Backfitting and crossed random effects

Art B. Owen

and

Trevor Hastie

and

Swarnadip Ghosh

https://arxiv.org/abs/2007.10612

Summary

Crossed random effects are common:

- Plant varieties $\times$ environments
- Customers $\times$ products
- Hospitals $\times$ dialysis centers
  - He, Kalbfleisch, Li, Li (2013) on readmissions

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

- Thesis and papers of Katelyn Gao

Our contribution

We can get $O(N)$ cost via backfitting

- Buja, Hastie, Tibshirani (1989)

Linear models today

GLMMs "real soon now"

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

- major thanks to Brad Klingenberg

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

- Query string:

  “covid”, “Biden”, · · · , “heteroscedasticity”

- Customer ID, URL, IP address

  these ‘churn’

Some differences

1) can have millions of levels
2) power law frequency
3) continually changing set of levels

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

Crossed random effects

Customers $\times$ products

Students $\times$ questions

Genotype $\times$ environments

Diners $\times$ restaurants

Review $\times$ book $\times$ mailing address $\times$ credit card #

I.E., two or more of these variables

More generally: many $\leftrightarrow$ many relationships

Simple model

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

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

Modest because

Only one layer

No latent factors to discover communities / genres

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 rows

$\Longrightarrow$ dependence comes in via columns

Borders on an $n = 1$ setting

Subjective difficulty ladder

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

Balanced and unbalanced

Simplest if all ij pairs observed.

Our use cases are unbalanced.

Notation

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

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

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

More indices

$$1 \leqslant r \leqslant R \qquad 1 \leqslant s \leqslant C$$

Sample sizes

$$N_{i\bullet} = \sum_{j=1}^{C} Z_{ij} \quad \text{'size' of row } i$$

$$N_{\bullet j} = \sum_{i=1}^{R} Z_{ij} \quad \text{'size' of col } j$$

$$N = \sum_{i=1}^{R} N_{i\bullet} = \sum_{j=1}^{C} N_{\bullet j} \quad \text{sample size}$$

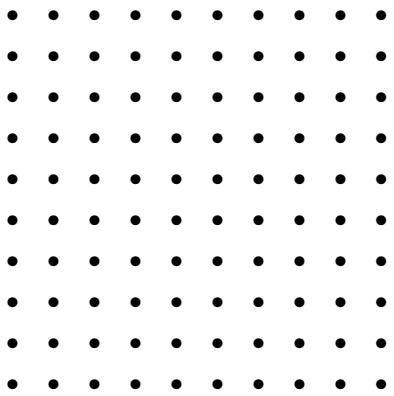
Sparsity

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

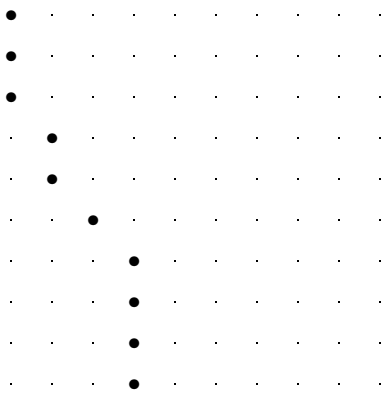
Observation patterns

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

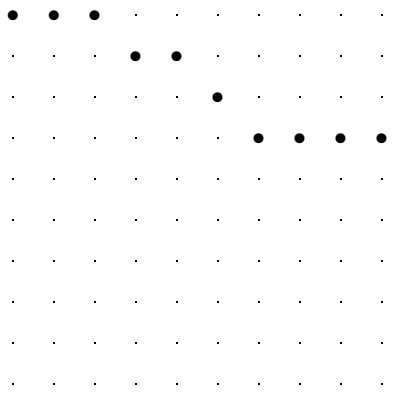
Balanced



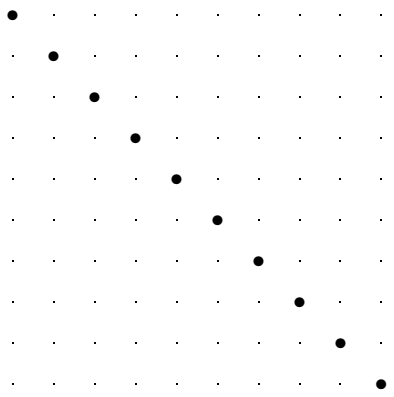
Row nested in col



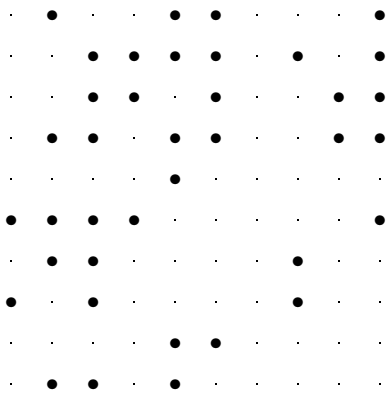
Col nested in row



IID



Arbitrary



Informative missingness

Movie / TV ratings biased high

Restaurant ratings biased towards extremes

Handling informative missingness

- Requires information from outside the data
- Pass for now
- Linear mixed models are hard enough already
- Maybe propensity methods will help

