq 1 \;\middle|\; y_{[1]} \right\} \\
&\leq \mathrm{E}\left\{ \frac{P_{\tau_A^2+\epsilon}(y_{[1]}, y_{[2]})}{P_{\tau_A^2}(y_{[1]}, y_{[2]})} \;\middle|\; y_{[1]} \right\} \\
&= \sum_{y_{[2]} \in Y_{[2]}} \frac{P_{\tau_A^2+\epsilon}(y_{[1]}, y_{[2]})}{P_{\tau_A^2}(y_{[1]}, y_{[2]})} p_{\tau_A^2}(y_{[2]} \mid y_{[1]}) \\
&= \sum_{y_{[2]} \in Y_{[2]}} \frac{P_{\tau_A^2+\epsilon}(y_{[1]}, y_{[2]})}{P_{\tau_A^2}(y_{[1]})} \\
&= \frac{P_{\tau_A^2+\epsilon}(y_{[1]})}{P_{\tau_A^2}(y_{[1]})}.
\end{aligned}$$

The outer probability above is the conditional probability under τ_A^2. This last ratio converges to zero in probability as described in Jiang (2013) for a regular likelihood and $\widetilde{N}_A \to \infty$. The same argument can be used for $\tau_A^2 - \epsilon$ and $\tau_B^2 \pm \epsilon$. Cramer consistency now follows from the dominated convergence theorem and standard arguments (e.g., Jiang, 2010, pp. 9–10). That is, there is a root of the row likelihood equation that is consistent for τ_A^2 and a root of the column likelihood equation that is consistent for τ_B^2. □

S1.3. *An alternative 'large rows' estimator*

If one does not find Cramer consistency sufficient, then it is possible to construct a consistent estimator using all of the data from a subset of rows with a large number $N_{i\bullet}$ of observations. We call those 'large rows'. Let $\theta_{ij} = x_{ij}^{\mathrm{T}}\gamma$. From Equation (4) of the manuscript, it follows that

$$\Pr(Y_{ij} = 1 \mid a_i) = \Phi(\delta_{0i} + \delta_{1i}\theta_{ij}) \tag{S2}$$

where $\delta_{0i} = a_i/(1 + \sigma_B^2)^{1/2}$ and $\delta_{1i} = (1 + \tau_A^2)^{1/2}$. As $N_{i\bullet} \to \infty$ the MLEs of this model will be consistent so long as the variance of θ_{ij} is not too small. It is enough to have that variance bounded away from zero and infinity. See Section 9.2.2 of Amemiya (1985). While we do not have θ_{ij} we

do have consistent estimates $\hat{\theta}_{ij} = x_{ij}^{\mathrm{T}}\hat{\gamma}$ of it. As a result, we can obtain a consistent estimate of τ_A^2 from probit regression within one single row of data so long as $N_{i\bullet} \to \infty$ for that row.

This estimator fits in with the composite likelihood approach because composite likelihoods are often based on conditional likelihoods. The intercept in (S2) provides a consistent estimate of $a_i/(1+\sigma_B^2)^{1/2}$. If we square this we get a consistent estimate of τ_A^2 which we could then average over rows. However the slope δ_{1i} is more attractive as it estimates the same quantity in every row. Under standard assumptions, we have

$$\hat{\delta}_{1i} = (1+\tau_A^2)^{1/2} + O_p(N_{i\bullet}^{-1/2}).$$

Then we can get a consistent estimate of τ_A^2 from just one row with $N_{i\bullet} \to \infty$ or by pooling consistent estimates of τ_A^2 from many such rows. The analogous operation on columns gives us an estimate of τ_B^2. Next we give conditions under which the number of large rows will diverge to infinity as $N \to \infty$.

THEOREM S3. *Let there be $R = N^\rho$ rows and $C = N^\kappa$ columns in the data set with $0 < \rho, \kappa < 1$. Assume that $\epsilon_R = \max_i N_{i\bullet}/N \to 0$ and $\epsilon_C = \max_j N_{\bullet j}/N \to 0$. If $\alpha > 0$ satisfies $\rho + \alpha < 1$ then the number of rows i with $N_{i\bullet} > N^\alpha$ goes to infinity as $N \to \infty$. If $\alpha > 0$ satisfies $\kappa + \alpha < 1$ then the number of columns j with $N_{\bullet j} > N^\alpha$ goes to infinity as $N \to \infty$.*

Proof. We prove the case for rows as the argument for columns is the same. The average number of observations per row is $N^{1-\rho}$. Let R_0 be the number of rows i with $N_{i\bullet} < N^\alpha$ and let $R_1 = R - R_0$ be the number of rows i with $N_{i\bullet} \geq N^\alpha$. We need to show that $R_1 \to \infty$.

If $R_0 = 0$ then $R_1 = N^\rho \to \infty$. If $R_0 > 0$ then we may define $\bar{N}_0$ to be the average value of $N_{i\bullet}$ for rows with $N_{i\bullet} < N^\alpha$. By the pigeonhole principle, there must be at least one row with $N_{i\bullet} \geq N^\alpha$ and so let $\bar{N}_1$ be the average value of $N_{i\bullet}$ for rows with $N_{i\bullet} \geq N^\alpha$. Solving $R = R_0 + R_1$ and $N = R_0\bar{N}_0 + R_1\bar{N}_1$ together yields

$$R_1 = \frac{N - R\bar{N}_0}{\bar{N}_1} \geq \frac{N - RN^\alpha}{\bar{N}_1} \geq \frac{N - RN^\alpha}{\epsilon_R N} = \frac{1}{\epsilon_R}\big(1 - N^{\alpha+\rho-1}\big).$$

Therefore the number of rows with $N_{i\bullet} \geq N^\alpha$ diverges to infinity as $N \to \infty$. □

Every very large row provides a consistent estimate of τ_A^2 and aggregating consistent estimates by a weighted average would remain consistent without regard to the correlation pattern among those estimates. We still require $\max(\epsilon_R, \epsilon_C) \to \infty$ in order to get weak consistency of $\hat{\gamma}$. The large row argument lets us replace Cramer consistency of τ_A^2 and τ_B^2 by convergence in probability for them.

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 because we believe that there is valuable information in the other rows. Under the independent but not identically distributed Bernoulli model for Z_{ij} from Ghosh et al. (2022), every row eventually is large enough to provide a consistent estimate.

S2. FULL SET OF SIMULATION RESULTS

The following pages show the graphs that summarize the eight settings of the simulation study described in the article. For each simulation setting there are three graphs: 1. mean square errors of β_0 and β_1 estimates; 2. mean square errors of σ_A and σ_B estimates; 3. sample distributions of σ_A and σ_B estimates. The graphs corresponding to setting IMB-NUL-HI are also shown in the

article as Figures 2-4. In these supplemental materials, the captions are kept to a minimum. For more details on the graphs, see the captions of Figures 2-4 in the article.

S2.1. *Mean Square Errors of $\hat{\beta}$*

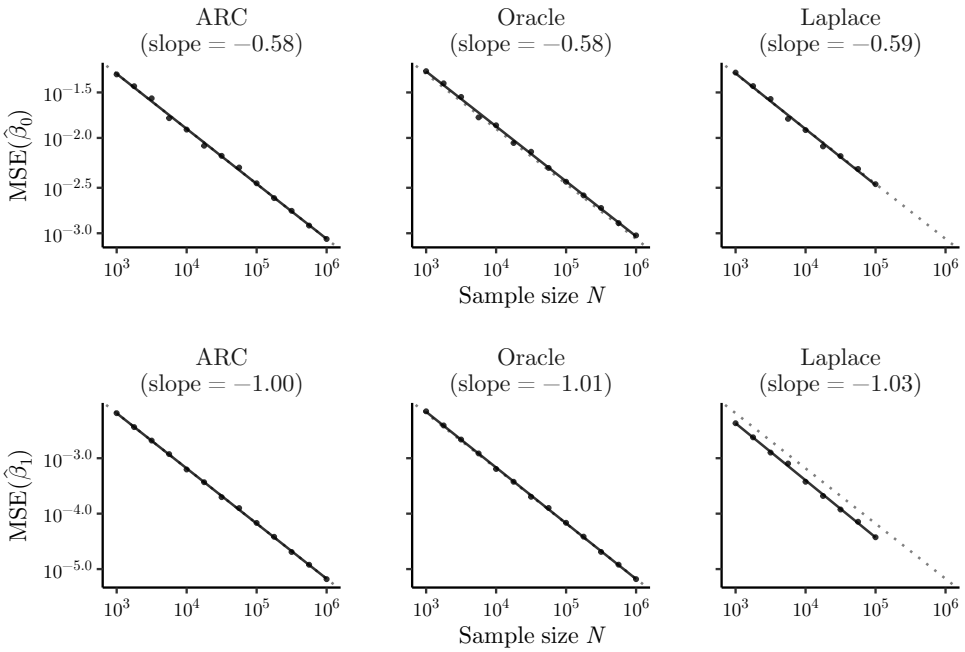

Fig. S1. Mean square errors of β_0 and β_1 estimates for setting BAL-NUL-HI.

