Method of moments for large crossed linear mixed models

Art B. Owen
Stanford University
and
Katelyn Gao
Stanford University

Motif

Statistics faces two challenges

1) Data sets are getting bigger, and

2) findings are not reproducing.

Ideally

Problem 1 would fix problem 2.

Unfortunately

This does not happen when the data contain a huge tangle of correlations as in crossed random effects.

Summary

- e-commerce generates large crossed random effects data
- GLMMs are appropriate, but the MLE has superlinear cost, $N^{3/2}$
- MCMC appears to cost superlinearly too (e.g., $N^{3/2}$ for Gibbs)

Likelihood and Bayes both have high costs.

Let's go back to the '90s*

The method of moments helps when data sets grows faster than computation.

*1890s

Method of moments

1) Costs $O(N)$
2) No parametric distribution assumed
3) No tuning parameters
4) No convergence diagnostics
5) Easy to do in parallel

Drawbacks

1) Can give $\hat{\sigma}^2 < 0$
2) May require extended precision
3) May be inefficient (statistically)

E-commerce data

Logs look like

$$(x, y,\ \ i, j, \cdots, r, s)$$

Variables x, y

Y : Click Y/N $ spent Rating 1:5 stars Liked Y/N $\cdots$

X : experimental A/B time of day page load speed home city $\cdots$

Factors i, j, $\cdots$, r,s

- customer ID or cookie or IP address
- URL
- product ID (e.g., SKU)
- query string
- tweet or product review or news article

Other data

While e-commerce leads in data size, other areas have growing data set sizes.

Crossed random effects are as fundamental as

$$\text{Many} \to \text{Many}$$

mappings.

Agricultural yield

Plants: Cultivars $\times$ environments.

Animals: $\male \times \female$.

Education examples

Almost but not quite Many $\to$ One

From Raudenbush (1993)

Schools $\times$ neighborhoods

Students $\times$ teachers (multiyear data)

Factor vs categorical variable

Typical categorical variable: 3 kinds of iris flower, 50 US states

Big categorical variable

- millions of levels
- power law frequency (e.g., queries "Adele" $\cdots$ "heteroscedasticity")
- can be new levels every day
- many hapax legomena (appear once only)

Fixed vs. random effects

Some factors turn over (churn) much faster than others. E.g. cookies

It is better to learn something about the distribution from which levels are sampled, than to memorize facts about specific factor levels.

Stitch Fix

A stylist selects clothing and send 5 items to clients

Clients buy some and return others

The data are:

(client ID, garment ID, features X, rating Y)

Rating Y from 1 to 10.

Binary Y (bought vs returned) is also of interest.

The features can be about the garment or the client or 'joint'.

Enormous thanks to Brad Klingenberg for data.

Linear model

Client i and item j:

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}, \qquad i, j \in \mathbb{N}$$

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

$$\mathbb{E}(Y_{ij}^4) < \infty \qquad x_{ij} \in \mathbb{R}^p$$

Pattern of observed vs missing

$$Z_{ij} = \begin{cases} 1, & Y_{ij} \text{ observed} \\ 0, & \text{else} \end{cases}$$

Previous paper: Gao & O (2017)

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$$

Estimated $\sigma_A^2, \sigma_B^2, \sigma_E^2$

Informative missingness

Netflix: movie ratings are probably biased high.

Yelp: restaurant ratings are probably biased towards extremes.

Handling informative missingness

- Requires information from outside the data at hand,
- and / or untestable assumptions.
- Every case is different.

Propensity

Imbens & Rubin (2015)

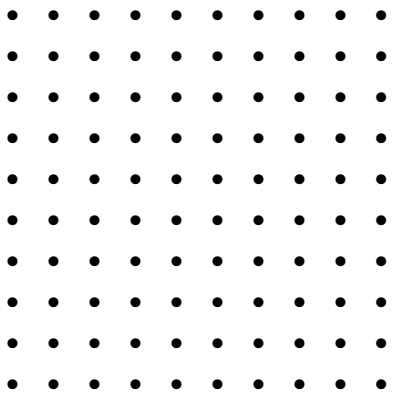
Propensity models might be compatible with our approach. Defer consideration.

Even without informative missingness the problem is a challenge.

