# Complexity of Crossed Random Effects

Art B. Owen

with

Swarnadip Ghosh

and

Trevor Hastie

# Summary

Crossed random effects are common:

- Plant varieties $\times$ environments
- Customers $\times$ products
- Hospitals $\times$ dialysis centers

"Many to many" mappings

$N = 5{,}000{,}000$ clothing ratings from Stitch Fix

$N = 100{,}000{,}000$ movie ratings from Netflix

Simple models cost $O(N^{3/2})$ (or worse)

- Thesis and papers of Katelyn Gao

## Recent contributions with Hastie & Ghosh

Generalized least squares

- Iterations cost $O(N)$ via backfitting
- Needs $O(1)$ iterations

Logistic regression

- Iterations cost $O(N)$

# Categorical variables

- treated vs untreated patients
- 3 kinds of iris flower
- 50 US states

Some factors have many more levels

# Factors

- Product SKU

  e.g., dust filter for a Hoover Max Extract Pressure Pro model 60

  maybe millions of levels

- Query string

  “Pfizer”, “Zelensky”, · · · , “heteroscedasticity”

  very unequal (power law) frequency

- Customer ID, URL, IP address

  these ‘churn’

  appear, change popularity, disappear

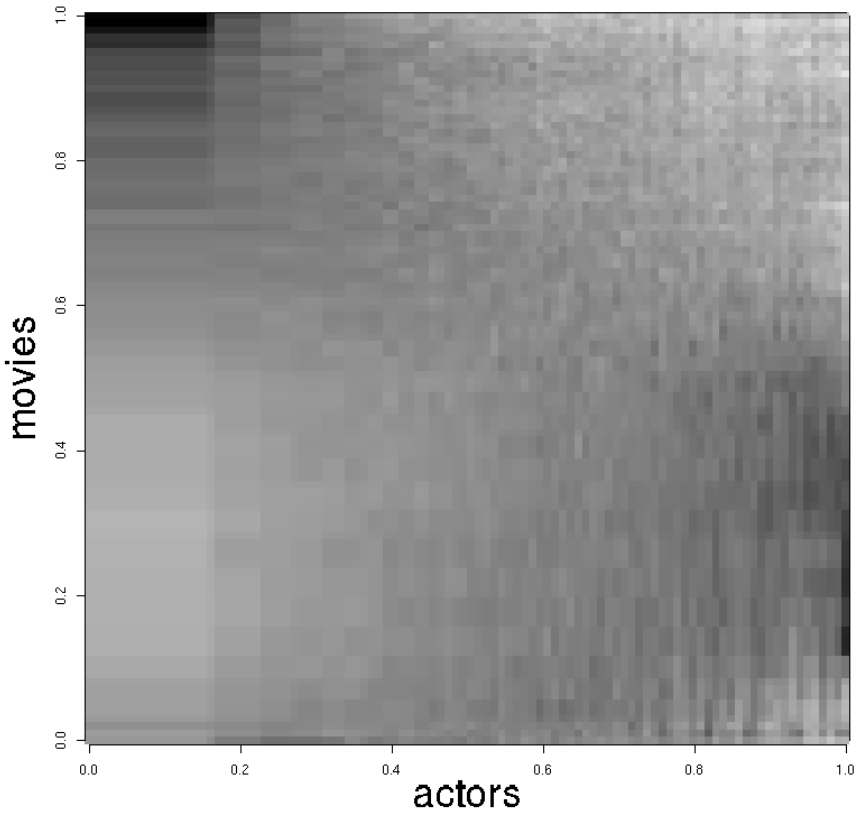
- Batch number (reagents, battery cells, rolls of vinyl)

  gone!

We **might** want to treat factors as random effects.

· · · because they introduce correlations.

# IMDB movies & actors




From work with Justin Dyer

Data courtesy of Jure Leskovic

Actor distribution has many 1s

# Simple model

$$Y_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}$$

$$a_i \sim (0, \sigma_A^2) \qquad b_j \sim (0, \sigma_B^2) \qquad \varepsilon_{ij} \sim (0, \sigma_E^2) \qquad \text{indep.}$$

## A humble model

Only one layer

No latent factors to discover communities / genres

$\sim$70 years old (before ICs)

## Already challenging

- Gaussian likelihood costs $O(N^{3/2})$ to evaluate **once**
  Gao & O (2019)
- Gibbs takes $O(N^{1/2})$ iterations at cost $O(N)$ each
  Gao & O (2017)

# Generalization

- Statistics runs on replication
- Easy from IID data
- Ok for hierarchical data
  - dependent with clusters
  - independent between

## Crossed effects

Hold out some customers

$\Longrightarrow$ correlated with held-ins (same products)

Almost an $n = 1$ setting

## Subjective generalization difficulty ladder

IID $\prec$ hierarchical $\prec$ time series $\prec$ crossed effects $\prec$ networks

# Notation

'Rows' $i = 1, \ldots, R$

'Columns' $j = 1, \ldots, C$

$Z_{ij} = 1 \iff (\boldsymbol{x}_{ij}, Y_{ij})$ observed $(0$ else$)$

## Sample sizes

$$N_{i\bullet} = \sum_{j=1}^{C} Z_{ij} \qquad N_{\bullet j} = \sum_{i=1}^{R} Z_{ij}$$

