Method of moments for large crossed random effects data

Art B. Owen, Stanford University
and Katelyn Gao, Stanford University

Work still in progress; first paper soon.

Summary

- e-commerce generates large crossed random effects data
- GLMMs* are appropriate, but have superlinear cost
- and so does every MCMC we looked at

Let's go back to the '90s (1890s)

1) Method of moments costs $O(N)$
2) Weak assumptions
3) No tuning parameters
4) No convergence diagnostics
5) Works in parallel
6) Can give $\hat{\sigma}^2 < 0$ ☹
7) May require extended precision ☹

*Generalized linear mixed models, e.g., Doug Bates' work.

E-commerce data

Logs look like

$$(i, j, \ldots, r, s, \;\; x, y)$$

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

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

Variables x, y

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

X : time of day page load speed home city experimental var $\cdots$

Stitch Fix

garments $\times$ customers $\times$ curators

Brad Klingenberg

Factor $\not\simeq$ categorical variable

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

Big categorical variable

- very large number of levels
- can appear with a power law frequency (Taylor Swift $\cdots$ heteroscedasticity)
- usually we have not seen them all yet
- can be many hapax legomena ('onesies')

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.

Most basic model

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}, \quad i, j \in \mathbb{N}$$

$$a_i \overset{\text{iid}}{\sim} (0, \sigma_A^2)$$

$$b_j \overset{\text{iid}}{\sim} (0, \sigma_B^2)$$

$$\varepsilon_{ij} \overset{\text{iid}}{\sim} (0, \sigma_E^2)$$

$$\mathbb{E}(Y_{ij}^4) < \infty$$

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

Informative missingness?

Consider that later

Ok, later starts now

Netflix: people (mostly) rate movies they watch, and those are (mostly) movies they expect to like. So there should be an upward bias.

Yelp: motivation to rate increases for very low ratings and for very high ones. So there should be increased variance.

Handling informative missingness

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

The uncertainty in a solution will depend in part on the data and in part on the outside information/assumptions. Our methods may help quantify uncertainty arising from the data.

Roadmap

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

$$Y_{ij} = X_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij} \qquad \textbf{soon}$$

$$\mathrm{logit}(\mathbb{P}(Y_{ij} = 1)) = X_{ij}^{\mathsf{T}}\beta + a_i + b_j \qquad \textbf{later}$$

$$Y_{ij} = X_{ij}^{\mathsf{T}}\beta + a_i + b_j + u_i^{\mathsf{T}} v_j + \varepsilon_{ij} \qquad \textbf{maybe}$$

$$\mathrm{logit}(\mathbb{P}(Y_{ij} = 1)) = X_{ij}^{\mathsf{T}}\beta + a_i + b_j + u_i^{\mathsf{T}} v_j \qquad \textbf{maybe}$$

Comparison

The models get more useful as we proceed from top to bottom. Also more difficult.

The basic model

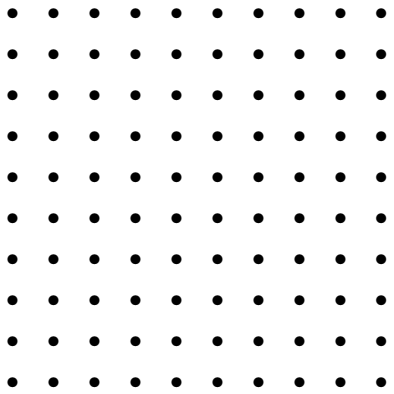
Knowing σ_A^2, σ_B^2, σ_E^2 would help you shrink towards row, column and overall mean.

It is the simplest model with no satisfactory treatment.