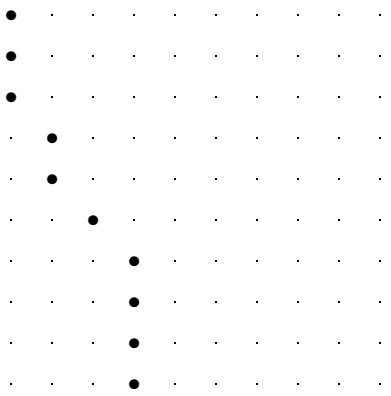
Observation patterns

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

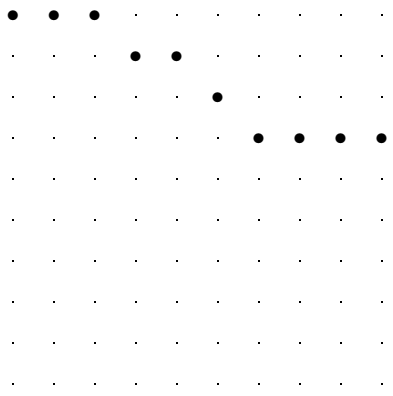
Crossed



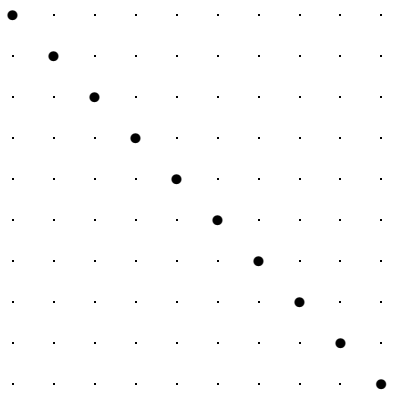
Row nested in col



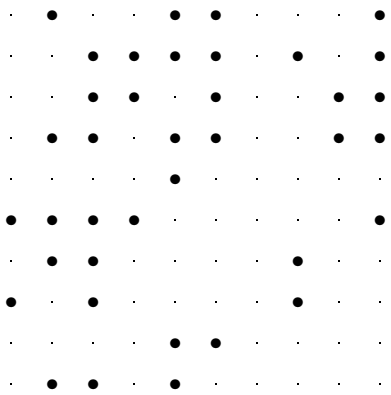
Col nested in row



IID

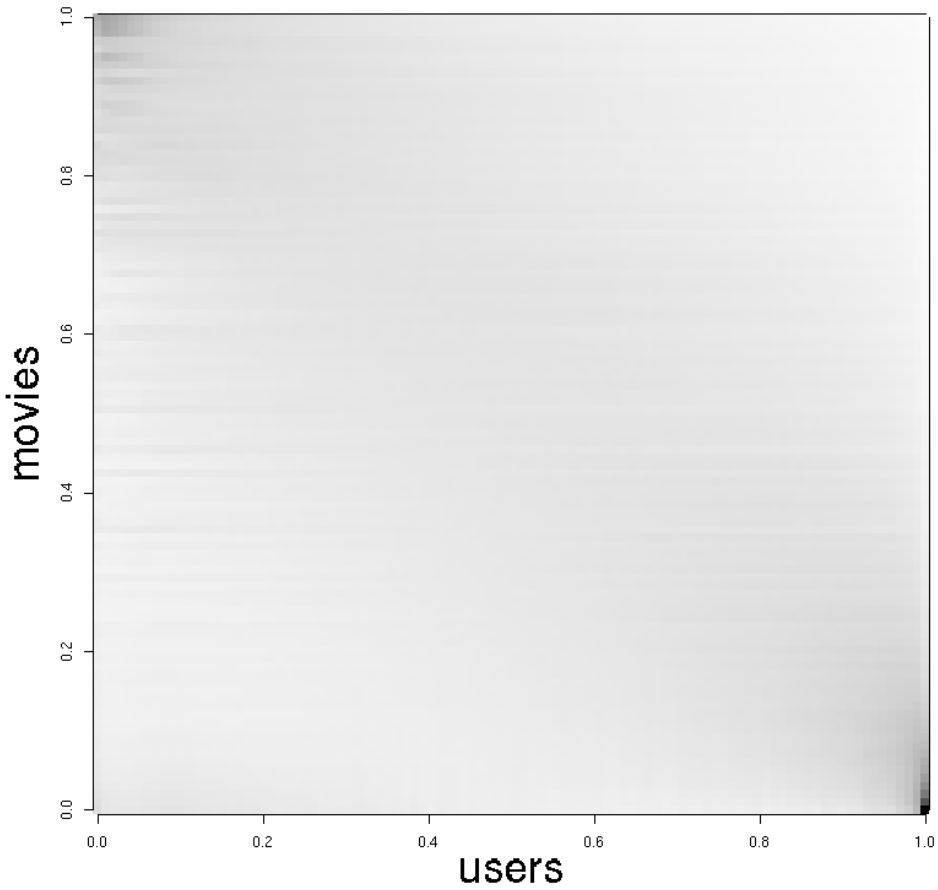


Arbitrary



Netflix

Bennett and Lanning (2007)