$$N = \sum_{i=1}^{R} N_{i\bullet} = \sum_{j=1}^{C} N_{\bullet j}$$

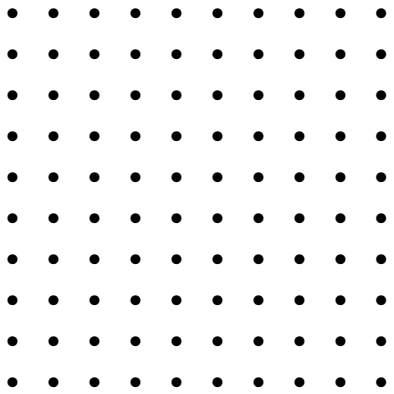
## Sparsity

$$1 \ll R, C \ll N \ll R \times C$$

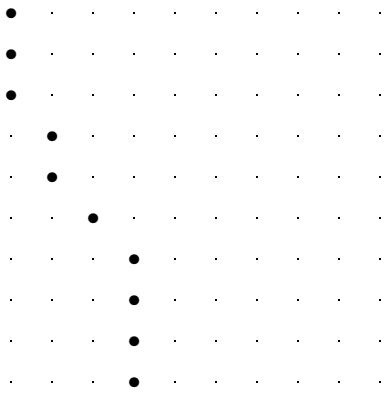
# Observation patterns

Solid for $Z_{ij} = 1$ dot / invisible for $Z_{ij} = 0$

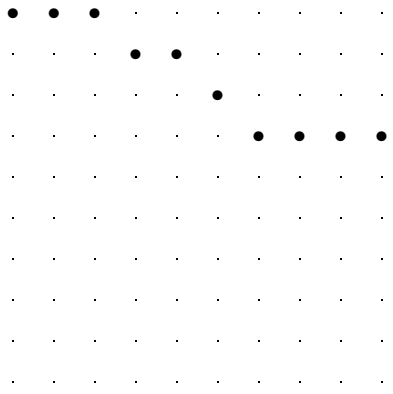
Balanced



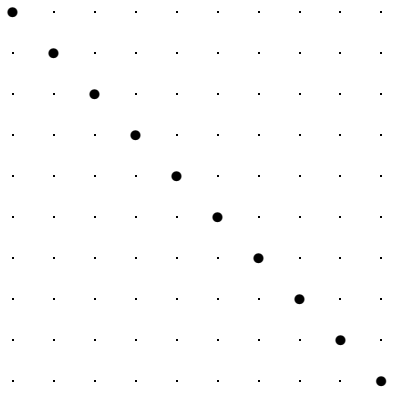
Row nested in col



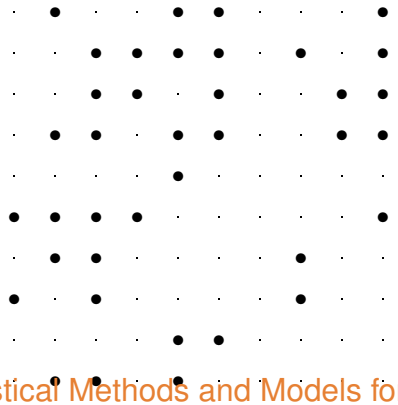
Col nested in row



IID



Arbitrary



# Informative missingness

Movie / TV ratings biased high

Restaurant ratings biased towards extremes

## Handling informative missingness

pass for now

needs info from outside the data

crossed effects hard enough already

# OLS and GLS

Ordinarily least squares and generalized least squares

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

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

## In compatible order

| | |
|---|---|
| $\mathcal{X} \in \mathbb{R}^{N\times p}$ | rows are $\boldsymbol{x}_{ij}$ |
| $\mathcal{Y} \in \mathbb{R}^{N}$ | elements are $Y_{ij}$ |
| $\mathcal{V} \in \mathbb{R}^{N\times N}$ | $\mathrm{Cov}(\mathcal{Y})$ |

## Nota Bene

$\boldsymbol{x}_{ij}, \beta \in \mathbb{R}^p$ $p$ not large and not growing with $N$

# Two problems with OLS

OLS is **inefficient**:

$$\mathrm{Var}(\hat{\beta}_{\mathrm{OLS}}) \succcurlyeq \mathrm{Var}(\hat{\beta}_{\mathrm{GLS}})$$

low power

OLS is **naive**:

$$\widehat{\mathrm{Var}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS}}) \preccurlyeq \mathrm{Var}(\hat{\beta}_{\mathrm{OLS}}) \qquad \text{(in expectation)}$$

false discoveries

At least

OLS is consistent

# Toy example of $\mathcal{V} = \mathrm{Cov}(\mathcal{Y})$

$R = 3$ rows and $C = 4$ columns with $N = 8$ observations

$$\begin{array}{c} \mathrm{Z} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 1 & \cdot & \cdot \\ \cdot & 1 & 1 & 1 \\ 1 & 1 & \cdot & 1 \end{bmatrix} \end{array}$$

Labels $\ell = 1, \ldots, N$

$$\begin{array}{c} Z \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 2 & \cdot & \cdot \\ \cdot & 3 & 4 & 5 \\ 6 & 7 & \cdot & 8 \end{bmatrix} \end{array}$$

E.g. observation $\ell = 7$ is in row $i = 3$ and column $j = 2$.

# Correlation structure

Correlations come from common rows or columns

Row correlations of $a_i$, $R = 3$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | · | · | · | · | · | · |
| 2 | 1 | 1 | · | · | · | · | · | · |
| 3 | · | · | 1 | 1 | 1 | · | · | · |
| 4 | · | · | 1 | 1 | 1 | · | · | · |
| 5 | · | · | 1 | 1 | 1 | · | · | · |
| 6 | · | · | · | · | · | 1 | 1 | 1 |
| 7 | · | · | · | · | · | 1 | 1 | 1 |
| 8 | · | · | · | · | · | 1 | 1 | 1 |

Column correlations of $b_j$, $C = 4$

| | 1 | 6 | 2 | 3 | 7 | 4 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | · | · | · | · | · | · |
| 6 | 1 | 1 | · | · | · | · | · | · |
| 2 | · | · | 1 | 1 | 1 | · | · | · |
| 3 | · | · | 1 | 1 | 1 | · | · | · |
| 7 | · | · | 1 | 1 | 1 | · | · | · |
| 4 | · | · | · | · | · | 1 | · | · |
| 5 | · | · | · | · | · | · | 1 | 1 |
| 8 | · | · | · | · | · | · | 1 | 1 |

# $\mathrm{Cov}(\mathcal{Y})$ in row order

$$\mathcal{V} = \sigma^2_A \begin{pmatrix} 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \end{pmatrix} + \sigma^2_B \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 \end{pmatrix} + \sigma^2_E I$$

We need $\mathcal{V}^{-1}\mathcal{X}$.