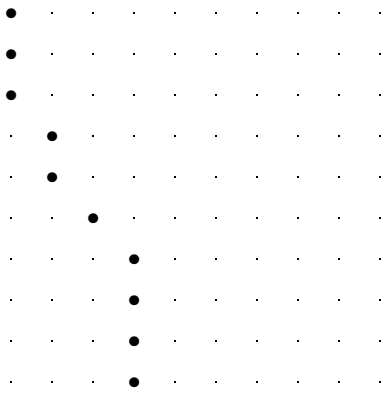
Observation patterns

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

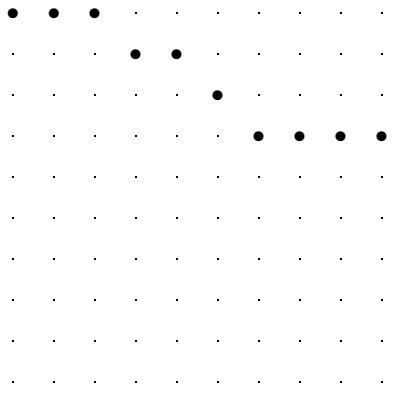
Crossed



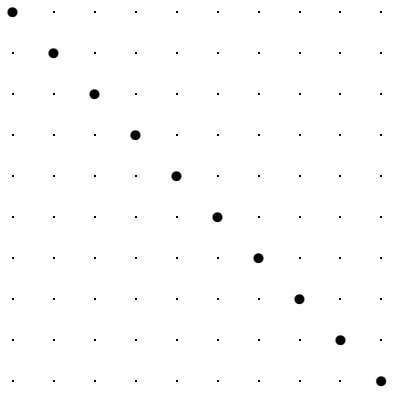
Row nested in col



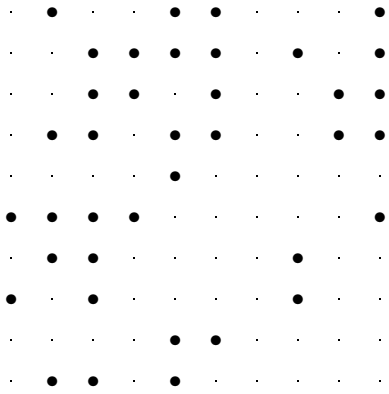
Col nested in row



IID

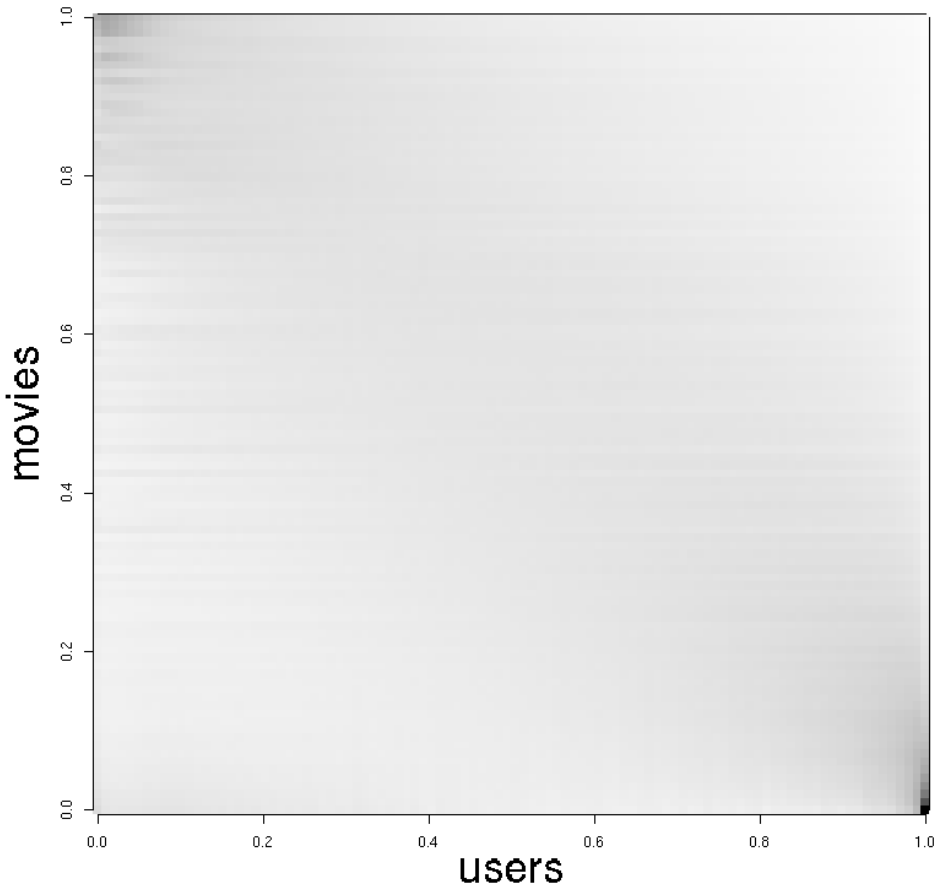


Arbitrary



Netflix

Bennett and Lanning (2007)