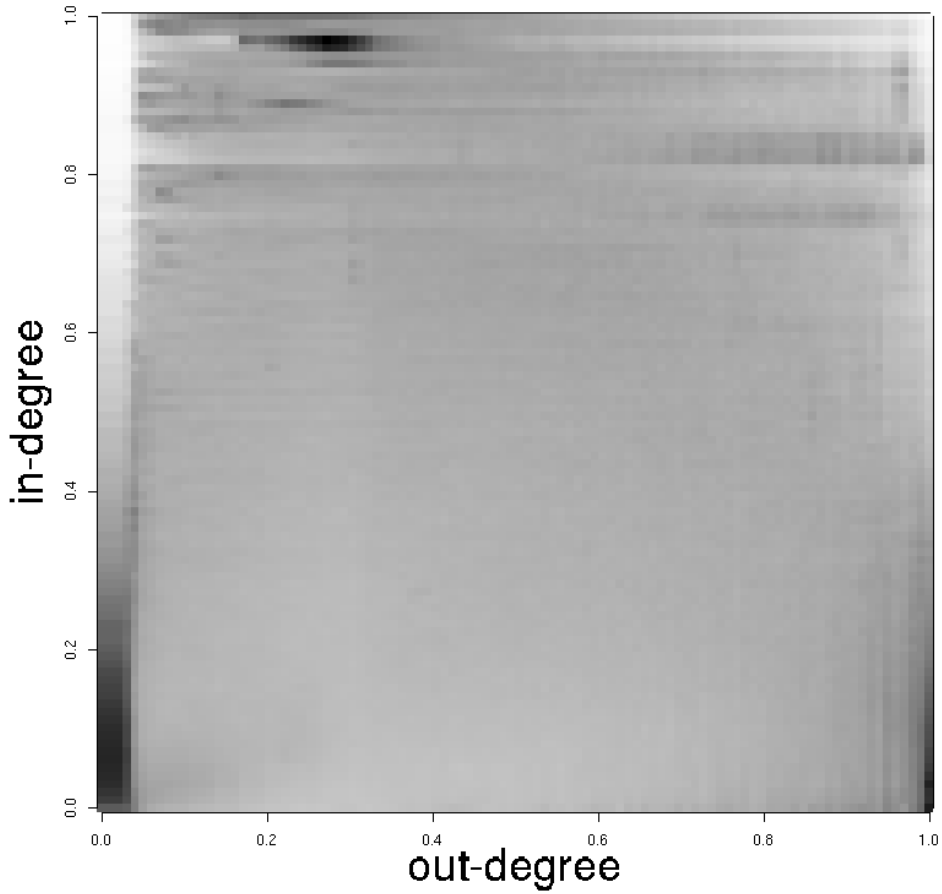
Users sorted by # ratings

Movies sorted too

Image from thesis of Justin Dyer

This is a copula

Wikipedia

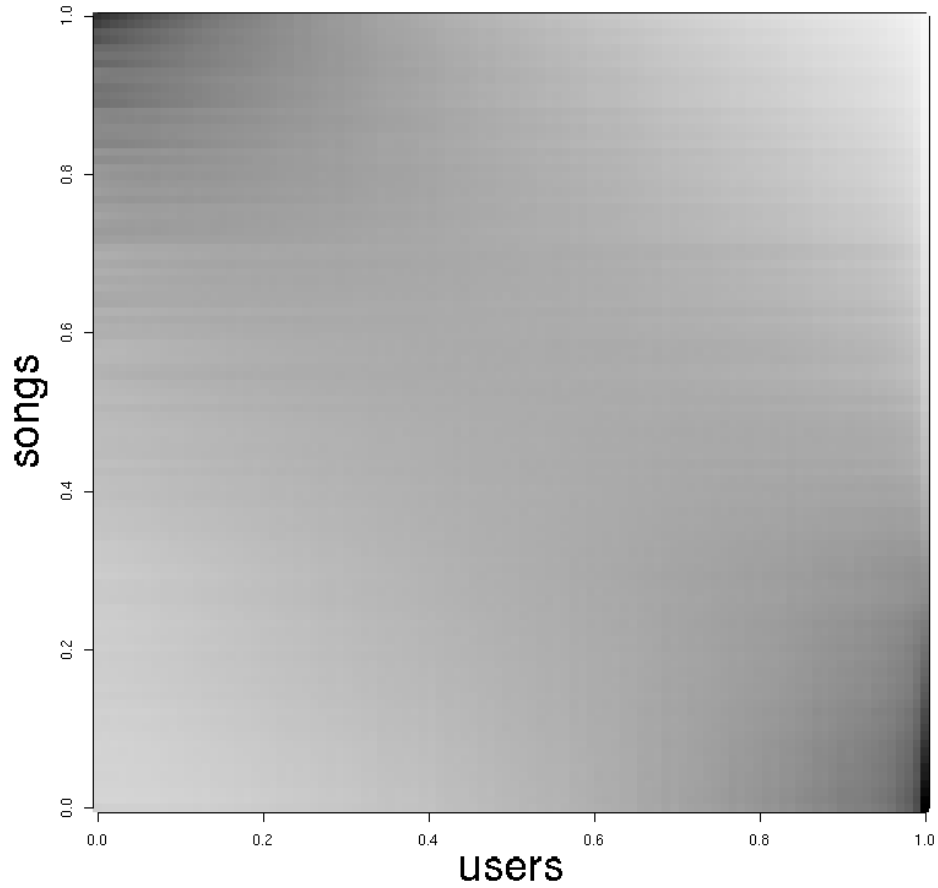


Z_{ij} is 1 if article i links to j

Data from David Gleich

Image from thesis of Justin Dyer

Yahoo! songs



Data from Yahoo! Webscope

Image from thesis of Justin Dyer

Sample size quantities

$$Z_{ij} = 1 \iff Y_{ij} \text{ observed}$$

Derived quantitiies

$$N = \sum_i \sum_j Z_{ij} \qquad 1 \leqslant N < \infty$$

$$N_{i\bullet} = \sum_j Z_{ij} \qquad \text{'size' of row } i$$

$$N_{\bullet j} = \sum_i Z_{ij} \qquad \text{'size' of col } j$$

$$R = \sum_i 1_{N_{i\bullet}>0} \qquad \text{\# unique observed rows}$$

$$C = \sum_j 1_{N_{\bullet j}>0} \qquad \text{\# unique observed cols}$$

Criteria

1) **Must** get $\hat{\beta}$ within $O(N)$ time and $O(R+C)$ space.