Sherman-Morrison-Woodbury does not do it.

# Linear mixed models

Naive algebra costs $O(N^3)$

Actual algebra costs $O((R+C)^3)$ Bates (2014)

$$RC \gg N \implies \max\{R, C\} \gg \sqrt{N} \implies (R+C)^3 \gg N^{3/2}$$

## Upshot

**Crossed:** superlinear cost.

**Nested:** blocks $\implies$ linear cost.

## LMM computation

The best is Doug Bates' most recent Julia code.

It costs $O(N^{3/2})$ to evaluate the likelihood **once**.

# What about MCMC?

- Plain Gibbs takes $O(\sqrt{N})$ iterations at $O(N)$ cost Gao & O (2017)
- quite unlike successes in the nested case, e.g.
  Yu & Meng (2011) interweaving, Gelman et al. STAN
- Problems with: block Gibbs, reparameterization, Langevin, MALA,
  Indep sampler, RWW, RWM subsampling, pCN
- Bates et al. (2015) took Bayes out of `lme4` $\cdots$ comput'n deemed unreliable

## Literature check

**Nested:** Lots of MCMC papers, theory and applied, hierarchical models.

**Crossed:** Very few MCMC papers.

# State of the art

1) Cameron, Gelbach & Miller (2011)

   OLS, non-naive via Huber-White

2) Gao & O method of moments, inefficient but not naive.

   Efficient if $\sigma^2_A \approx 0$ or $\sigma^2_B \approx 0$

3) Papaspiliopoulos, Roberts, & Zanella (2020). hereafter PRZ

   Collapsed Gibbs sampler. [Right way to do Bayes.]

Similar idea Johndrow (personal communication).

## PRZ (2020)

✓ Analytically integrate out intercept.

   Huge improvement.

× Assume all $N_{i\bullet} = N/R$ and all $N_{\bullet j} = N/C$.

Then mixing time is $O(1 \times \text{unknown})$

   Unknown mixing time of Gibbs walk on $Z_{ij}$:

   on later slide

# Backfitting

Robinson (1991) proves GLS $\equiv$

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

for observation matrices

$$\mathcal{Z}_A \in \{0,1\}^{N\times R} \qquad \mathcal{Z}_B \in \{0,1\}^{N\times C}$$

and 'ridge' penalties

$$\lambda_A = \frac{\sigma_E^2}{\sigma_A^2} \qquad \lambda_B = \frac{\sigma_E^2}{\sigma_B^2}$$

## Basic backfit

Update $\boldsymbol{a}$ then $\boldsymbol{b}$ then $\beta$ then $\boldsymbol{a}$ then $\boldsymbol{b}$ etc.

$=$ Block coordinate descent $=$ Gauss-Seidel

# For one random effect

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

So solve

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

Solve normal equations $\cdots$

$$\begin{aligned}
\mathcal{Z}_A \hat{\boldsymbol{a}} &= \left[\mathcal{Z}_A \left(\mathcal{Z}_A^{\mathsf{T}} \mathcal{Z}_A + \lambda_A I_R\right)^{-1} \mathcal{Z}_A^{\mathsf{T}}\right] (\mathcal{Y} - \mathcal{X}\hat{\beta}) \\
&\equiv \mathcal{S}_A (\mathcal{Y} - \mathcal{X}\hat{\beta}) \qquad \text{“smoother matrix” } \mathcal{S}_A \\
\hat{\beta} &= (\mathcal{X}^{\mathsf{T}} (I_N - \mathcal{S}_A) \mathcal{X})^{-1} \mathcal{X}^{\mathsf{T}} (I_N - \mathcal{S}_A) \mathcal{Y}
\end{aligned}$$

NB: $\mathcal{S}_A$ shrinks row averages towards zero

# Two effects

For generic response $\mathcal{R} \in \mathbb{R}^N$, we alternate

$$\mathcal{Z}_A \hat{\boldsymbol{a}} \leftarrow \mathcal{S}_A(\mathcal{R} - \mathcal{Z}_B \hat{\boldsymbol{b}})$$

$$\mathcal{Z}_B \hat{\boldsymbol{b}} \leftarrow \mathcal{S}_B(\mathcal{R} - \mathcal{Z}_A \hat{\boldsymbol{a}})$$

It converges

Buja, Hastie, Tibshirani (1990).

to a two-factor smoother $\mathcal{S}_{AB}$

that we use to update $\hat{\beta}$

Details in

Ghosh, Hastie & Owen Annals Stat (2022)

## Cost

$O(N)$ per iteration.

$\Longrightarrow$ We must bound the # iterations.

# Improved backfitting

Impose $\sum_i a_i = \sum_j b_j = 0$

At each iteration

Speeds up convergence

Avoids identifiability problem with intercept

# Stitch Fix

Stylists select clothing and send $5$ items to clients

Clients buy some and return others

## Variables

- Customer $i$
- Garment $j$
- Features $\boldsymbol{x}_{ij}$ price, size, materials, brand, ZIP code $\cdots$

## Model

$$Y_{ij} \sim \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}$$

## Response

$Y_{ij}$: size ok 0/1 or liked (10 point scale)

Enormous thanks to Brad Klingenberg for data.

# Stitch Fix data

$N = 5{,}000{,}000$ ratings by $R = 762{,}752$ clients on $C = 6{,}318$ items.

Tiny data subset; very old data

Ratings $Y_{ij}$ are on a 10 point scale.

## Predictors

Match$_{ij} \in [0, 1]$, a prediction from some baseline model

(not representative of all their algos).

Whether item is 'Edgy' or 'Boho'.

Same for client.

Material type: leather, fur, acrylic, $\cdots$, wool.