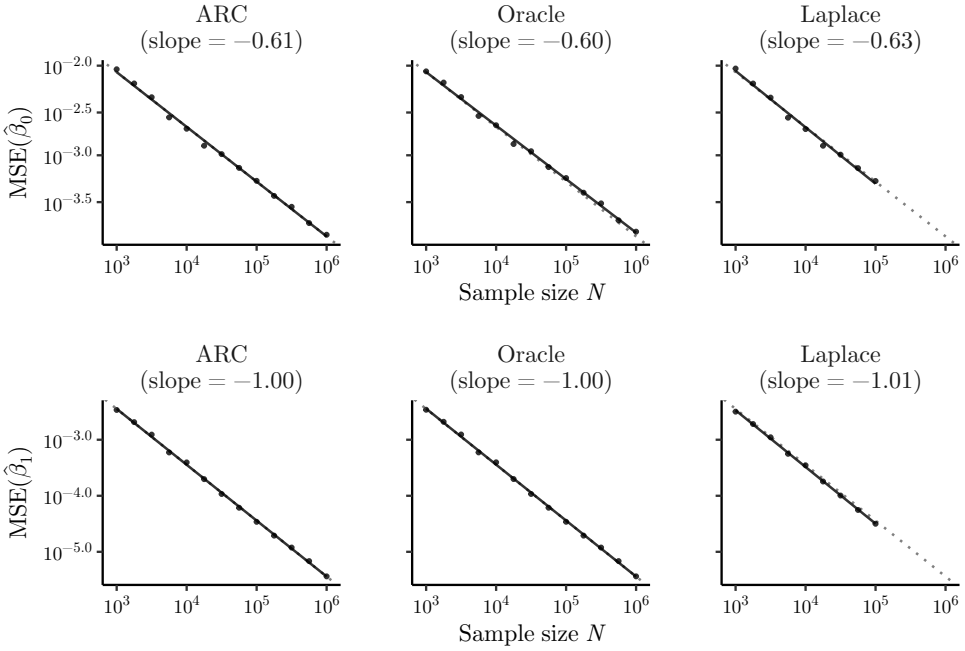

Fig. S2. Mean square errors of β_0 and β_1 estimates for setting BAL-NUL-LO.

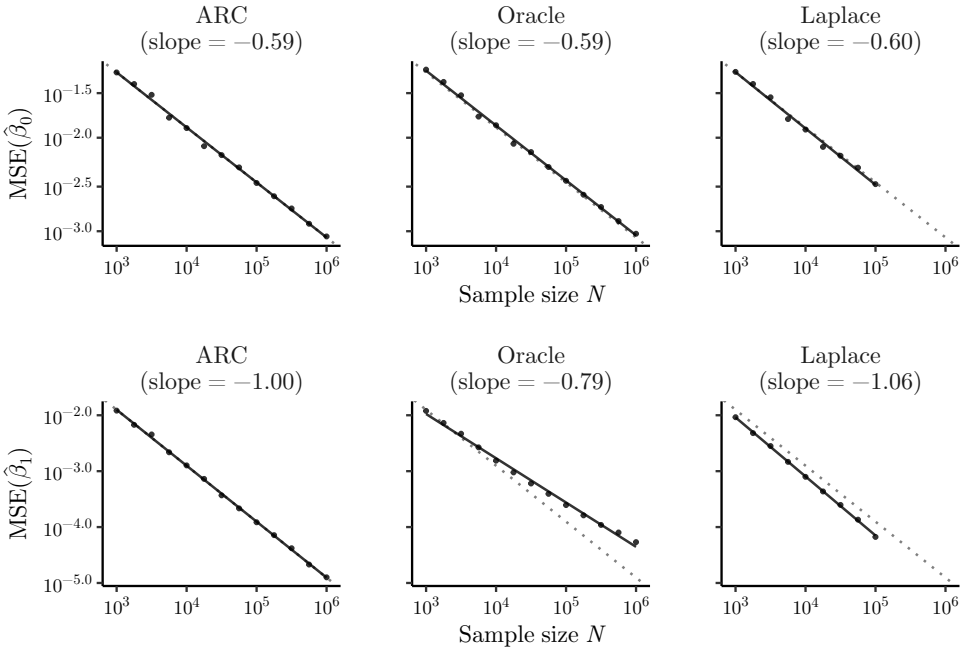

Fig. S3. Mean square errors of β_0 and β_1 estimates for setting BAL-LIN-HI.

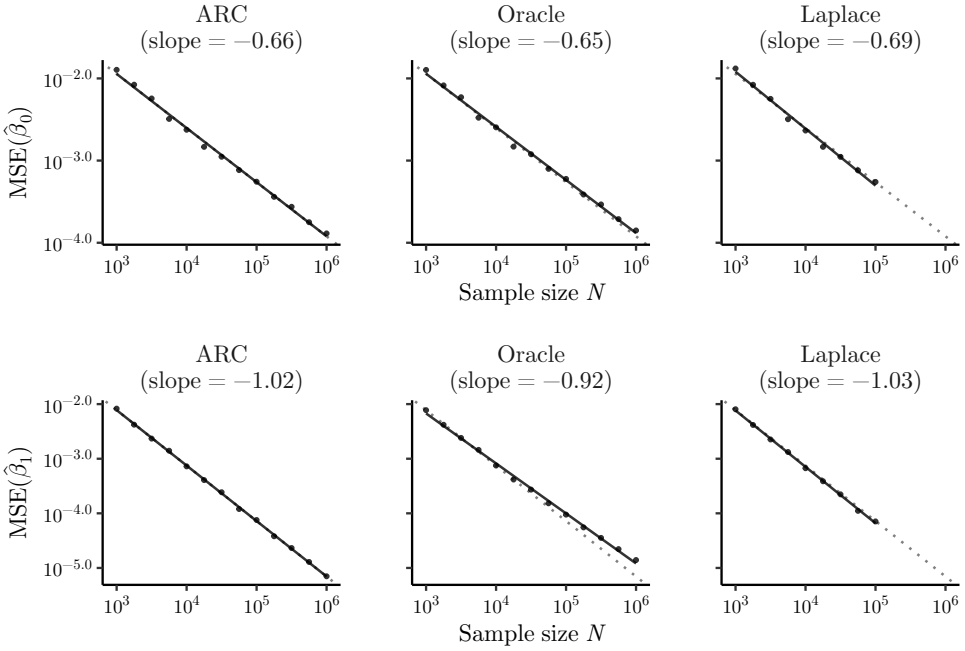

Fig. S4. Mean square errors of β_0 and β_1 estimates for setting BAL-LIN-LO.