2) **Should** get a reliable $\widehat{\mathrm{Var}}(\hat{\beta})$ accounting for random effects.

3) **Prefer** a statistically efficient $\hat{\beta}$.

Relative importance

- Item 1 is a hard constraint.
- For item 2, a mildly conservative variance estimate should be ok for large N.

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \begin{cases} 2 \times \mathrm{Var}(\hat{\beta}) & \Longrightarrow \text{ OK} \\ \mathrm{Var}(\hat{\beta})/1000 & \Longrightarrow \text{ not OK} \end{cases}$$

- We sacrifice item 3. E.g., 50% efficiency could be ok for large N.

Ordinary least squares

Put x_{ij} into $X \in \mathbb{R}^{N\times p}$ and Y_{ij} into $Y \in \mathbb{R}^N$ (compatible ij order).

$$\begin{aligned}\hat\beta_{\mathrm{OLS}} &= (X^{\mathsf T}X)^{-1}X^{\mathsf T}Y\\ &= \Big(\sum_{ij} Z_{ij}x_{ij}x_{ij}^{\mathsf T}\Big)^{-1}\sum_{ij} Z_{ij}x_{ij}Y_{ij}\end{aligned}$$

Cost is $O(Np^2+p^3)$. Normal eqns. May use extended precision.

Variance

$$\mathrm{Var}(\hat\beta_{\mathrm{OLS}}) = (X^{\mathsf T}X)^{-1}X^{\mathsf T}\mathrm{Cov}(Y)X(X^{\mathsf T}X)^{-1}$$

NOT $(X^{\mathsf T}X)^{-1}\sigma^2$.

CAN be computed in $O(N)$ time. (Woodbury formula.)

OLS is not efficient. Gauss-Markov

Stitch Fix data

We got $N = 5{,}000{,}000$ ratings by $R = 762{,}752$ clients on $C = 6{,}318$ items.
This a subset of their customer / inventory base.

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

Predictors include Match$_{ij}$, a prediction from some baseline model
(not representative of all their algos).

Also whether item is 'Edgy' or 'Boho'. Same for client. Also indicators of material, leather, fur, acrylic, . . . , wool.

$p = 30$, including intercept.

Caveats

- A data analysis would consider many models. We will look at just one.
- People should be skeptical about coefficients. We use them to illustrate the consequences of crossed random effects.

From OLS results

The coefficient for Match was 5.05. A naive OLS standard error was 0.012.
Accounting for row and column correlations we estimate a true standard error of 0.146 for this OLS estimate; about 12.5x larger.

Effective N about $12.5^2 \doteq 150$ fold smaller.
$\sim 33{,}000$.

For Netflix data

$N \doteq 10^8$. Only about $17{,}000$ movies.
Unequally weighted: n_{eff} for movies $\approx 2{,}000$.

Generalized least squares

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

$$V = \mathrm{Cov}(Y) \in \mathbb{R}^{N \times N}$$

Correlation structure

$$V_{ij,i'j'} = \mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \sigma_A^2 1_{i=i'} + \sigma_B^2 1_{j=j'} + \sigma_E^2 1_{i=i'} 1_{j=j'}$$

We need to be able to compute $V^{-1}x$ for $x \in \mathbb{R}^N$.

Do it $p+1$ times to get $V^{-1}X$ and $V^{-1}Y$.

Illustration of $\mathrm{Cov}(Y)$

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

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

Apply labels 1 to 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 7 is in row $i = 3$ and column $j = 2$.

Generalized least squares

$$\hat{\beta}_{\text{GLS}} = (X^\mathsf{T} V^{-1} X)^{-1} X^\mathsf{T} V^{-1} Y \qquad N = 8, \quad R = 3, \quad C = 4$$

Row effect correlations, $R = 3$

	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 effect correlations, $C = 4$

	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

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

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

We need $V^{-1}X$. These 3 matrices are not simultaneously diagonalizable. Cholesky decomp of an $(R+C) \times (R+C)$ matrix. Costs $O(R^3 + C^3)$ to solve. Sherman-Morrison-Woodbury does not help further.

Linear mixed models

These require solving systems of equations $(R+C)\times(R+C)$.

The Cholesky decompositions cost $O(R^3+C^3)$.

See Bates (2014), Raudenbush (1993) for LMM.

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

Upshot

Crossed: linear mixed models and GLMMs have superlinear cost.

Nested: linear models have a block diagonal structure. Linear cost.

LMM computation

The best is Bates' most recent Julia code.

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

The number of iterations varies with N.

It crashed (for us) at some N in the millions. Crashed on Stitch Fix data.

We think it uses more than $O(R+C)$ working memory.

What about MCMC?

- empirically it mixes slowly for crossed data
- quite unlike successes in the nested case, e.g.
  Yu & Meng (2011) interweaving, Gelman et al. STAN
- we can prove it mixes slowly in some special cases (balanced Gaussian)
  Gibbs takes $O(N^{1/2})$ iterations of cost $O(N)$ each.
- we see numerically that unbalance does not help much if at all

To define an MCMC algorithm, we need to specify more about the data. We consider Gaussian effects a_i, b_j and ε_{ij}. Several priors for σ_A^2, σ_B^2 and σ_E^2.

Literature check

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

Crossed: Very few MCMC papers.