$p = 30$, including intercept.

# Analysis

- For one regression model
- Explore OLS naivete
- Explore OLS inefficiency

## Variance components

Gao & O (2019)

Moments $\rightarrow$ consistent $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$ and $\hat{\sigma}^2_E$ in $O(N)$ work

# Model for ratings

For each observed client-item pair $(i, j)$:

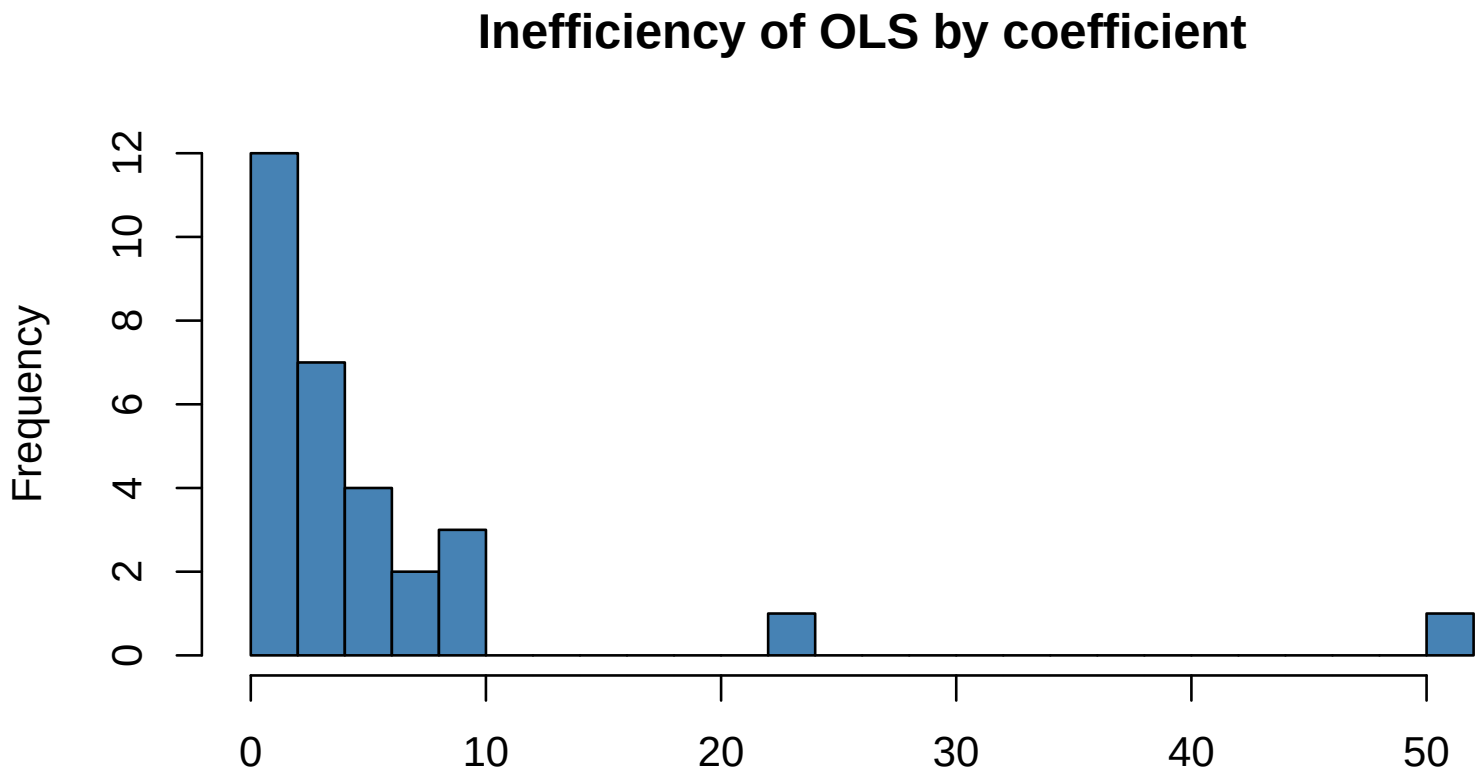
$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1 \text{Match}_{ij} + \beta_2 \mathbb{I}\{\text{client edgy}\}_i + \beta_3 \mathbb{I}\{\text{item edgy}\}_j \\ &+ \beta_4 \mathbb{I}\{\text{client edgy}\}_i \times \mathbb{I}\{\text{item edgy}\}_j + \beta_5 \mathbb{I}\{\text{client boho}\}_i \\ &+ \beta_6 \mathbb{I}\{\text{item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

## Notes

- Categorical $\text{Material}_{ij} \implies$ Indicator variables (baseline = Polyester)
- $p = 30$
- Gao & O found edgy items to edgy clients worked
  (but even boho clients tended not to like boho items)

# Inefficiency of OLS

$$1 < \frac{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS},j})} < 51 \qquad j = 0, \ldots, 29$$