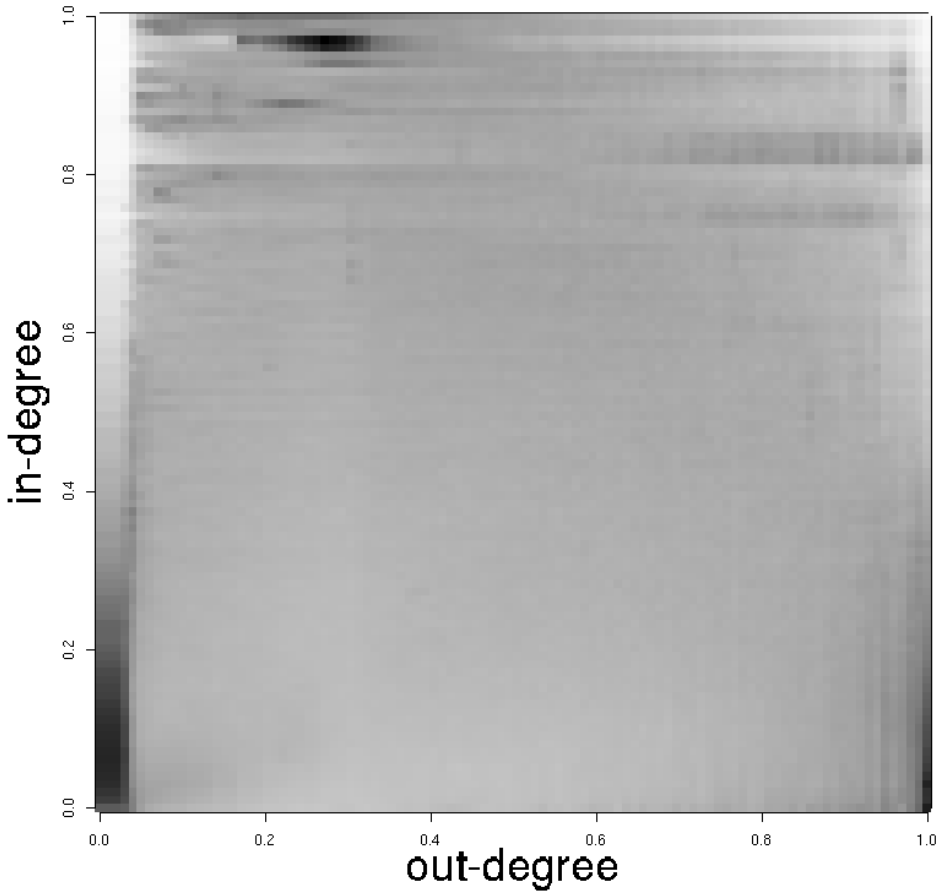


Users sorted by # ratings

Movies sorted too

Image from thesis of Justin Dyer

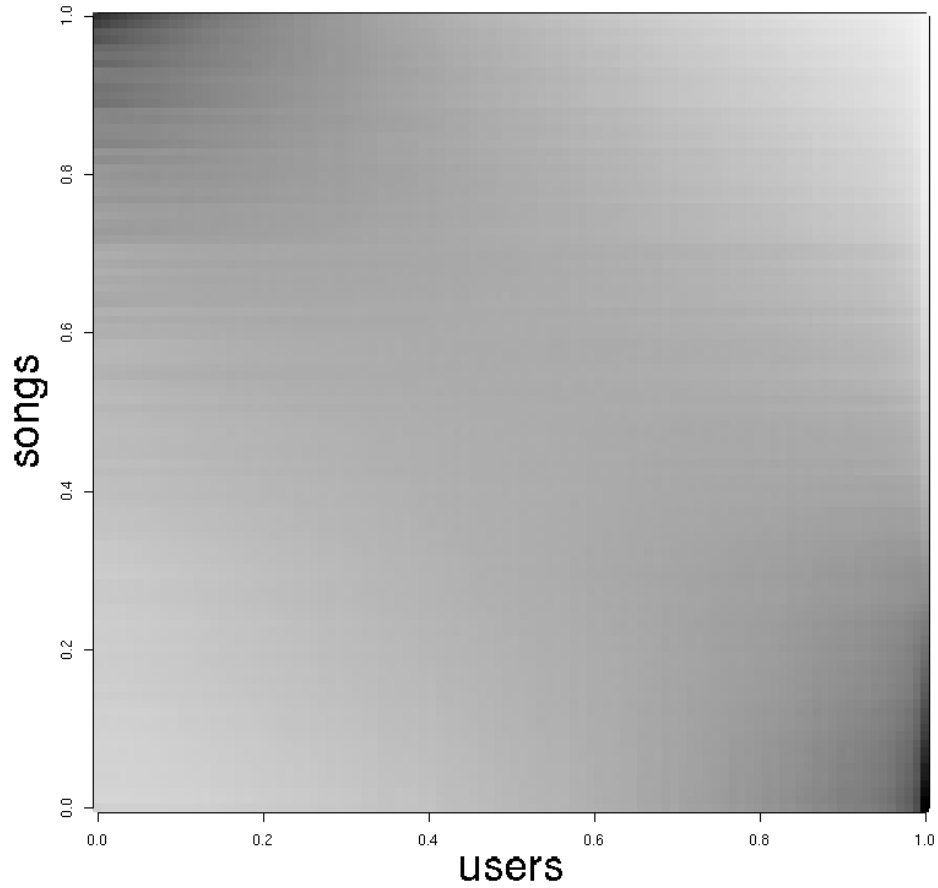
Wikipedia



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

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

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

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

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

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

Linear mixed models

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

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

Bates (2014), Raudenbush (1993)

$$N \leqslant RC \implies \max\{R, C\} \geqslant \sqrt{N} \implies \text{cost} > cN^{3/2}$$

Upshot

Crossed: linear mixed models and GLMMs have superlinear cost.

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

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)
- 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

Nested: Lots of MCMC papers, theory and applied.

Crossed: Very few MCMC papers.

Convergence rates