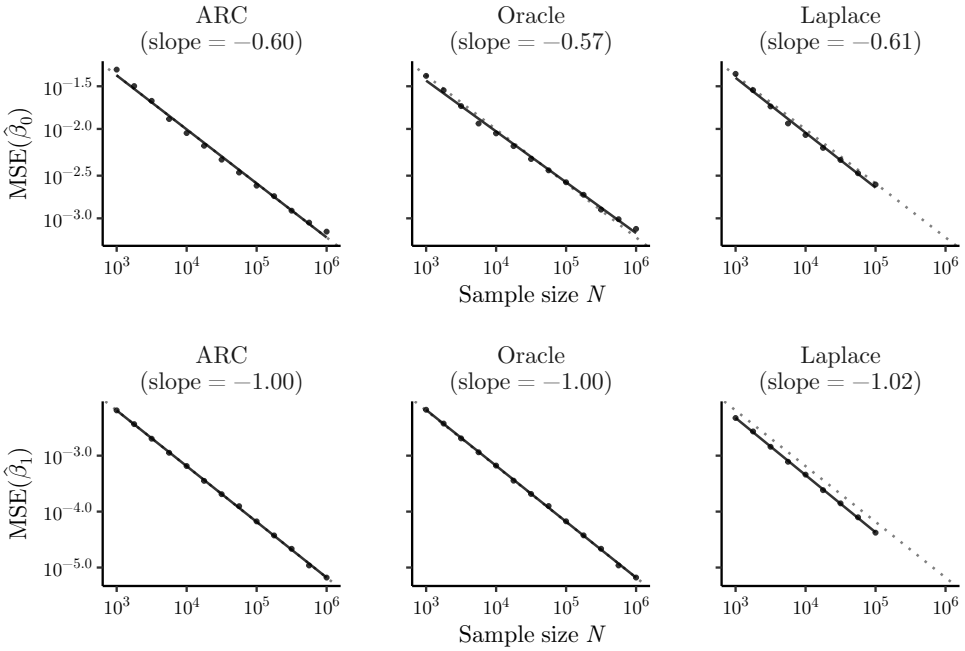

Fig. S5. Mean square errors of β_0 and β_1 estimates for setting Imb-Nul-Hi.

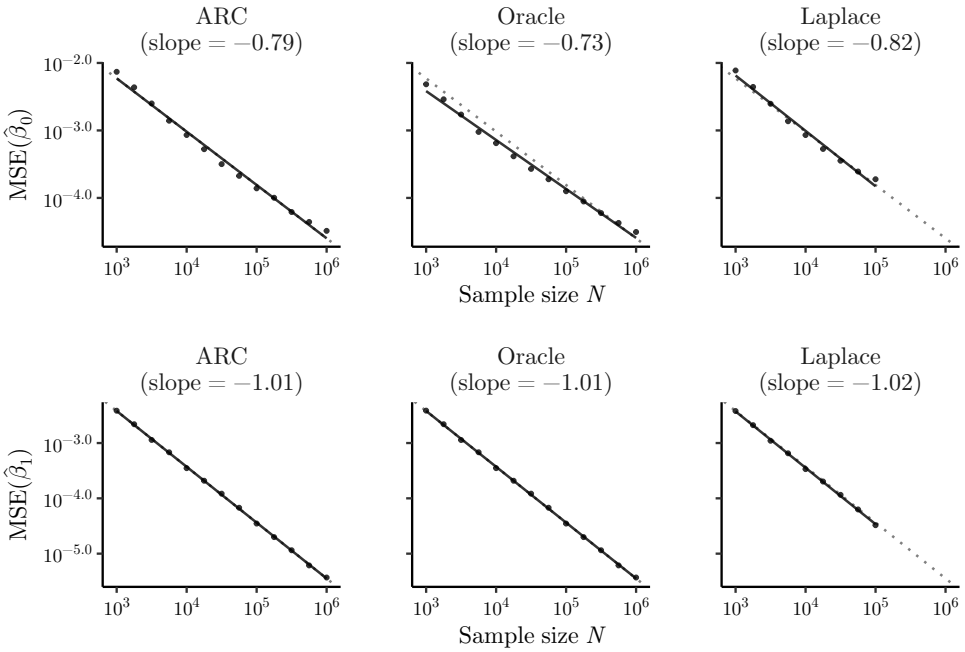

Fig. S6. Mean square errors of β_0 and β_1 estimates for setting Imb-Nul-Lo.

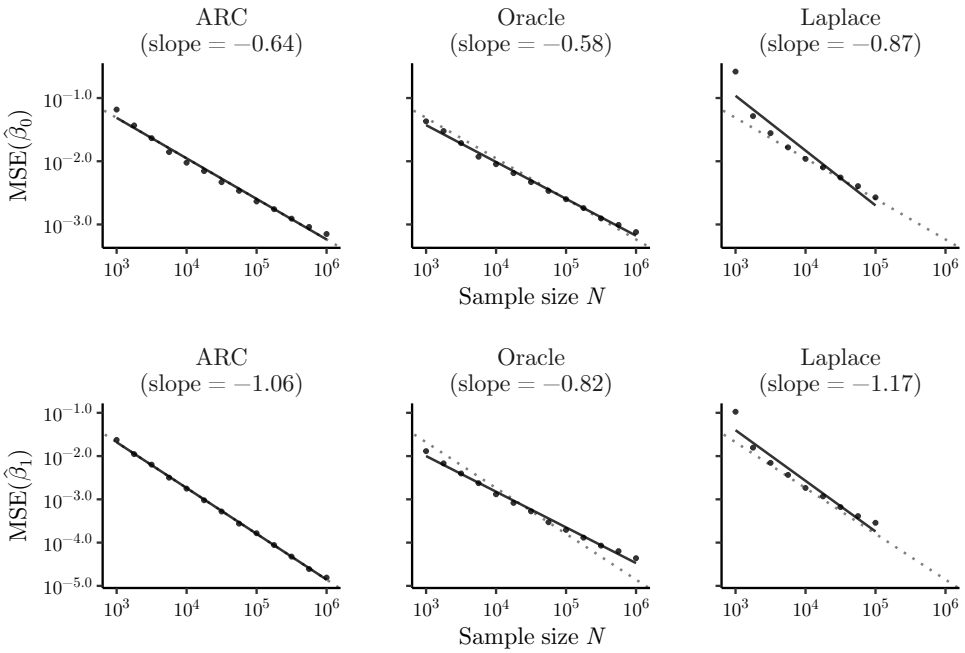

Fig. S7. Mean square errors of β_0 and β_1 estimates for setting Imb-Lin-Hi.

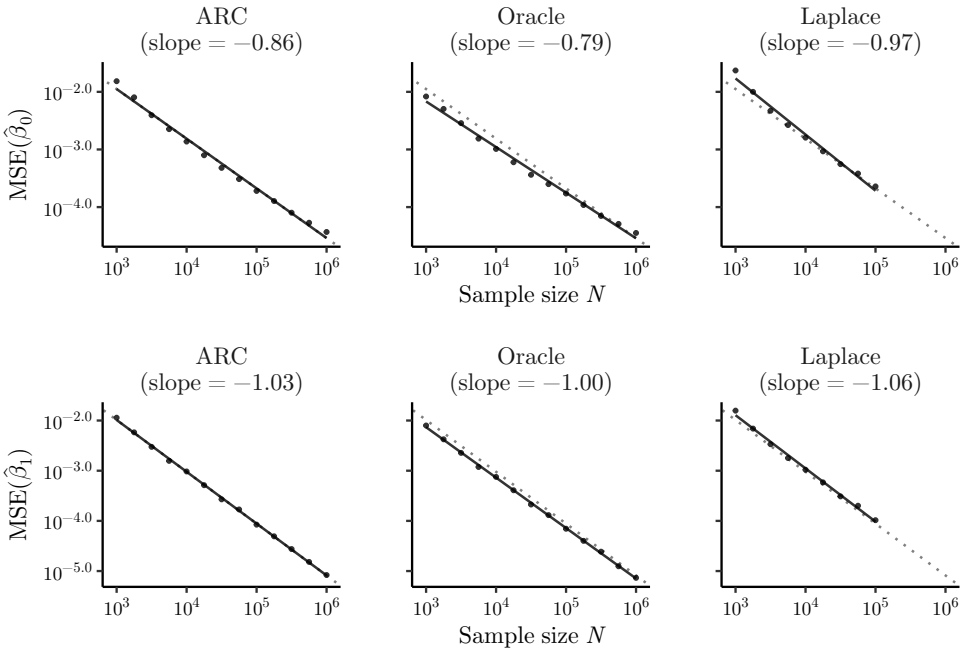

Fig. S8. Mean square errors of β_0 and β_1 estimates for setting Imb-Lin-Lo.