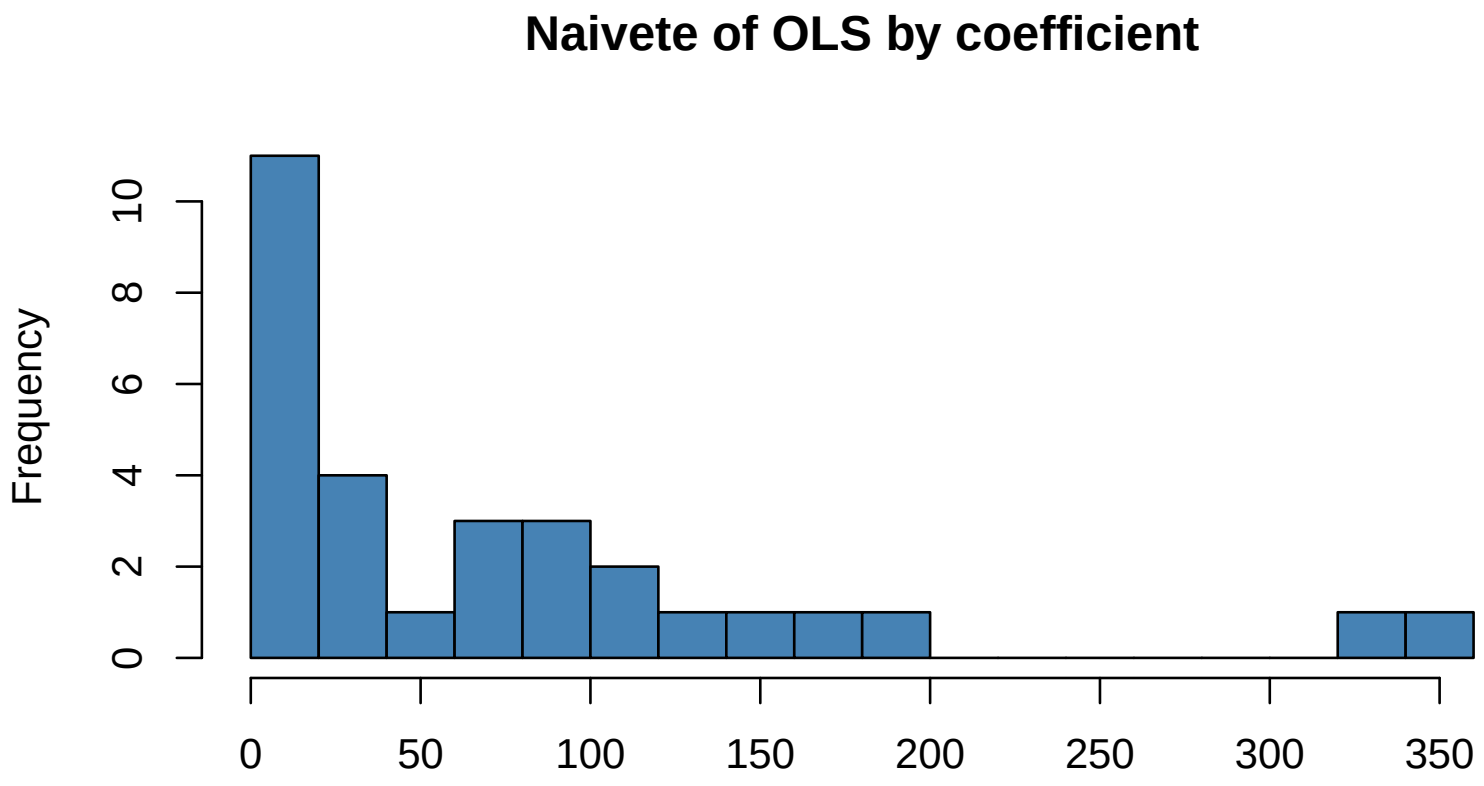



E.g., ratio $= 5$ is like ignoring $80\%$ of information

# Naivete of OLS

$$1.75 < \frac{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})} < 350 \qquad j = 1, \ldots, 30$$

Worst is material $=$ Modal, next is Tencel.






Naive by $100 \implies$ CIs $10$x too narrow $\implies$ Want $95\%$ get $15.5\%$

# Convergence

Backfitting update to $\boldsymbol{b}$:

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

If it converges, so does $\boldsymbol{a}$. Solution:

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

We want spectral radius$(M) \leqslant 1 - \delta$ for $\delta > 0$

## Sufficient condition

$$\|M\|_p \equiv \sup_{\boldsymbol{b} \neq 0} \frac{\|M\boldsymbol{b}\|_p}{\|\boldsymbol{b}\|_p} \leqslant 1 - \boldsymbol{\delta} \qquad \text{some } 1 \leqslant p \leqslant \infty \text{ and } \delta > 0$$

# Main results

Conditions on $Z_{ij}$ for which

$$\mathbb{P}\big(\|M\|_1 < 1 - \delta\big) \to 1$$

as sampling increases.

For that we need

Model with $R, C \to \infty$

$N \ll RC$

$Z_{ij}$ to control $\|M\|_1$

Why $\|\cdot\|_1$?

More tractable than $\|\cdot\|_2$ or spectral norm

# The model

Problem size is $S \to \infty$

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

$$N \ll RC = S^{\rho+\kappa}$$

## Sampling pattern

$$Z_{ij} \overset{\text{ind}}{\sim} \text{Bern}(p_{ij}) \qquad 1 \leqslant i \leqslant R \quad 1 \leqslant j \leqslant C$$

$$\frac{S}{RC} \leqslant p_{ij} \leqslant \Upsilon \frac{S}{RC} \qquad 1 \leqslant \Upsilon < \infty$$

The good

- Unequal random $N_{i\bullet}$ and $N_{\bullet j}$
- Unequal $\mathbb{E}(N_{i\bullet})$ and $\mathbb{E}(N_{\bullet j})$

The disappointing

- Still very close to equal
- Does not include tiny $N_{i\bullet}$ and $N_{\bullet j}$

$$S \leqslant \mathbb{E}(N) \leqslant \Upsilon S$$

# Main result

Theorem 4.3. If

| | |
|---|---|
| $0 < \rho, \kappa < 1$ | $1 \ll R, C \ll N$ |
| $\rho + \kappa > 1$ | $N \ll RC$ |
| $2\rho + \kappa < 2$ | controls $N_{i\bullet}$ |
| $3\rho + 4\kappa < 4$ | gets column overlaps & controls $N_{\bullet j}$ |