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

$\boldsymbol{x}_{ij}, \beta \in \mathbb{R}^p$ $\quad p$ 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**:

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

false discoveries

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

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

$$
\begin{array}{c} Z \\ 1 \\ 2 \\ 3 \end{array}
\begin{array}{c}
\begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\
\left[\begin{array}{cccc}
1 & 1 & \cdot & \cdot \\
\cdot & 1 & 1 & 1 \\
1 & 1 & \cdot & 1
\end{array}\right]
\end{array}
$$

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

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

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

$$\mathcal{V} = \sigma_A^2 \begin{pmatrix} 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 \end{pmatrix} + \sigma_B^2 \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 \end{pmatrix} + \sigma_E^2 I$$

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

Sherman-Morrison-Woodbury does not do it.

Sad trombone sound.

Linear mixed models

Naive algebra costs $O(N^3)$

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

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

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 $\sigma_A^2 \approx 0$ or $\sigma_B^2 \approx 0$

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 $\boldsymbol{a}$ etc.

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

For one random effect

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

So solve

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

Normal equations

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

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

We get

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

NB: $\mathcal{S}_A$ does shrunken row averages

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).

Limit equals two-factor smoother matrix that we can apply to $\mathcal{Y}$ and columns of $\mathcal{X}$.

Details in

Ghosh, Hastie & Owen arXiv:2007.10612

Cost

$O(N)$ per iteration.

$\Longrightarrow$ We must bound the # iterations.

Centering

Usually $\mathcal{X}$ has an intercept

Then the solution has

$$\sum_{i=1}^{R} \hat{a}_i = 0 = \sum_{j=1}^{C} \hat{b}_j. \qquad (*)$$

Why make the iterations discover $(*)$?

let's just bake it in

Analagous to collapsed sampling

Simple centering

After each iteration:

$$a_i \leftarrow a_i - \bar{a}, \qquad \text{for} \quad \bar{a} = \frac{1}{R}\sum_{i=1}^{R} a_i$$

$$b_j \leftarrow b_j - \bar{b}$$

Principled centering

$$\min_{\boldsymbol{a}} \|\mathcal{R} - \mathcal{Z}_A \boldsymbol{a}\|^2 + \lambda_A \|\boldsymbol{a}\|^2 \qquad \text{subject to } \sum_{i=1}^{R} a_i = 0$$

Theorem 3.2

$$\mathcal{R}_{i\bullet} = \sum_{j=1}^{C} Z_{ij} \mathcal{R}_{ij} \qquad \text{row sums}$$

$$w_i = \frac{(N_{i\bullet} + \lambda_A)^{-1}}{\sum_r (N_{r\bullet} + \lambda_A)^{-1}} \qquad \text{normalized weights}$$

Then

$$\hat{a}_i \leftarrow \frac{\mathcal{R}_{i\bullet} - \sum_r w_r \mathcal{R}_{r\bullet}}{N_{i\bullet} + \lambda_A}$$

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.

This a subset of their customer / inventory base.

we got no identifying features

Data is $\geqslant 5$ years old.

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

- We only look at 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

We get λ_A and λ_B

Thesis Gao (2017)