MCMC continued

1) For Gibbs and intercept only and balanced Gaussian data using Roberts and Sahu (1997) we prove that it takes $O(\sqrt{N})$ iterations to converge at $O(N)$ cost each. Cost is $O(N^{3/2})$.

2) For scattered missing data we compute the Roberts and Sahu (1997) rate and see similar slow mixing.

3) Similar problems for

   (a) random walk Metropolis,

   (b) Langevin

   (c) Metropolis adjusted Langevin,

   (d) RWM with subsampling

   (e) Conditional augmentation

   (f) pCN Hairer

4) For consensus MCMC of Scott et al. (2013). We cannot split data into independent parts.

MCMC and solving equations

Gibbs samples one variable at a time.

Resembles solving a quadratic, one at a time.

To solve $Vx = b$,

write $V = M - N$ with M invertible. Then

$$x \leftarrow M^{-1}(Nx + b).$$

To each M, N there is a Markov chain updating $x \sim \mathcal{N}(0, V)$.

Colin Fox:

convergence rate in the chain

$=$ convergence rate in the linear solver

Then computing $V^{-1}x$ will be about as hard as simulating $\mathcal{N}(0, V)$.

GLS for row effects only

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + \varepsilon_{ij}$$

$\mathrm{Cov}(Y)$ has R blocks, each $N_{i\bullet} \times N_{i\bullet}$, 'diagonal plus rank one':

$$\mathrm{Cov}(Y) \equiv V_R = \sigma_E^2 I + \sigma_A^2 \begin{pmatrix} 11^{\mathsf{T}} & 0 & \cdots & 0 \\ 0 & 11^{\mathsf{T}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 11^{\mathsf{T}} \end{pmatrix}$$

We can invert V_R and do GLS in $O(N)$ cost, via **Woodbury** formula.

We need estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$.

Column effects only

Same situation by symmetry.

What we do

Total error: $\eta_{ij} = a_i + b_j + \varepsilon_{ij}$

1) Get $\hat{\beta}_{\mathrm{OLS}}$ and $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}_{\mathrm{OLS}}$.

2) Estimate $\sigma_A^2, \sigma_B^2, \sigma_E^2$ from $\hat{\eta}_{ij}$ (by moments below).

3) Get either $\hat{\beta}_{\mathrm{RLS}}$ or $\hat{\beta}_{\mathrm{CLS}}$ using $\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\sigma}_E^2$.

4) Re-estimate $\sigma_A^2, \sigma_B^2, \sigma_E^2$ from $\hat{\eta}_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}}\hat{\beta}_{\mathrm{RLS}}$ (or $\hat{\beta}_{\mathrm{CLS}}$)

5) Estimate $\mathrm{Var}(\hat{\beta}_{\mathrm{RLS}})$ using latest $\hat{\sigma}_A^2, \hat{\sigma}_B^2, \hat{\sigma}_E^2$ (or $\mathrm{Var}(\hat{\beta}_{\mathrm{CLS}})$)

Total cost

Is $O(N)$ time and $O(R + C)$ space.

Fixed number of passes over data.

Details for steps 2,3,4,5 next.

Method of moments

$$U_a(\hat\beta) = \sum_i S_{i\bullet}, \qquad S_{i\bullet} = \sum_j Z_{ij}(\hat\eta_{ij} - \bar{\hat\eta}_{i\bullet})^2$$

$$U_b(\hat\beta) = \sum_j S_{\bullet j}, \qquad S_{\bullet j} = \sum_i Z_{ij}(\hat\eta_{ij} - \bar{\hat\eta}_{\bullet j})^2$$

$$U_e(\hat\beta) = \sum_{ij} Z_{ij}(\hat\eta_{ij} - \bar{\hat\eta}_{\bullet\bullet})^2,$$

3 equations 3 unknowns

$$\mathbb{E}\begin{pmatrix} U_a \\ U_b \\ U_e \end{pmatrix} = M \begin{pmatrix} \sigma_A^2 \\ \sigma_B^2 \\ \sigma_E^2 \end{pmatrix} \quad \text{for} \quad M = \begin{pmatrix} 0 & N-R & N-R \\ N-C & 0 & N-C \\ N^2 - \sum_i N_{i\bullet}^2 & N^2 - \sum_j N_{\bullet j}^2 & N^2 - N \end{pmatrix}$$

$M \in \mathbb{R}^{3\times 3}$ depends on Z_{ij}

$$\begin{pmatrix} \hat\sigma_A^2 \\ \hat\sigma_B^2 \\ \hat\sigma_E^2 \end{pmatrix} = M^{-1} \begin{pmatrix} U_a \\ U_b \\ U_e \end{pmatrix}$$

More details

These variances can be computed in one pass Chan, Golub, Leveque (1983).

They take $O(N)$ time and $O(R+C)$ space.

Gao & O. (2017)

Henderson

This is a U-statistic version of Henderson I. Searle, Casella, McCulloch.

The sample variances are conveniently represented as U-statisics.

See thesis Gao (2017) for more.

$\mathbf{Cov}((U_a, U_b, U_e)^\mathsf{T})$

These come from the intercept-only model

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$$

Applicable to $Y_{ij} - x_{ij}^\mathsf{T}\hat{\beta}_{\mathrm{OLS}}$ when $\hat{\beta}_{\mathrm{OLS}}$ is consistent.