S2.2. *Mean Square Errors of $\hat{\sigma}_A$ and $\hat{\sigma}_B$*

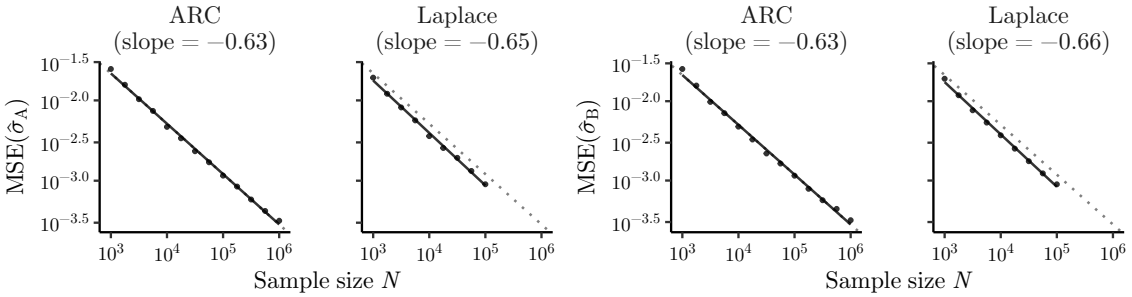

Fig. S9. Mean square errors of σ_A and σ_B estimates for setting BAL-NUL-HI.

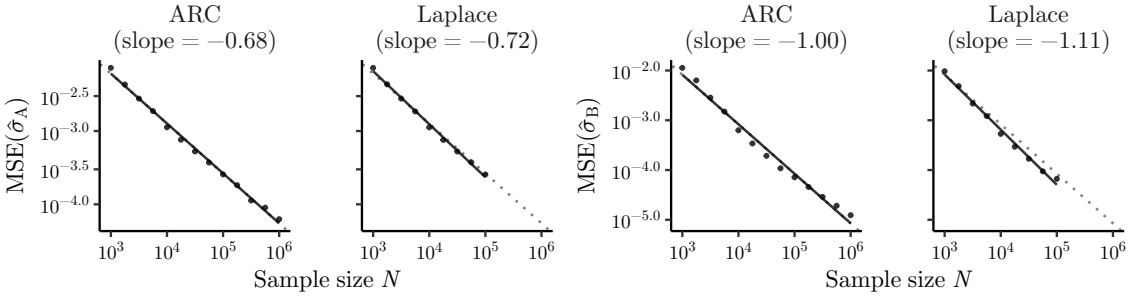

Fig. S10. Mean square errors of σ_A and σ_B estimates for setting BAL-NUL-LO.

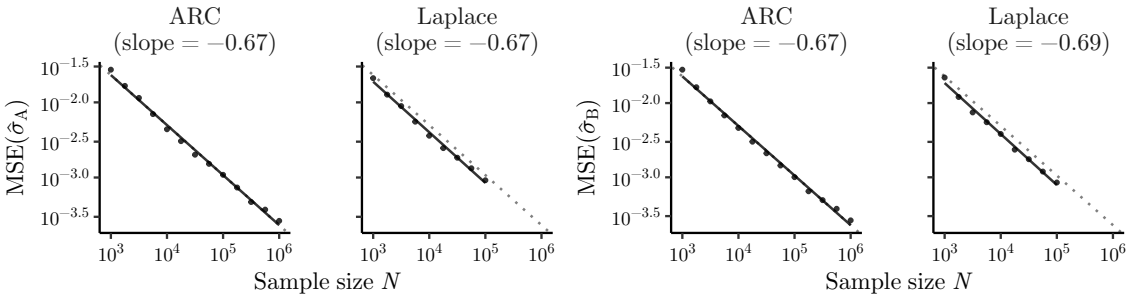

Fig. S11. Mean square errors of σ_A and σ_B estimates for setting BAL-LIN-HI.

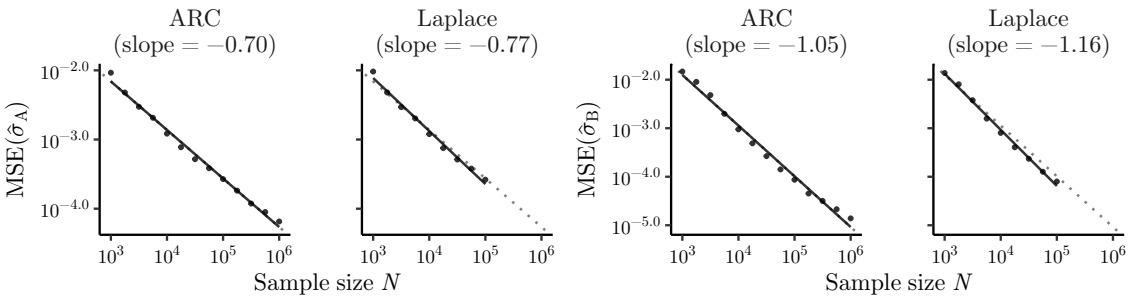

Fig. S12. Mean square errors of σ_A and σ_B estimates for setting BAL-LIN-LO.

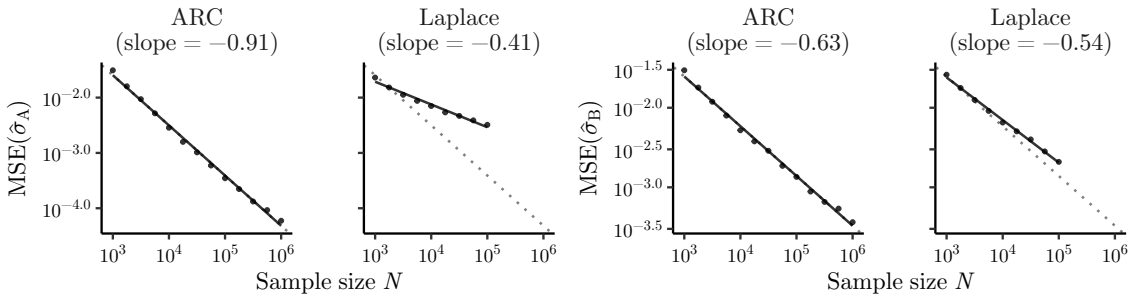

Fig. S13. Mean square errors of σ_A and σ_B estimates for setting Imb-Nul-Hi.

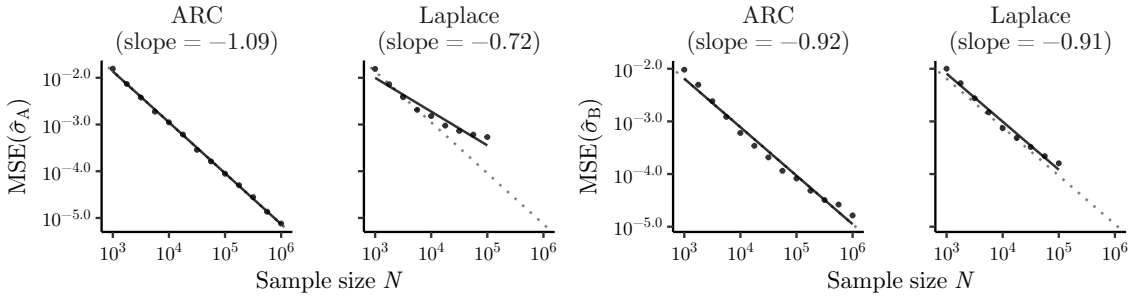

Fig. S14. Mean square errors of σ_A and σ_B estimates for setting Imb-Nul-Lo.

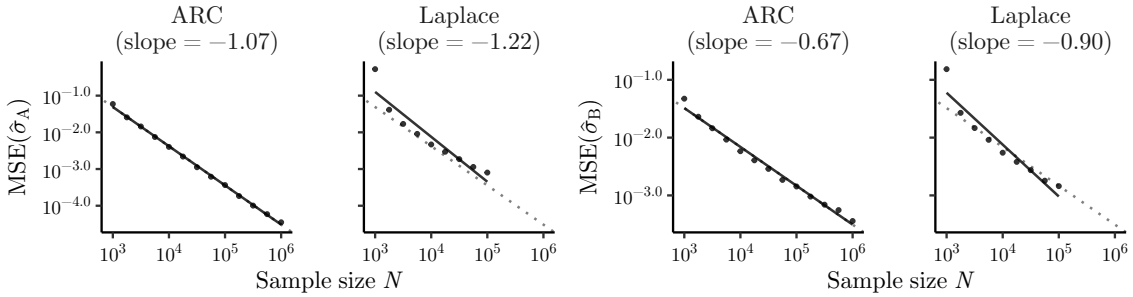

Fig. S15. Mean square errors of σ_A and σ_B estimates for setting Imb-Lin-Hi.