$\hat{\sigma}^2_A$ etc asymptotically normal

Model for ratings

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

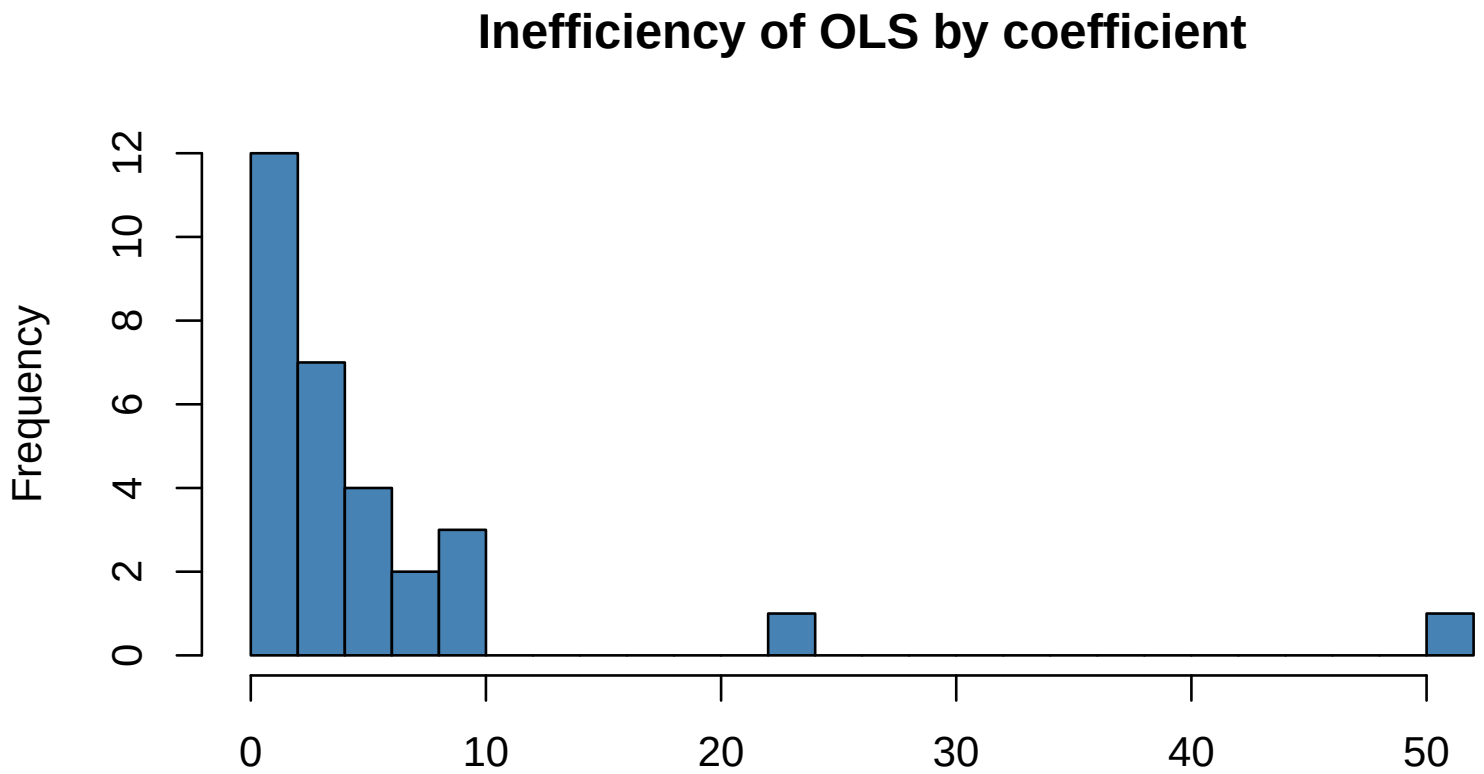
$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1 \text{Match}_{ij} + \beta_2 \mathbb{I}\{\text{client edgy}\}_i + \beta_3 \mathbb{I}\{\text{item edgy}\}_j \\ &+ \beta_4 \mathbb{I}\{\text{client edgy}\}_i \times \mathbb{I}\{\text{item edgy}\}_j + \beta_5 \mathbb{I}\{\text{client boho}\}_i \\ &+ \beta_6 \mathbb{I}\{\text{item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