Roberts and Sahu (1997)

Let $\theta^{(n)}$ for $n \in \mathbb{N}$ be a Markov chain with stationary distribution h. Its convergence rate is the minimum number ρ such that

$$\lim_{n\to\infty} \mathbb{E}_h\Big(\big(\mathbb{E}_h(f(\theta^{(n)}) \mid \theta^{(0)}) - \mathbb{E}_h(f(\theta))\big)^2\Big) r^{-n} = 0$$

holds for all $\mathbb{E}_h(f(\theta)^2) < \infty$ and all $r > \rho$.

Upshot

If ρ is close to 1, L^2 convergence is slow, i.e., ρ^n

Gibbs

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}, \quad \text{with independent}$$

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

Suppose $\mu, \sigma_A^2, \sigma_B^2, \sigma_E^2, Y$ are known for $i = 1, \ldots, R$ and $j = 1, \ldots, C$

Gibbs updates for a and b

$$a^{(t+1)} \sim \mathcal{N}\left(\frac{\sigma_A^2 \sum_j (Y_{ij} - \mu - b_j^{(t)})}{\sigma_E^2 + C\sigma_A^2}, \frac{\sigma_A^2 \sigma_E^2}{\sigma_E^2 + C\sigma_A^2} I_R \right)$$

$$b^{(t+1)} \sim \mathcal{N}\left(\frac{\sigma_B^2 \sum_j (Y_{ij} - \mu - a_i^{(t+1)})}{\sigma_E^2 + R\sigma_B^2}, \frac{\sigma_B^2 \sigma_E^2}{\sigma_E^2 + R\sigma_B^2} I_C \right)$$

Gibbs ctd.