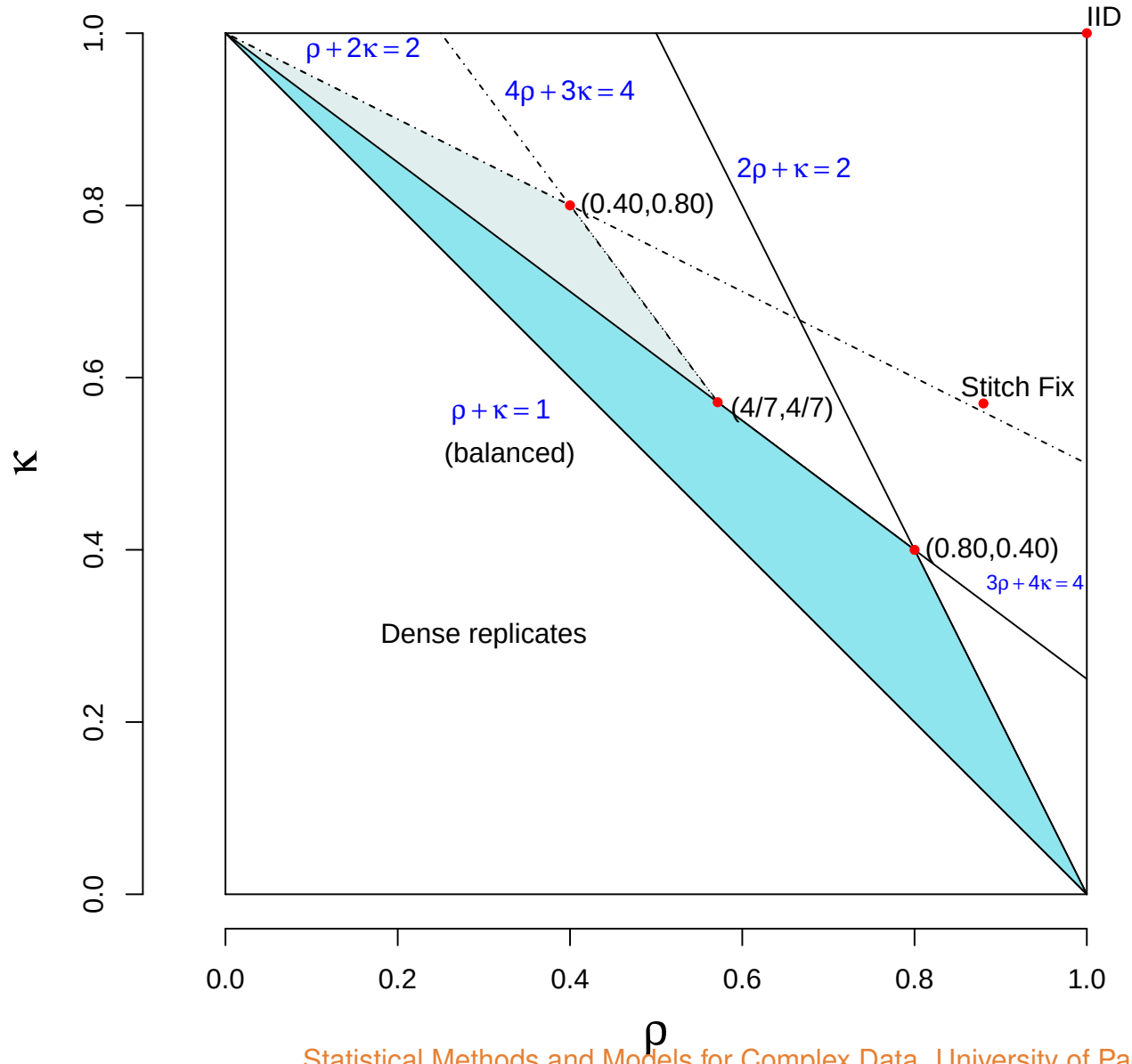
Then for any $\epsilon > 0$

$$\mathbb{P}\big(\|M\|_1 \leqslant \Upsilon^2 - \Upsilon^{-2} + \epsilon\big) \to 1$$

For $\Upsilon^2 - \Upsilon^{-2} < 1$

$$\Upsilon < \sqrt{\frac{1+\sqrt{5}}{2}} \doteq 1.27$$

# Domain


IID
ρ + 2κ = 2
4ρ + 3κ = 4
2ρ + κ = 2
(0.40,0.80)
Stitch Fix
(4/7,4/7)
ρ + κ = 1
(balanced)
(0.80,0.40)
3ρ + 4κ = 4
Dense replicates
κ
ρ
0.0
0.2
0.4
0.6
0.8
1.0

# The proof

Update: $b^{(k+1)} \leftarrow Mb^{(k)} + \eta \quad$ for $\quad M \in \mathbb{R}^{C \times C}$

$$\|M\|_1 = \max_{1 \leqslant s \leqslant C} \sum_{j=1}^{C} |M_{js}|$$

## Steps

1) Write $M_{js}$ in terms of $Z_{ij}$ and $\sigma_A^2/\sigma_E^2$ and $\sigma_B^2/\sigma_E^2$
2) Derive simultaneous almost sure bounds on $N_{i\bullet}$, $N_{\bullet j}$ etc. (Hoeffding)
3) Propagate the bounds to $\|M\|_1$

# Actual Stitch Fix norms

$$Z \in \{0,1\}^{762{,}752 \times 6318} \qquad M \in \mathbb{R}^{6318 \times 6318}$$

Three backfitting iterations:

$M^{(0)}, M^{(1)}\ M^{(2)}$

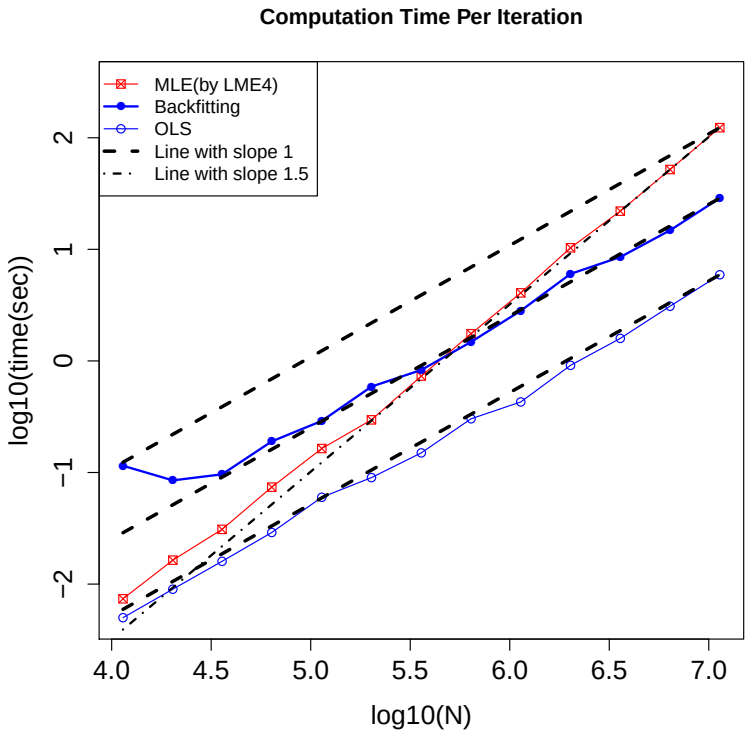
$\|M\|_1 < 1 - \delta$ sufficient but not necessary

$$\begin{pmatrix} \|M^{(0)}\|_1 & \|M^{(0)}\|_2 & |\lambda_{\max}(M^{(0)})| \\ \|M^{(1)}\|_1 & \|M^{(1)}\|_2 & |\lambda_{\max}(M^{(1)})| \\ \|M^{(2)}\|_1 & \|M^{(2)}\|_2 & |\lambda_{\max}(M^{(2)})| \end{pmatrix} = \begin{pmatrix} 31.9525 & 1.4051 & 0.6403 \\ 11.2191 & 0.4512 & 0.3338 \\ 8.9178 & 0.4541 & 0.3341 \end{pmatrix}$$

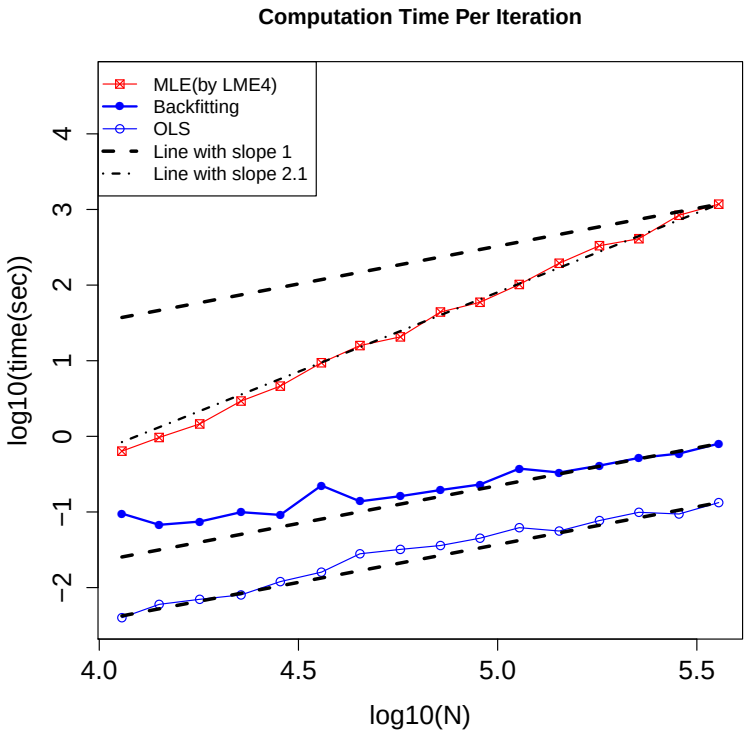
Iterations

6 iterations with threshold $10^{-8}$

# Timings




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




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

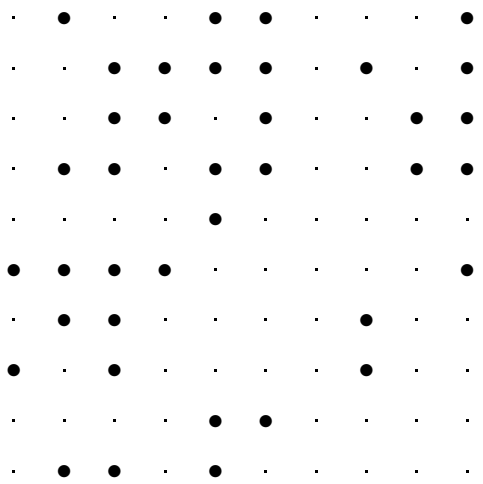
$$R^3, C^3 = O(N^{2.1})$$

# PRZ results

Mixing time for collapsed Gibbs

$$\rho_{\mathrm{PRZ}} = \frac{N\sigma_A^2}{N\sigma_A^2 + R\sigma_E^2} \times \frac{N\sigma_B^2}{N\sigma_B^2 + C\sigma_E^2} \times \rho_{\mathrm{AUX}}$$
$$\to \rho_{\mathrm{AUX}}$$

Auxilliary $\rho$



Mixing time for Gibbs sampler on $Z$:

new $i \mid j$ with prob $Z_{ij}/N_{\bullet j}$

new $j \mid i$ with prob $Z_{ij}/N_{i\bullet}$

Burnside process

They did not bound $\rho_{\mathrm{AUX}}$

# Improved mixing

Ghosh & Zhong (arxiv2109.02849)

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$$
$$a_i \sim \mathcal{N}(0, \sigma_A^2), \quad b_j \sim \mathcal{N}(0, \sigma_B^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2).$$

Using random matrix theory of Latala (2005) to study $\|M\|_2$

## Results

Mixing time $=$ Spectral Gap$^{-1}$ asymptotically bounded

Only need $\rho + \kappa/2 < 1$ and $\kappa + \rho/2 < 1$

Bigger triangle. Bigger $\Upsilon \approx 1.52$

If $\mathbb{E}(N_{i\bullet})$ and $\mathbb{E}(N_{\bullet j})$ constant then $\Upsilon$ can be greatly relaxed

# Logistic regression

Challenging integral

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

$$L(\beta \mid \boldsymbol{a}, \boldsymbol{b}) = \prod_{(i,j):Z_{ij}=1} \frac{\exp(\boldsymbol{x}_{ij}^\mathsf{T}\beta + a_i + b_j)^{Y_{ij}}}{1 + \exp(\boldsymbol{x}_{ij}^\mathsf{T}\beta + a_i + b_j)}$$

Laplace approximations and quasi-likelihood used

Hard questions remain

Consistency of the MLE is recent!