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

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

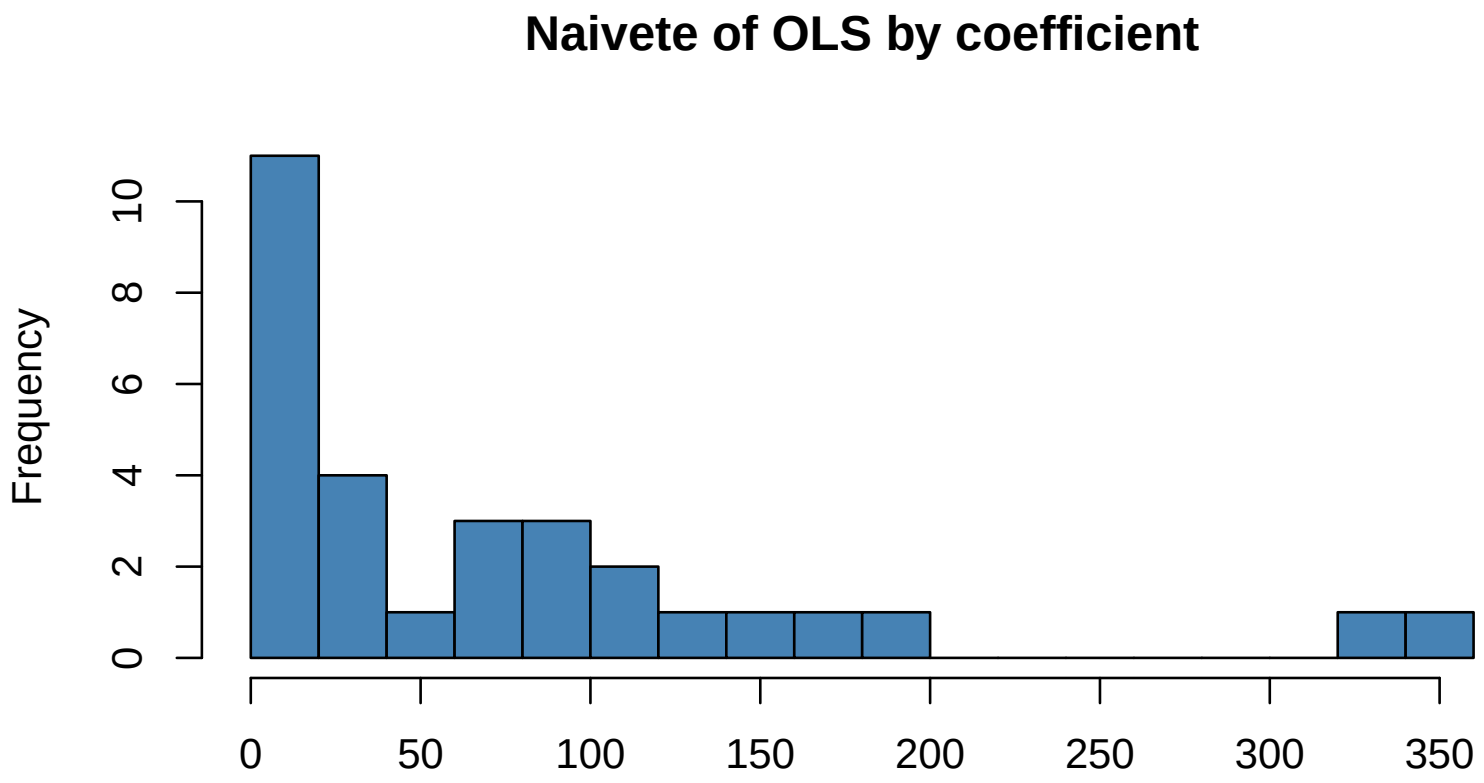


E.g., ratio $= 5$ is like ignoring 80% of information

Naivete of OLS

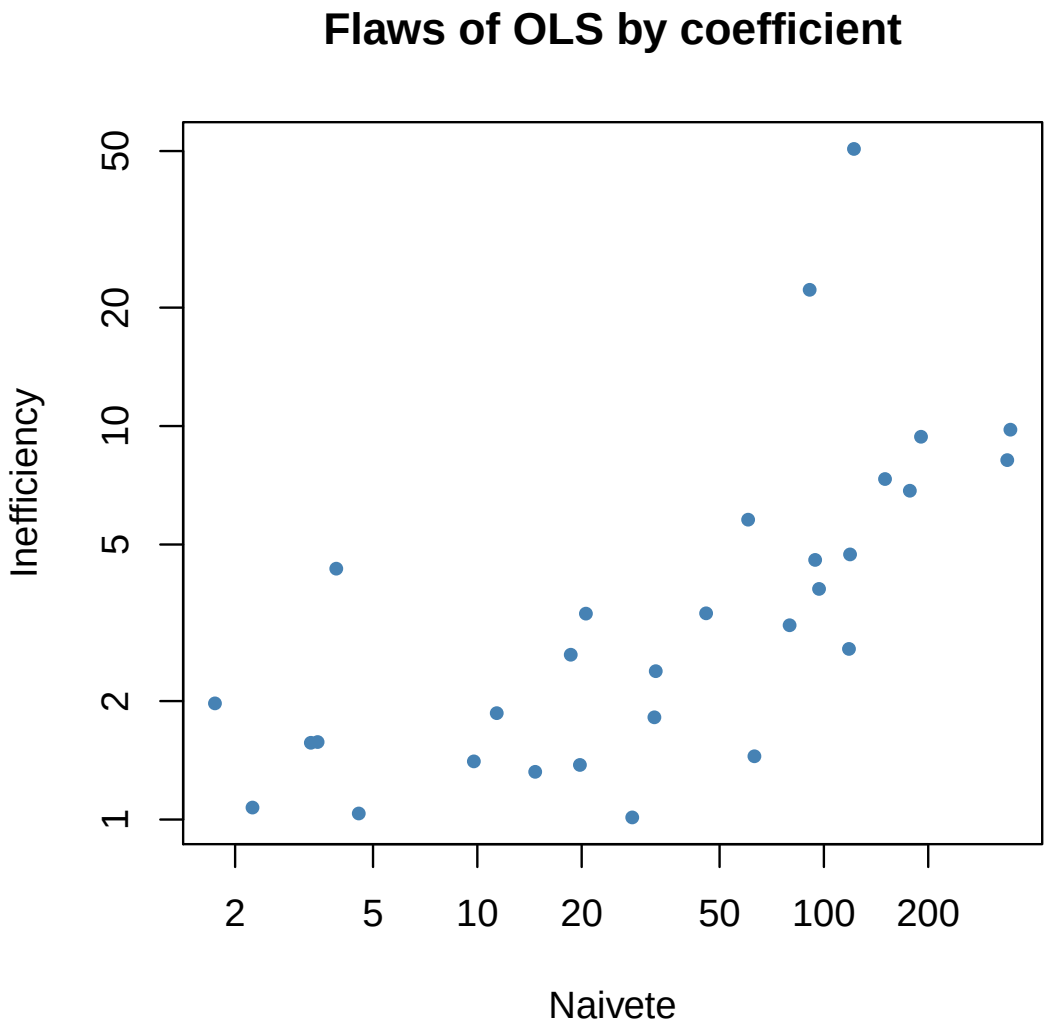
$$1.75 < \frac{\widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{OLS},j})}{\widehat{\mathrm{Var}}_{\mathrm{OLS}}(\hat{\beta}_{\mathrm{OLS},j})} < 350 \qquad j = 1, \ldots, 30$$

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



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

Inefficiency vs naivete



Convergence

Backfitting update to $\boldsymbol{b}$:

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

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

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

For finite iterations

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

Keeps spectral radius below $1 - \delta$.

Matrix M

$$a_i \leftarrow \frac{\sum_s Z_{is}(Y_{is} - b_s)}{N_{i\bullet} + \lambda_A} \qquad b_j \leftarrow \frac{\sum_r Z_{rj}(Y_{rj} - a_r)}{N_{\bullet j} + \lambda_B}$$

Leads to

For $1 \leqslant j, s \leqslant C$

$$b_j \leftarrow \eta_j + \sum_{s=1}^{C} M_{js} b_s$$

$$M_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_{r=1}^{R} \frac{Z_{rs}(Z_{rj} - N_{\bullet j}/R)}{N_{r\bullet} + \lambda_A}$$

Centering

Above is $M^{(0)}$.

Simple centering $M^{(1)}$.

Principled centering $M^{(2)}$.

Main results

Conditions on Z_{ij} for which

$$\mathbb{P}\big(\|M^{(2)}\|_1 < 1 - \delta\big) \to 1$$

as sampling increases. Similar for $M^{(1)}$.