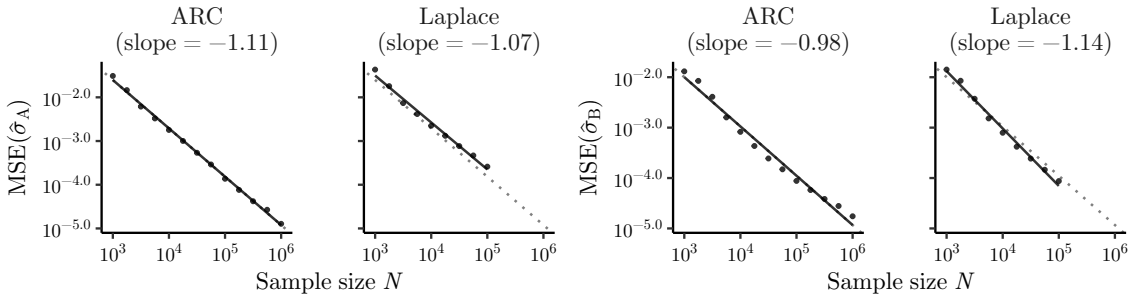

Fig. S16. Mean square errors of σ_A and σ_B estimates for setting Imb-Lin-Lo.

S2.3. *Sample Distributions of $\hat{\sigma}_A$ and $\hat{\sigma}_B$*

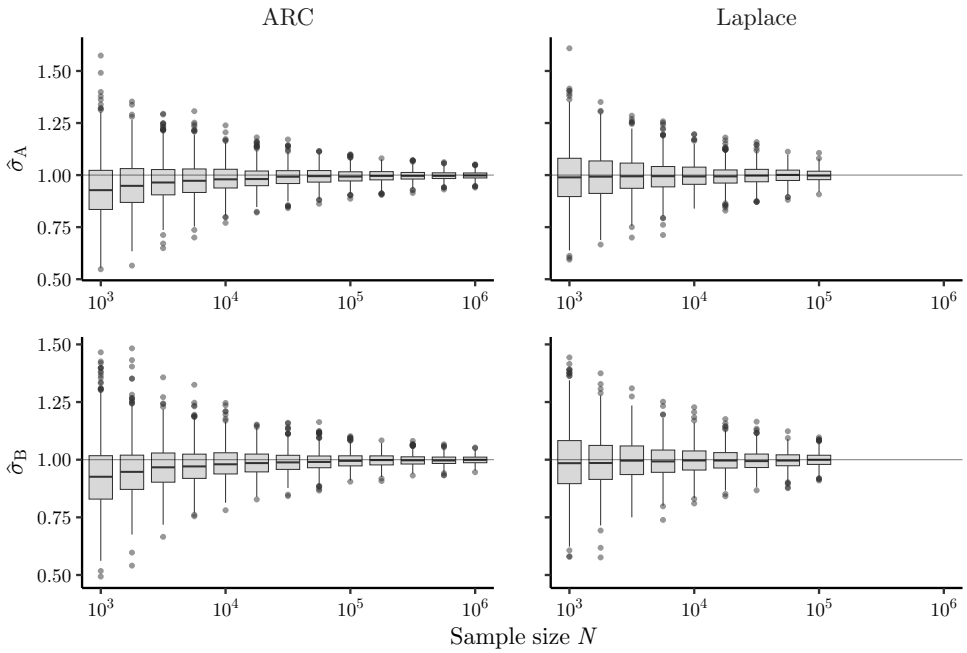

Fig. S17. Boxplots of σ_A and σ_B estimates for setting BAL-NUL-HI.

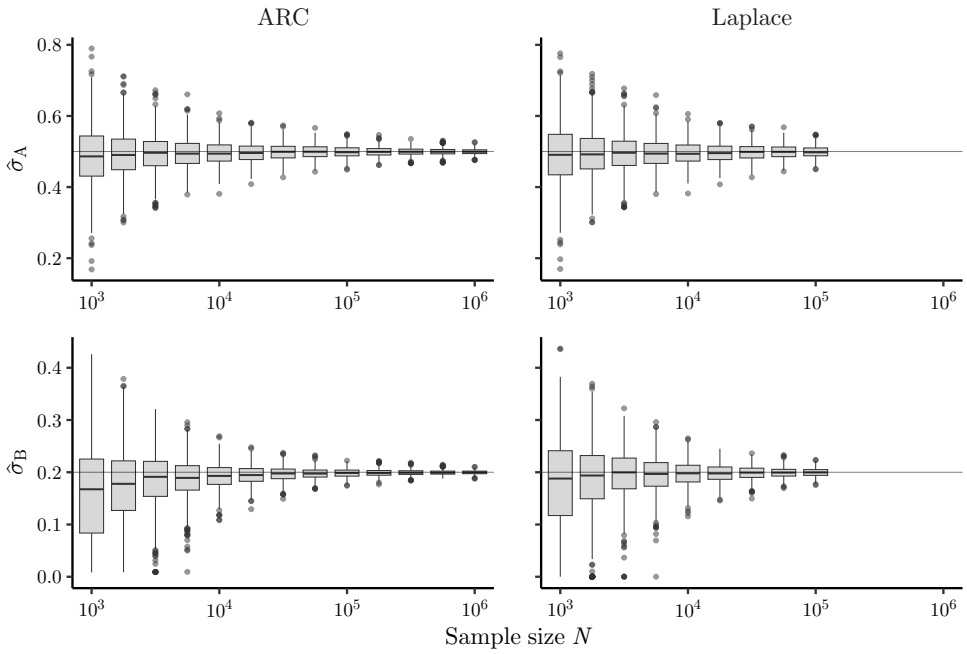

Fig. S18. Boxplots of σ_A and σ_B estimates for setting BAL-NUL-LO.

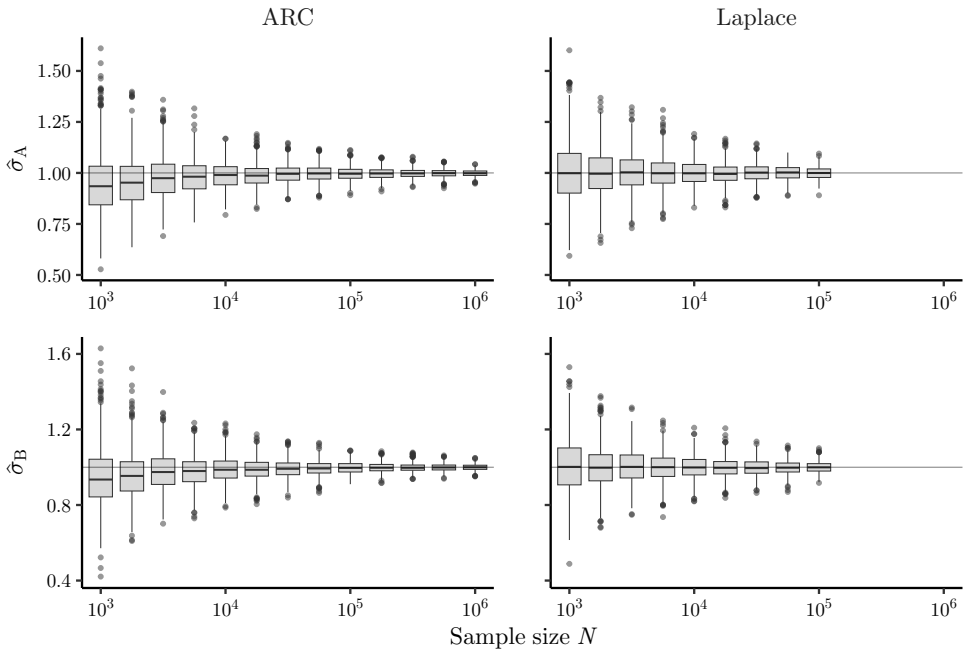

Fig. S19. Boxplots of σ_A and σ_B estimates for setting BAL-LIN-HI.