Jiang (2013)

What about Laplace?

# Quasi-likelihood

We use backfitting GLS as an inner loop

Outer loop is quasi-likelihood of Schall (1991)

Schall's algorithm costs $O(N^{3/2})$

- Takes trace of inverse of $(R+C)\times(R+C)$ matrix
- we ignore off diagonal elements to get $O(N)$
- and prove asymptotic equivalence

Iterations cost $O(N)$

Empirically few iterations; no proof

## 'Clubbed' updates

Update $\boldsymbol{a}$ & $\beta$ then $\boldsymbol{b}$ & $\beta$ then $\boldsymbol{a}$ & $\beta\cdots$

Brings large efficiency gain for categorical predictors

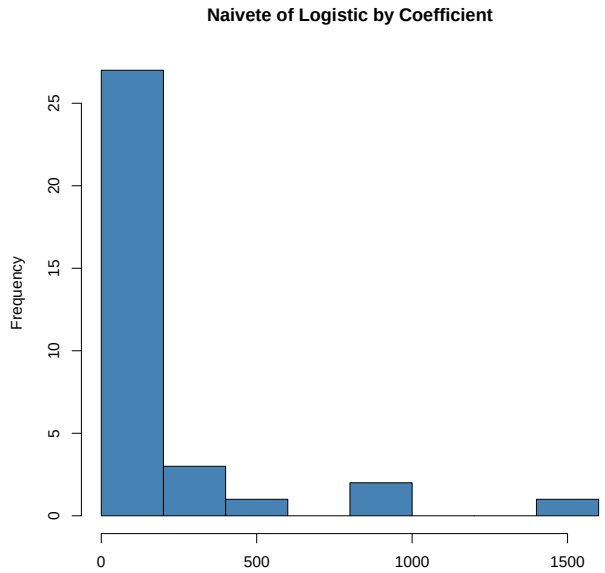
Critical in the GLMM setting

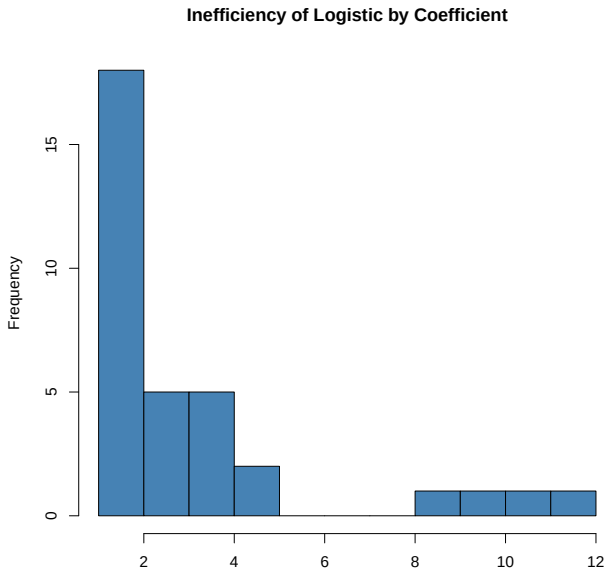
Ghosh, Hastie & O (2022) Electronic Journal Statistics

# Naivete and inefficiency

Some Stitch Fix data

Ghosh, Hastie & O arXiv:2105.13747




Severe naivete in logistic regression

# Probit

Work in progress with

C. Varin, R. Bello, S. Ghosh

Change: $\frac{e^w}{1+e^w} \rightarrow \Phi(w)$

Gaussian latent variables & Gaussian random effects

$\rightarrow$ some simplifications and efficiencies

# Thanks


- Co-authors Swarnadip Ghosh and Trevor Hastie
- Brad Klingenberg (Stitch Fix) data and discussions
- NSF IIS-1837931
- Invitation: Giovanna Capizzi, Annamaria Guolo
- Discussion: Alessandra Salvan, Nicola Sartori
- Hospitality: LCC Congressi

# Backup slides

1) is OLS a reasonable comparison?

2) why Bayes and frequentist rates are often the same

# OLS, really?

It seems like a straw man.

More likely to use $a_i$ and $b_j$ as fixed effects.

## Cost

$$O(N(R+C+p)^2) = O(N^2) \qquad \text{or worse}$$

So OLS on $p$ variables is feasible.

OLS treatment of effects as fixed is not.

## Hunch

people use learning algos that don't distinguish fixed vs random

# Bayes & frequentist iterations

Sampling $\mathcal{N}(0, \Sigma)$:

e.g. draw one $x_j$ at a time

convergence rate $\rho_\Sigma$

Minimizing $\boldsymbol{x}^\mathsf{T} Q \boldsymbol{x}$:

e.g. minimize over one $x_j$ at a time

convergence rate $\rho_Q$

Very generally $\Sigma = Q \implies \rho_S = \rho_Q$

Colin Fox also Amit & Grenander

## Conjectures

- this duality is why $N^{3/2}$ keeps popping up
- progress on Bayes $\implies$ progress on minimization
- and vice versa

# Ratio

## Inefficiency

$$\frac{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS},j})}$$

## Naivete

$$\frac{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})}$$

## Inefficiency / naivete

$$\frac{\widehat{\mathrm{Var}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{GLS},j})}$$

is the ratio of squared confidence interval widths