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 $N_{\bullet j}$

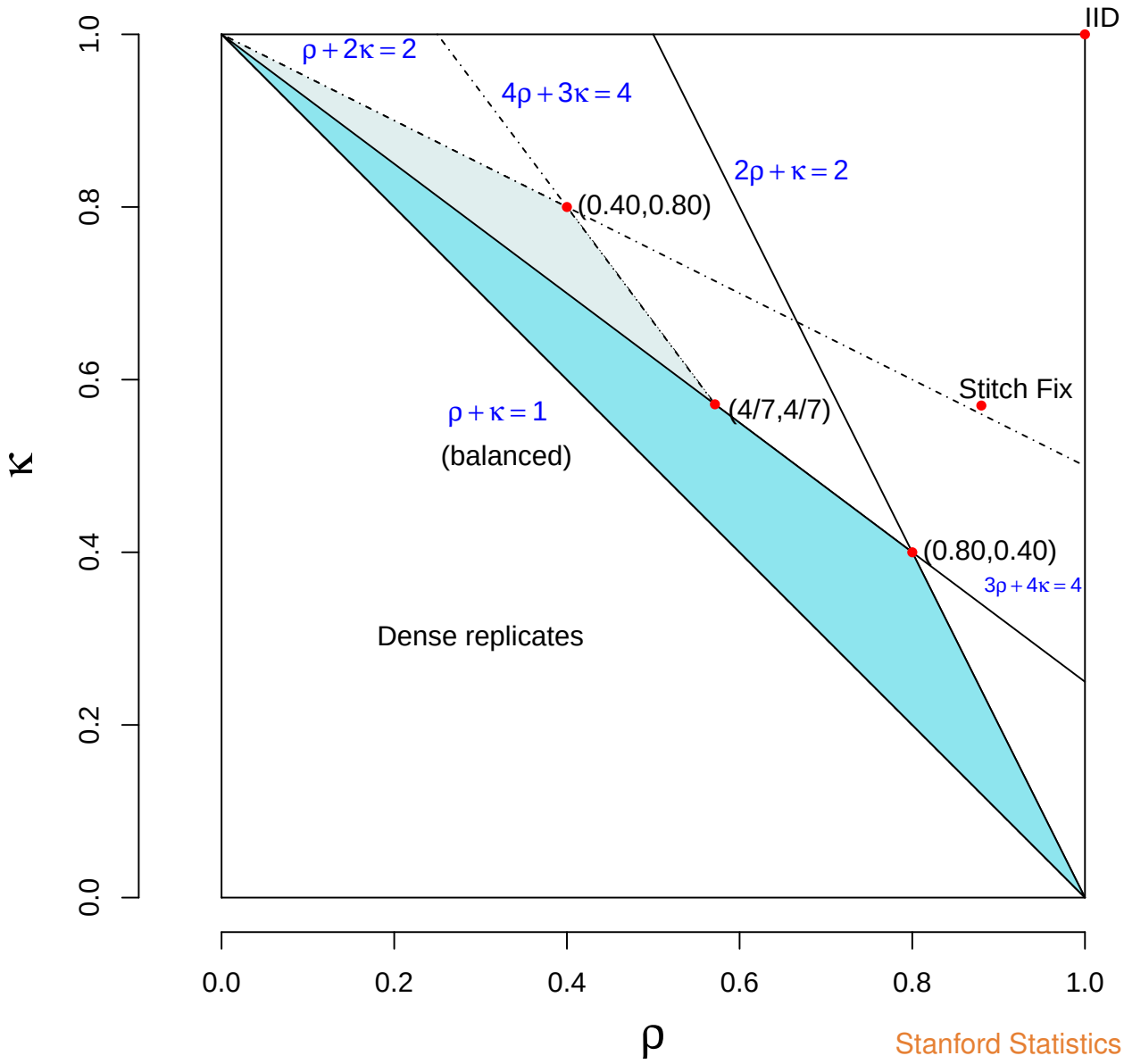
Then for any $\epsilon > 0$

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

Best Υ

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

Asymptotic domain



The proof

For $M \in \mathbb{R}^{C\times C}$

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

For $j \neq s$

$$M_{js} = \frac{1}{N_{\bullet j} + \lambda_B} \sum_{r=1}^{R} \frac{Z_{rj}}{N_{r\bullet} + \lambda_A} (Z_{rs} - \bar{Z}_{\bullet s}) \quad \text{where}$$

$$\bar{Z}_{\bullet s} = \sum_{i=1}^{R} \frac{Z_{is}}{N_{i\bullet} + \lambda_A} \Big/ \sum_{i=1}^{R} \frac{1}{N_{i\bullet} + \lambda_A}$$

Upshot

It's a bit of a mess

Sketch

Hoeffding inequalities keep

$$(1-\epsilon)S^{1-\rho} \leqslant N_{i\bullet} \leqslant (\Upsilon+\epsilon)S^{1-\rho}$$
$$(1-\epsilon)S^{1-\kappa} \leqslant N_{\bullet j} \leqslant (\Upsilon+\epsilon)S^{1-\kappa}$$

and co-observations for $j \neq s$

$$(1-\epsilon)S^{2-\rho-2\kappa} \leqslant \sum_{i=1}^{R} Z_{ij}Z_{is} \leqslant (\Upsilon^2+\epsilon)S^{2-\rho-2\kappa}$$