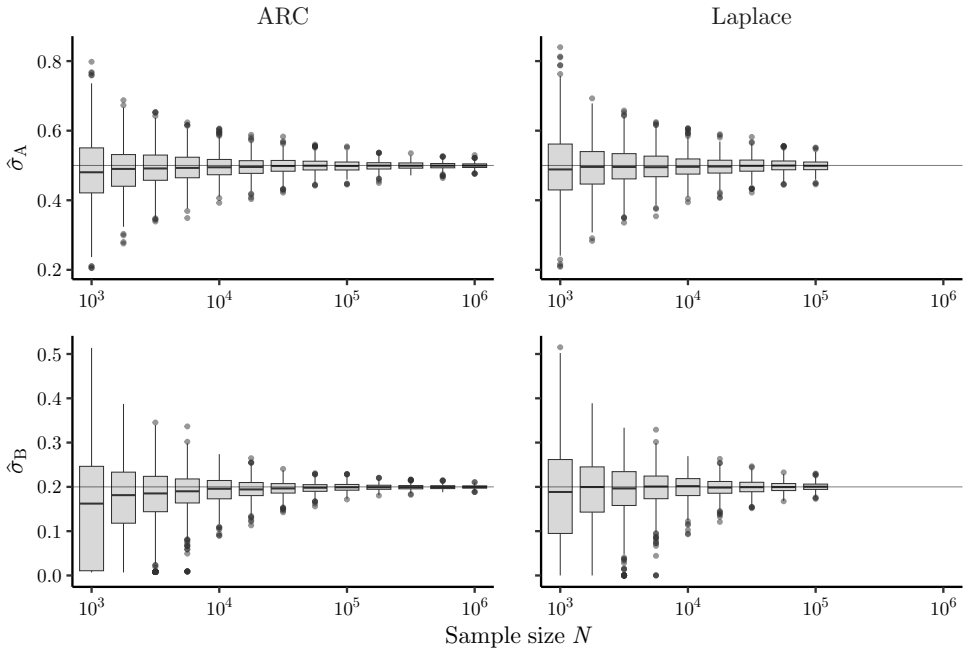

Fig. S20. Boxplots of σ_A and σ_B estimates for setting BAL-LIN-LO.

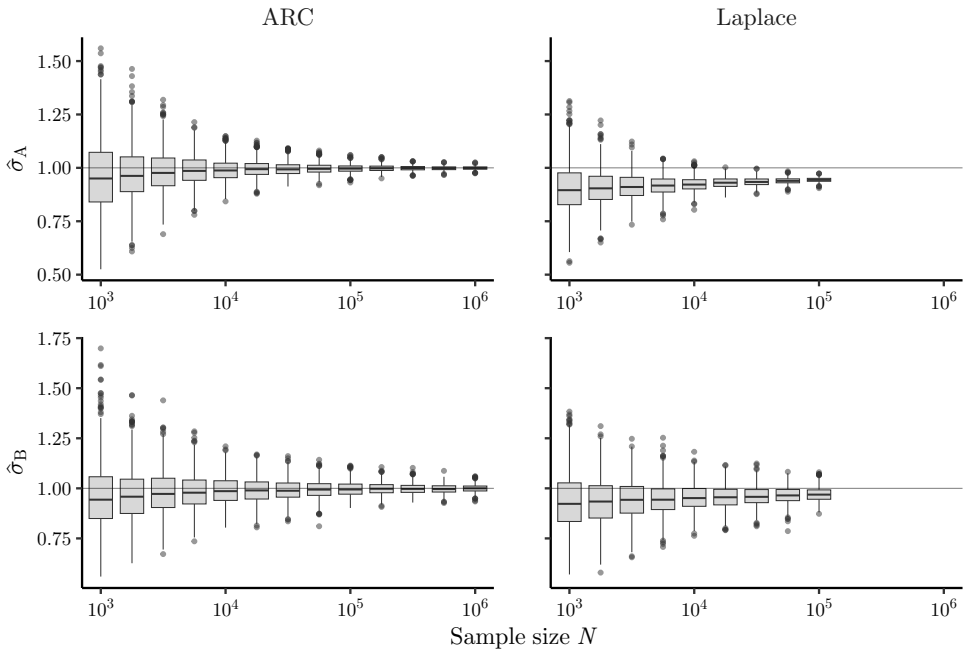

Fig. S21. Boxplots of σ_A and σ_B estimates for setting Imb-Nul-Hi.

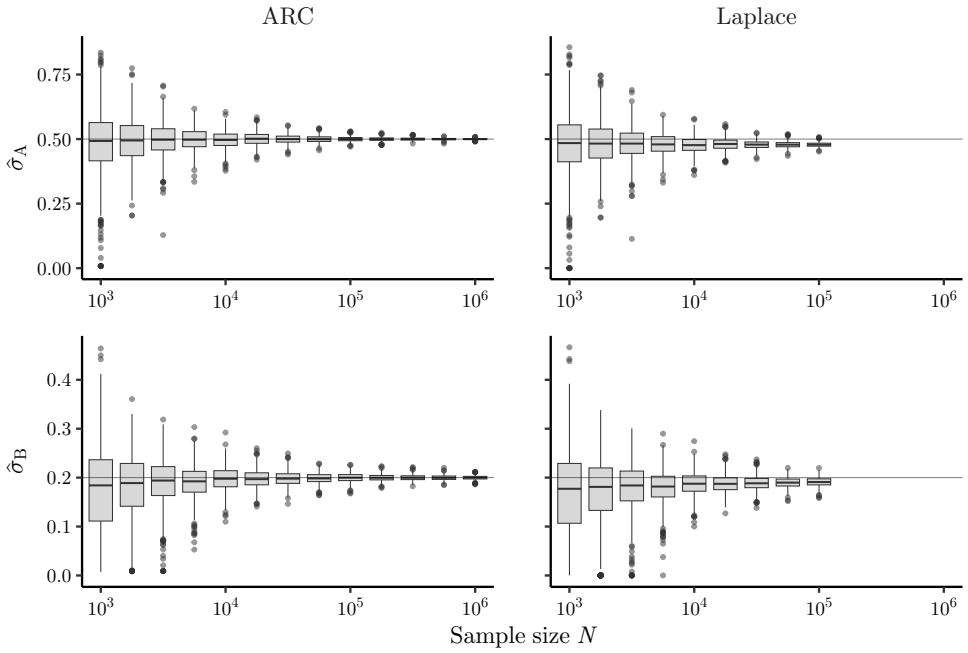

Fig. S22. Boxplots of σ_A and σ_B estimates for setting Imb-Nul-Lo.

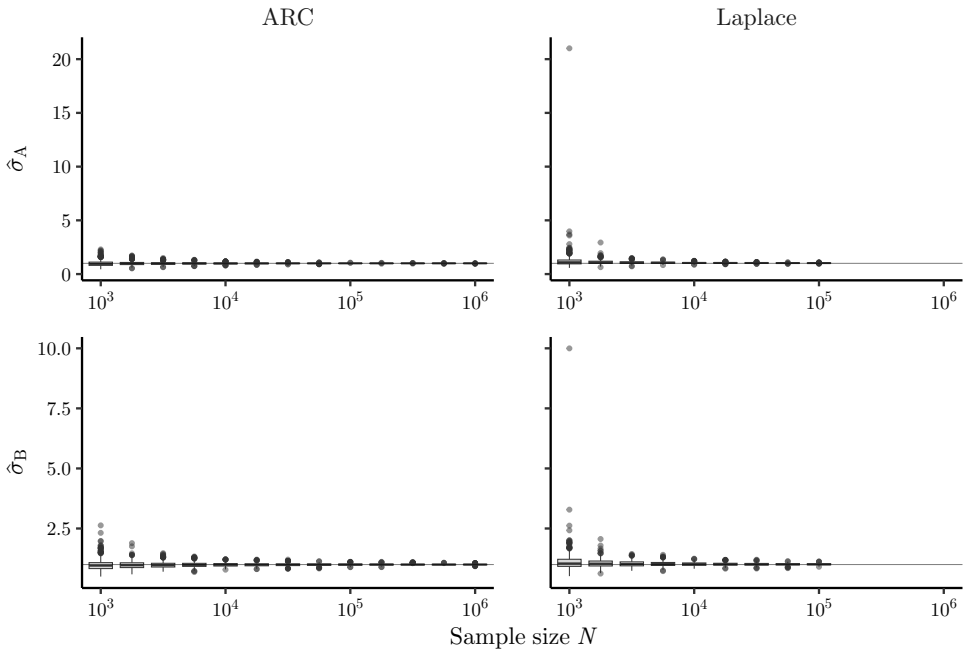

Fig. S23. Boxplots of σ_A and σ_B estimates for setting Imb-Lin-Hi.