Now we have

$$\hat{\sigma}_A^2, \quad \hat{\sigma}_B^2, \quad \hat{\sigma}_E^2$$

Extensive formulas and derivation omitted.

We use fourth moments to estimate kurtoses in order to get mildly conservative

$$\widehat{\mathrm{Cov}}\begin{pmatrix}\hat{\sigma}_A^2\\ \hat{\sigma}_B^2\\ \hat{\sigma}_E^2\end{pmatrix} = M^{-1}\,\widehat{\mathrm{Cov}}\begin{pmatrix}U_a\\ U_b\\ U_e\end{pmatrix}(M^{-1})^\mathsf{T}.$$

See Gao & O (2017)

Exact variances

$$\begin{aligned}\mathrm{Var}(U_a) &= \sigma_B^4(\kappa_B+2)\sum_{ir}(ZZ^{\mathsf{T}})_{ir}(1-N_{i\bullet}^{-1})(1-N_{r\bullet}^{-1})\\ &\quad+2\sigma_B^4\sum_{ir}N_{i\bullet}^{-1}N_{r\bullet}^{-1}(ZZ^{\mathsf{T}})_{ir}[(ZZ^{\mathsf{T}})_{ir}-1]+4\sigma_B^2\sigma_E^2(N-R)\\ &\quad+\sigma_E^4(\kappa_E+2)\sum_i N_{i\bullet}(1-N_{i\bullet}^{-1})^2+2\sigma_E^4\sum_i(1-N_{i\bullet}^{-1})\end{aligned}$$

$\mathrm{Var}(U_b)$ is similar, $\mathrm{Var}(U_e)$ is longer. Also covariances Gao & O (2017)

Under some conditions

$$\mathrm{Var}(\hat\sigma_A^2)\sim\sigma_A^4(\kappa_A+2)\sum_i N_{i\bullet}^2\;/\;N^2$$

$$\mathrm{Var}(\hat\sigma_B^2)\sim\sigma_B^4(\kappa_B+2)\sum_j N_{\bullet j}^2\;/\;N^2$$

$$\mathrm{Var}(\hat\sigma_E^2)\sim\sigma_E^4(\kappa_E+2)/N$$

As if the other components were zero. Asymptotically uncorrelated.

Row GLS or col GLS?

We do RLS (GLS via row random effects only) if

$$\sigma_A^2 \max_i N_{i\bullet} > \sigma_B^2 \max_j N_{\bullet j}$$

Otherwise we do CLS (via columns).

Motivation comes from Kantorovich inequality about worst case inefficiencies.

In hindsight

we could just do both $\hat{\beta}_{\mathrm{RLS}}$ and $\hat{\beta}_{\mathrm{CLS}}$