and then

much interval arithmetic to propagate bounds to $\|M\|_1$

$[a,A]+[b,B] \subseteq [a+b, A+B]$

For $a,b>0$ $\quad [a,A]/[b,B] \subseteq [a/B, A/b]$

$|[a,A]| \subseteq [0, \max(|a|,|A|)]$

Actual Stitch Fix norms

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

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

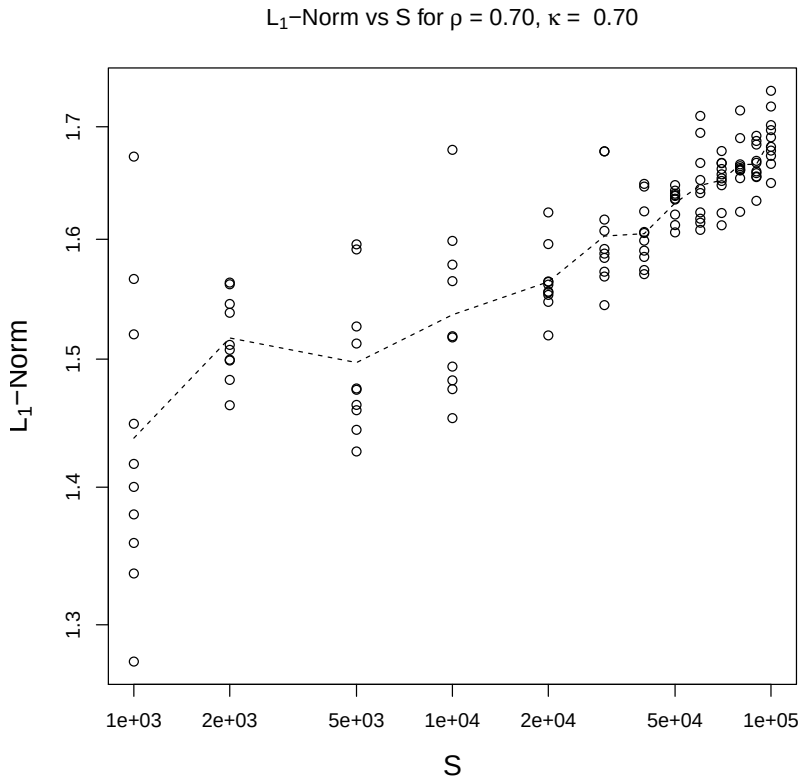
$$\begin{pmatrix} \|M^{(0)}\|_1 & \|M^{(0)}\|_2 & |\lambda_{\max}(M^{(0)})| \\ \|M^{(1)}\|_1 & \|M^{(1)}\|_2 & |\lambda_{\max}(M^{(1)})| \\ \|M^{(2)}\|_1 & \|M^{(2)}\|_2 & |\lambda_{\max}(M^{(2)})| \end{pmatrix} = \begin{pmatrix} 31.9525 & 1.4051 & 0.6403 \\ 11.2191 & 0.4512 & 0.3338 \\ 8.9178 & 0.4541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