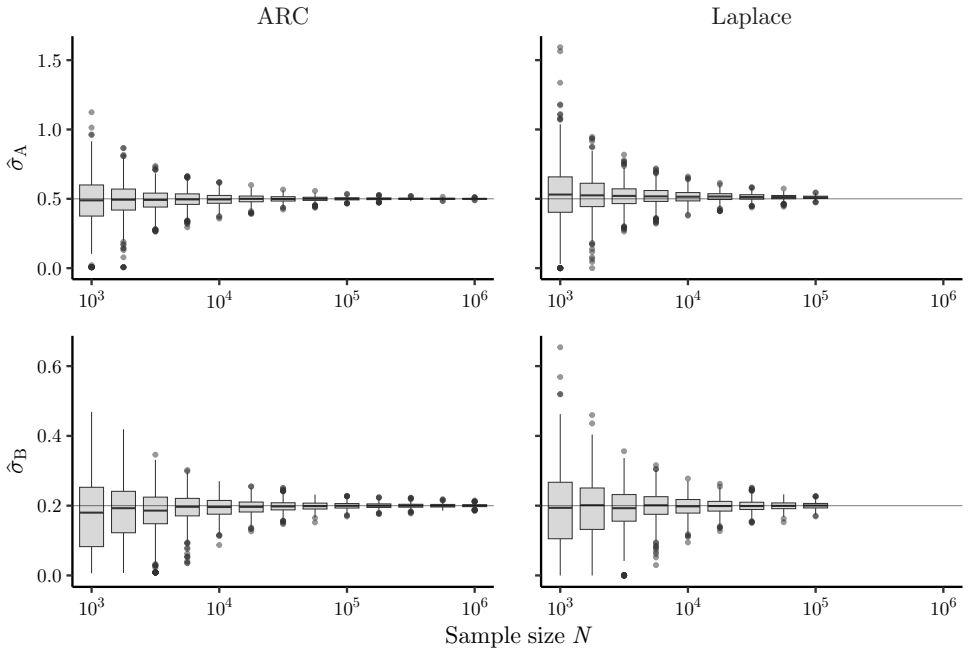

Fig. S24. Boxplots of σ_A and σ_B estimates for setting Imb-Lin-Lo.

S3. Maximum pairwise likelihood estimation

We consider here an alternative estimator given by the pairwise likelihood estimation introduced in Bellio & Varin (2005). Their method was proposed for a logit link, but it readily applies to probit models. Here we investigate a particular version of pairwise likelihood, along the lines of the all-row-column method. We leave the all part unchanged, and employ the pairwise likelihood approach to replace the row-wise and column-wise likelihood for the estimation of the two variance components. We estimate σ_A^2 and σ_B^2 by considering all the pairs sharing the same random effects for rows or for columns. Letting $\rho_A = \sigma_A^2/(1+\sigma_A^2+\sigma_B^2)$ be the intra-cluster correlation for row effects, we define the row-pairwise likelihood as

$$L_{2;\,\mathrm{row}}(\hat{\gamma}, \rho_A) = \prod_{i=1}^{R} \prod_{(j,k)\,|\,i} p(y_{ij}, y_{ik}; \hat{\gamma}, \rho_A), \tag{S3}$$

where $(j,k) \mid i$ is the set of pairs of column indices for observations with row index i. The bivariate probabilities involved are

$$p(y_{ij}, y_{ik}; \gamma, \rho_A) = \begin{cases} \Phi_2(x_{ij}^{\mathrm{T}}\gamma, x_{ik}^{\mathrm{T}}\gamma; \rho_A), & \text{if } y_{ij}=1, y_{ik}=1, \\ \Phi(x_{ij}^{\mathrm{T}}\gamma) - \Phi_2(x_{ij}^{\mathrm{T}}\gamma, x_{ik}^{\mathrm{T}}\gamma; \rho_A), & \text{if } y_{ij}=1, y_{ik}=0, \\ \Phi(x_{ik}^{\mathrm{T}}\gamma) - \Phi_2(x_{ij}^{\mathrm{T}}\gamma, x_{ik}^{\mathrm{T}}\gamma; \rho_A), & \text{if } y_{ij}=0, y_{ik}=1, \\ 1 - \Phi(x_{ij}^{\mathrm{T}}\gamma) - \Phi(x_{ik}^{\mathrm{T}}\gamma) + \Phi_2(x_{ij}^{\mathrm{T}}\gamma, x_{ik}^{\mathrm{T}}\gamma; \rho_A), & \text{if } y_{ij}=0, y_{ik}=0, \end{cases}$$

with $\Phi_2(x, y; \rho)$ denoting the bivariate normal cumulative distribution function for standardized marginals with correlation equal to ρ. The column-pairwise likelihood is then defined in a similar manner, that is

$$L_{2;\,\mathrm{col}}(\hat{\gamma}, \rho_B) = \prod_{i=1}^{C} \prod_{(j,k)\,|\,i} p(y_{ji}, y_{ki}; \hat{\gamma}, \rho_B), \tag{S4}$$

where $(j,k) \mid i$ is the set of pairs of row indices for observations with column index i, and $\rho_B = \sigma_B^2/(1+\sigma_A^2+\sigma_B^2)$. As for the all-row-column method of the main paper, this strategy would provide a consistent estimator, but with an excessive computational cost. As noted in the main article, the number of observation pairs to consider grows as $\Omega\{\max(N^2/R, N^2/C)\}$ when there are N observations in R rows and C columns. For $R, C = o(N)$, this cost is superlinear in N. For an illustration, we extend the simulation study reported in the main paper, for the same Imb-Nul-Hi setting where, as for the estimator based on the first-order Laplace's approximation, the maximum sample size was set equal to 10^5 due to the high computational cost. While the precision of the estimator based on the pairwise likelihood is similar to that of the all-row-column estimator (Figures S26–S28), the superlinear cost is apparent (Figure S25), ruling out this solution for large-scale settings.

S4. Binary regression standard errors for the Stitch Fix data

Table S1 and Figure S29 compare the naive standard errors based on the classic probit output with two nonparametric methods that account for dependence in the probit model with crossed random effects fitted to the Stitch Fix Data with the all-row-column method. The nonparameteric methods are the two-way cluster-robust sandwich standard errors discussed in §2.4 of the manuscript and the Pigeonhole bootstrap standard errors of Owen (2007). Table S1 and Figure S29 show the close agreement between the sandwich and Pigeonhole bootstrap standard errors. On

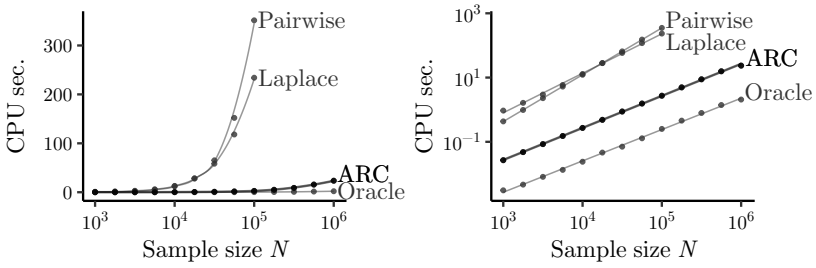

Fig. S25. Average computational costs for setting Imb-Nul-Hi .