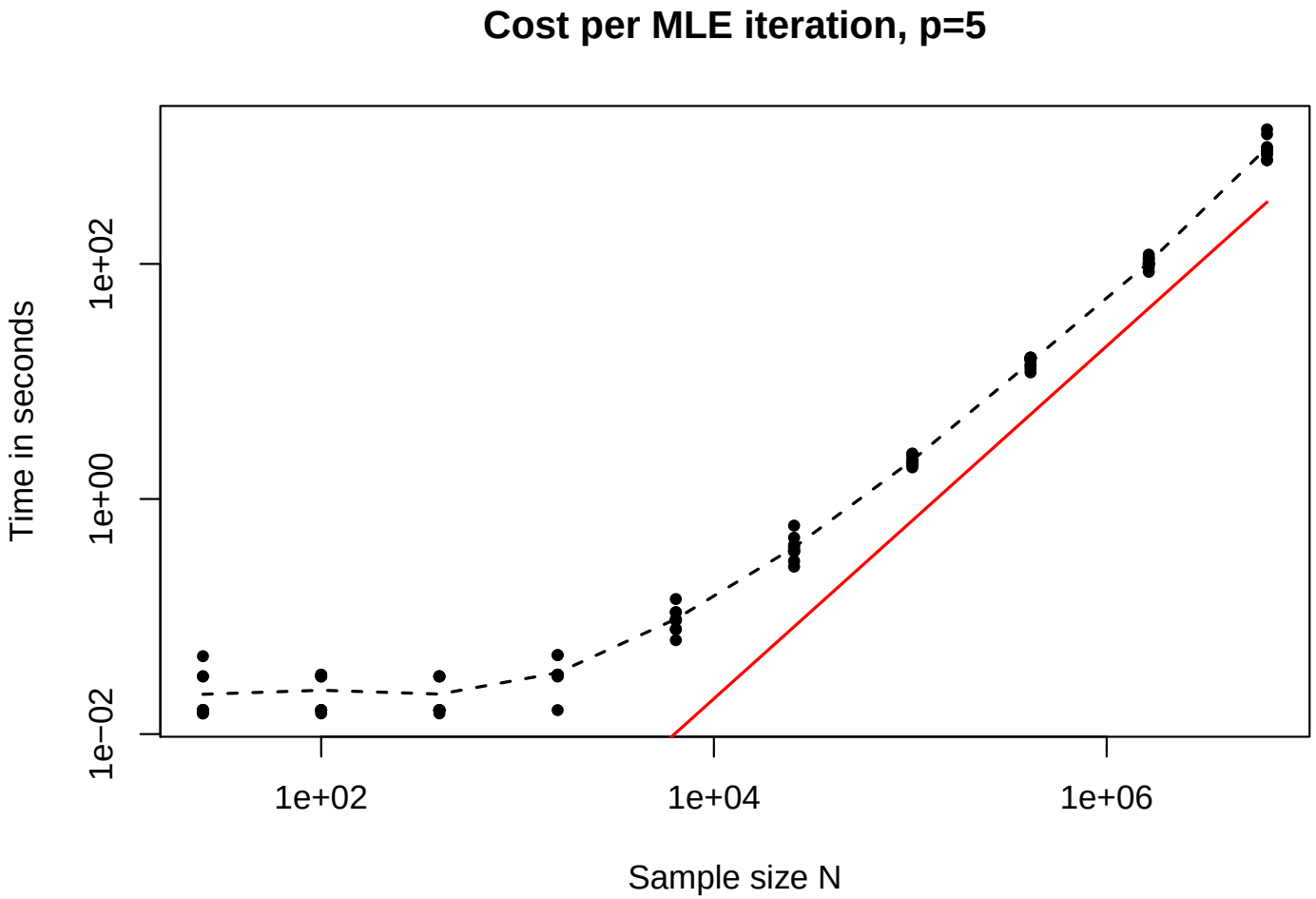
Row based GLS

Woodbury $\implies$ $\hat{\beta}_{\mathrm{RLS}}$ in $O(N)$ time.

Woodbury and $\hat{\sigma}^2_A, \hat{\sigma}^2_B, \hat{\sigma}^2_E \implies \widehat{\mathrm{Var}}_{\mathrm{GLS}}(\hat{\beta}_{\mathrm{RLS}})$ in $O(N)$ time.

This $\widehat{\mathrm{Var}}$ accounts for σ^2_A, σ^2_B and σ^2_E,
even though $\hat{\beta}_{\mathrm{RLS}}$ is derived from just σ^2_A and σ^2_E

Cost of one MLE iteration



MixedModels.jl Bates (2016)

$R = C, \quad N = RC/4, \quad$ random $Z_{ij}, \quad \sigma_A^2 = 2, \quad \sigma_B^2 = 1/2, \quad \sigma_E^2 = 1$

LMM vs $\hat{\beta}_{\mathrm{RLS}}$, efficiency

In our simulations LMM has about double the efficiency for β and σ^2_E.

About equal efficiency for σ^2_A and σ^2_B.

Non efficiency issues

LMM cannot extend to large N.

LMM confidence intervals assume normality.

Model for ratings

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

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

Notes

- Material_{ij} is a categorical variable, that we replaced by indicator variables. We chose 'Polyester', the most common material, as the baseline.
- We can study whether edgy items work best for edgy clients.
- We will interpret regression coefficients at face value, though of course such interpretations have usual caveats and may need followup testing.

	$\hat{\beta}_{\text{OLS}}$	$\widehat{\text{se}}_{\text{OLS}}()$	$\widehat{\text{se}}(\hat{\beta}_{\text{OLS}})$	$\hat{\beta}$	$\widehat{\text{se}}(\hat{\beta})$
Intercept	4.635^*	0.00539	0.05808	5.110^*	0.01250
Match	5.048^*	0.01174	0.1464	3.529^*	0.02153
$\mathbb{I}\{$c. edgy$\}$	0.0010	0.00244	0.0046	0.00186	0.003831
$\mathbb{I}\{$i. edgy$\}$	-0.3358^*	0.00425	0.03730	-0.3328^*	0.01542
$\mathbb{I}\{$c. edgy$\}$					
$*\mathbb{I}\{$i. edgy$\}$	0.3925^*	0.00622	0.01352	0.3864^*	0.006432
$\mathbb{I}\{$c. boho$\}$	0.1386^*	0.00226	0.004354	0.1334^*	0.003622
$\mathbb{I}\{$i. boho$\}$	-0.5499^*	0.00598	0.03049	-0.6261^*	0.01661
$\mathbb{I}\{$c. boho$\}$					
$*\mathbb{I}\{$i. boho$\}$	0.3822^*	0.00756	0.01057	0.3837^*	0.007697
Acrylic	-0.06482^*	0.00377	0.03804	-0.01627	0.02149
Angora	-0.01262	0.00784	0.09631	0.07271	0.05837
Bamboo	-0.04593	0.06215	0.2437	0.05420	0.1716
Cashmere	-0.1955^*	0.02484	0.1593	0.01354	0.1176

Boho?

Edgy items don't do well, unless they are sent to edgy customers.

WIth Boho, we see

$$0.13 \times \text{clientboho} - 0.63 \times \text{itemboho} + 0.38 \times \text{clientboho} \times \text{itemboho}$$

The negative does not get compensated.

Maybe this works when the algo making Match_{ij} is very confident.

We repeated the analysis keeping only client-item pairs with a high level of match, but the pattern persisted.