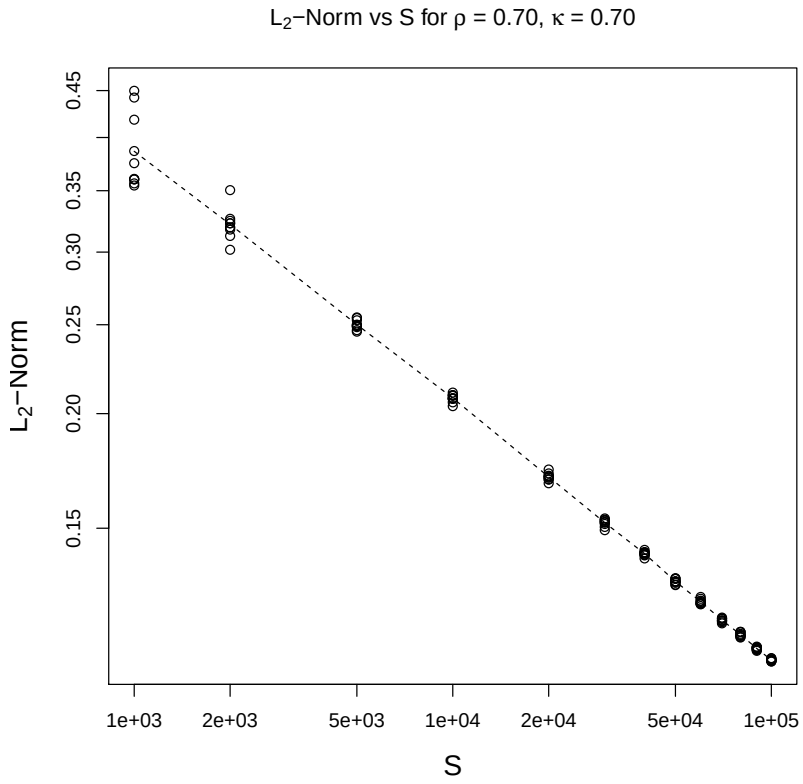
Iterations

6 iterations with threshold 10^{-8}

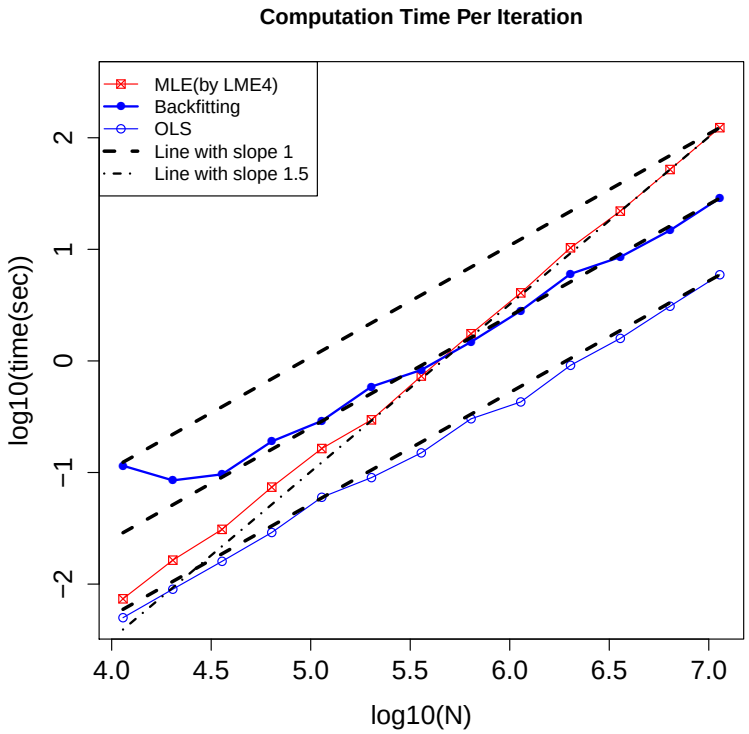
Simulated norms

$\rho = \kappa = 0.70$ outside our domain

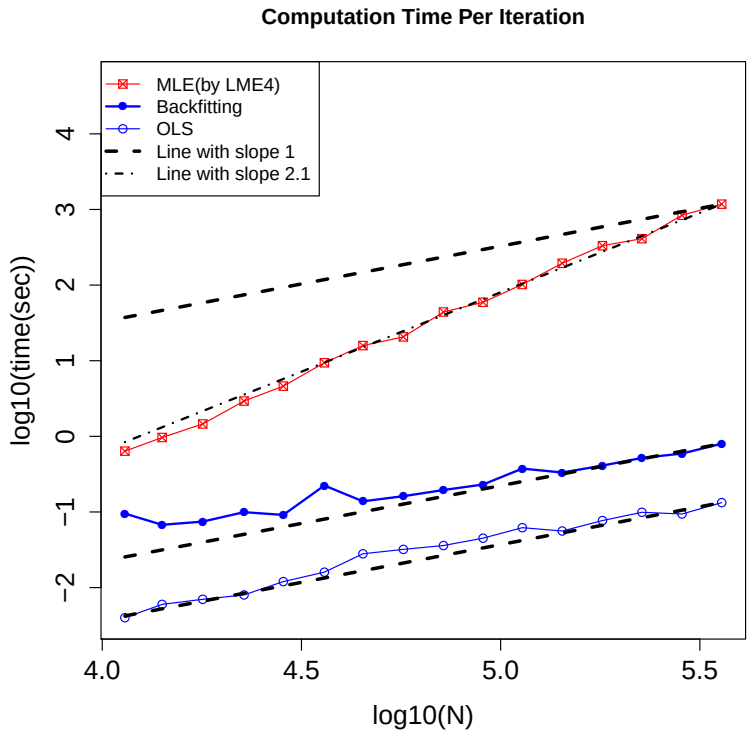




Timings



(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

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

Auxilliary ρ

Mixing time for Gibbs sampler on Z:

choose i given j with probability $Z_{ij}/N_{\bullet j}$

choose j given i with probability $Z_{ij}/N_{i\bullet}$

Further steps

- logistic regression (soon)
- higher way tables
- SVD-like interactions
- heteroscedastic effects

Thanks

- Co-authors Swarnadip Ghosh and Trevor Hastie
- Brad Klingenberg (Stitch Fix) data and discussions
- NSF IIS-1837931
- Giacomo Zanella, discussions
- Paromita Dubey, host
- Cindy Kirby, logistics

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 ρ_Σ

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

e.g. minimize over one x_j at a time

convergence rate ρ_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