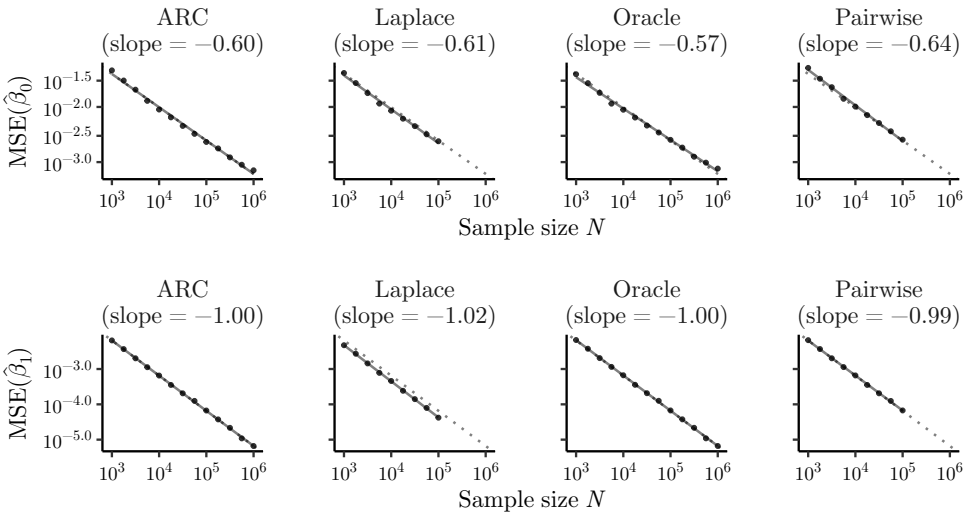

Fig. S26. Mean square errors of β_0 and β_1 estimates for setting Imb-Nul-Hi .

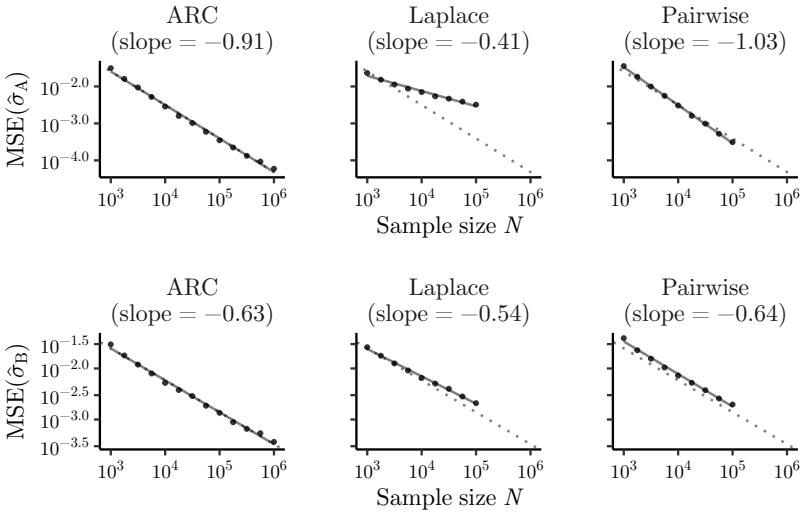

Fig. S27. Mean square errors of σ_A and σ_B estimates for setting Imb-Nul-Hi .

the other hand, the classic probit standard errors largely underestimate the uncertainty of the estimates.

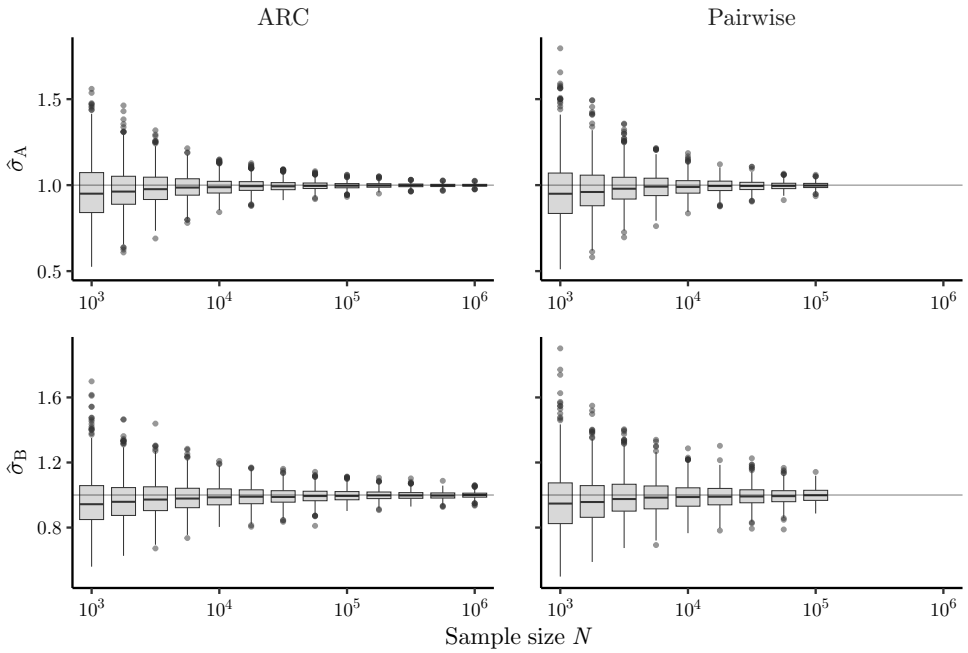

Fig. S28. Boxplots of σ_A and σ_B estimates in the setting Imb-Nul-Hi .

Table S1. *Binary regression standard errors for the Stitch Fix data, computed three ways.*

Variable	Level	Naive	Sandwich	Pigeonhole
Intercept		1.61	4.84	5.07
Client fit	loose or oversize	0.17	0.95	0.96
	straight or tight	0.18	0.57	0.63
Edgy	yes	0.14	0.47	0.50
Boho	yes	0.14	0.41	0.49
Chest		0.05	0.08	0.11
Size		0.03	0.09	0.10
Material	leather or animal	1.17	9.65	9.24
	regenerated	0.25	4.62	4.24
	vegetable	0.15	4.59	4.23
Item fit	loose or oversized	0.32	6.42	6.32
	straight or tight	0.26	3.74	3.45

Naive are the classical standard errors for the maximum likelihood estimates of the marginal probit model multiplied by $(1+\hat{\sigma}_A^2+\hat{\sigma}_B^2)^{1/2} \doteq 1.8$ to report them in the fixed effects scale of the probit model with crossed random effects. Sandwich are the two-way cluster-robust sandwich standard errors. Pigeonhole are non-parametric bootstrap standard errors described in Owen (2007). All standard errors are multiplied by 100 as in Figure S29.

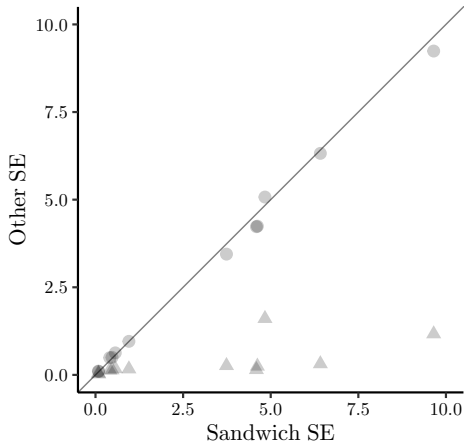

Fig. S29. Stitch Fix Application: two-way cluster-robust sandwich standard errors (x axis), naïve standard errors (y axis, triangles) and Pigeonhole nonparametric bootstrap standard errors (y axis, circles). The reference line is the first quadrant bisector $y = x$. All standard errors are multiplied by 100 as in Table S1.

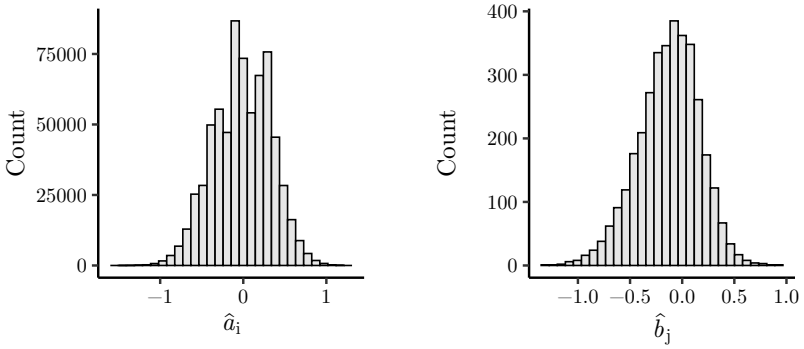

Fig. S30. Stitch Fix Application: distribution of the estimated customer ($\hat{a}_i$) and item ($\hat{b}_j$) random effects.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard university press.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Bellio, R. & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling* **5**, 217–227.

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

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer.

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

Lumley, T. & Mayer Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. Tech. rep., University of Washington.

Ochi, Y. E. & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531–543.

Owen, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics* **1**, 386–411.