Asymptotics

Sample size quantities:

$$\epsilon_R = \max_i N_{i\bullet}/N$$

$$\epsilon_C = \max_j N_{\bullet j}/N$$

For a covariance matrix $V \in \mathbb{R}^{p\times p}$:

$$\mathcal{I}(V) = \text{Min eigenvalue}(V)$$

$$\mathcal{I}_0(V) = \text{Min eigenvalue}(V_{2:p,2:p}) \quad \text{(to exclude the intercept)}$$

For $\hat{\beta}_{\mathrm{OLS}} \xrightarrow{\mathrm{p}} \beta$, we need

- $\epsilon_R \to 0$
- $\epsilon_C \to 0$
- $\mathcal{I}(X^{\mathsf{T}}X/N) \geqslant c > 0$

Variance components

Consistency for $\hat{\sigma}^2_A, \hat{\sigma}^2_B, \hat{\sigma}^2_E$ from some $\hat{\beta}$:

We need

- $\hat{\beta} \xrightarrow{\text{p}} \beta$
- $\max(\epsilon_R, \epsilon_C) \to 0$
- $\max\left(\frac{R}{N}, \frac{C}{N}\right) \leqslant \theta < 1$
- $M_N \equiv \max_{ij} Z_{ij} \|x_{ij}\|^2$ bounded.*

*Bounded x_{ij} could be weakened.

Row least squares, consistency

Challenge: handling the contribution of b_j (column random effects) to row-based GLS estimate $\hat{\beta}_{\mathrm{RLS}}$.

We need

- $\hat{\sigma}_A^2 \xrightarrow{\mathrm{p}} \sigma_A^2, \quad \hat{\sigma}_E^2 \xrightarrow{\mathrm{p}} \sigma_E^2 > 0$
- $\max(\epsilon_R, \epsilon_C) \to 0$
- $\mathcal{I}_0\Big(\frac{1}{N}\sum_{ij} Z_{ij}(x_{ij} - \bar{x}_{i\bullet})(x_{ij} - \bar{x}_{i\bullet})^{\mathsf{T}}\Big) \geqslant c > 0$
- $\frac{1}{R^2}\sum_{ir}(ZZ^{\mathsf{T}})_{ir} N_{i\bullet}^{-1} N_{r\bullet}^{-1} \to 0$

The last one is at most $\frac{1}{R}\sum_i N_{i\bullet}^{-1}$ and can be much smaller.

CLT for $\hat{\beta}_{\mathrm{RLS}}$

$$\tilde{x}_{\bullet j} \equiv \frac{1}{N_{\bullet j}} \sum_i Z_{ij} \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/N_{i\bullet}} \bar{x}_{i\bullet} \qquad \text{Second order avg for col } j$$

$$k \equiv \sum_j N_{\bullet j}^2 (\bar{x}_{\bullet j} - \tilde{x}_{\bullet j}) \Big/ \sum_j N_{\bullet j}^2 \qquad \text{Wtd avg of col deviations}$$

We need

- $\hat{\sigma}_A^2 \xrightarrow{\mathrm{p}} \sigma_A^2, \quad \hat{\sigma}_E^2 \xrightarrow{\mathrm{p}} \sigma_E^2 > 0$
- $\mathcal{I}(\sum_i \bar{x}_{i\bullet} \bar{x}_{i\bullet}^{\mathsf{T}}) \to \infty$
- $\mathcal{I}_0(\sum_{ij} Z_{ij}(x_{ij} - \bar{x}_{i\bullet})(x_{ij} - \bar{x}_{i\bullet})^{\mathsf{T}}) \to \infty$
- $\mathcal{I}_0(\sum_j N_{\bullet j}^2 (\bar{x}_{\bullet j} - \tilde{x}_{\bullet j} - k)(\bar{x}_{\bullet j} - \tilde{x}_{\bullet j} - k)^{\mathsf{T}}) \,/\, \sum_j N_{\bullet j}^2 \to \infty$
- Plus two ugly technical conditions for the intercept.

Martingale central limit theorems

Apply to $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$, and $\hat{\sigma}^2_E$.

They require stronger and more complicated sufficient conditions.

See thesis of Katelyn Gao (2017)

Henderson's ladder

Henderson I: $\mu + a_i + b_j + \varepsilon_{ij}$.

Henderson II: $x_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}$. I.e., allow mixed effects.

Henderson III: allow interactions between fixed and random components.

Our analysis allows for non-zero kurtoses.

Searle, Casella, McCulloch (1992) on Henderson III

> "Computationally, the method can involve the inversion of large-sized matrices — of order equal to the number of levels of the effects in the model. This disadvantage will decline as today's computing power increases in speed and declines in cost (per arithmetic operation)."

That looked like a good bet in 1992, but then data growth accelerated.

1990s $\neq$ 1890s $\doteq$ 2010s

Further steps

- higher way tables
- SVD-like interactions
- heteroscedastic effects
- backfitting iterations, suggested by Trevor Hastie
- logistic regression

Eckles & O (2012) handle first three items for bootstrap sampling of means
(should apply to smooth fns of means)
It takes $O(BN)$ work for B bootstrap samples.

Also

What happens to predictions and their standard errors?

What happens for ensemble learners?

Thanks

- Co-author Katelyn Gao
- Brad Klingenberg (Stitch Fix) data and discussions
- Yahoo!, Netflix, David Gleich for data
- Justin Dyer for images
- NSF DMS-1407397
- NSF Graduate Research Fellowship grant DGE-114747*

· · · and of course

*Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.