Using methods from Roberts and Sahu (1997) we find that

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_E^2/R} \times \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/C} = \Big(1 - O\Big(\frac{1}{R}\Big)\Big)\Big(1 - O\Big(\frac{1}{C}\Big)\Big)$$

If R and C grow in proportion, then $\rho = 1 - O(1/\sqrt{N})$

Cost

$O(\sqrt{N})$ iterations and $O(N)$ cost per iteration $\implies$ superlinear

(The balanced case can actually be updated faster than $O(N)$
just cache the row and columns sums.)

Unbalanced case

Caching does not work for arbitrary Z. We can compute ρ numerically and still see ρ approaching 1 as R and C grow.

MCMC algorithms

Gibbs is not the only way.

Random walk Metropolis

Convergence time grows like $d = R + C = O(\sqrt{N})$ (at best)

Roberts and Rosenthal (2001) (prior a product distn)

Cost per iteration $O(N)$ total cost $O(N^{3/2})$

Metropolis adjusted Langevin

Convergence time grows like $d^{1/3} = (R + C)^{1/3}$

Roberts and Rosenthal (2001) (indep prior) Cost is $O(N(R + C)^{1/3})$ at best $O(N^{1+1/6})$.

Puzzler

Q: Why does MCMC work for high dimensional hierarchical models?

$\hat{A}$: Of course $o(d)$ is also $O(d)$, but **why** are nested models nice?

Simulations

Non-informative priors for μ, σ^2_A, σ^2_B, σ^2_E

Also for Gibbs: point priors at the true values

Algorithms

- Gibbs
- Block Gibbs (sample a, b jointly). Usually infeasible.
- Reparametrization. Conditional augmentation of van Dyk and Meng (2001)
- Random walk Metropolis (RWM)
- RWM with subsampling
- pCN Hairer et al. (2014) 'RWM with shrunken proposal means'
- Langevin
- MALA

All failed

Some had bad ACF(μ), others bad ACF(σ^2), some bad both.

Variance refresher (IID case)

For $Y_i \overset{\text{iid}}{\sim} (0, \sigma^2), i = 1, \ldots, n$

U-statistic version

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{2}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}(Y_i - Y_{i'})^2$$

Mean and variance

$$\mathbb{E}(s^2) = \sigma^2$$

$$\mathrm{Var}(s^2) = \sigma^4\left(\frac{2}{n-1} + \frac{\kappa}{n}\right)$$

$$\kappa = \mathbb{E}((Y-\mu)^4)/\sigma^4 - 3 \qquad \text{kurtosis}$$

Miller (1986)

U-statistics

$$U_a = \frac{1}{2}\sum_i\sum_j\sum_{j'} N_{i\bullet}^{-1} Z_{ij}Z_{ij'}(Y_{ij}-Y_{ij'})^2$$

$$U_b = \frac{1}{2}\sum_i\sum_j\sum_{i'} N_{\bullet j}^{-1} Z_{ij}Z_{i'j}(Y_{ij}-Y_{i'j})^2$$

$$U_e = \frac{1}{2}\sum_i\sum_j\sum_{i'}\sum_{j'} Z_{ij}Z_{i'j'}(Y_{ij}-Y_{i'j'})^2$$

3 equations 3 unknowns

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

M depends on Z_{ij}

Why $N_{i\bullet}^{-1}$?

$$U_a = \frac{1}{2}\sum_i\sum_j\sum_{j'} N_{i\bullet}^{-1} Z_{ij}Z_{ij'}(Y_{ij}-Y_{ij'})^2$$

$$\vdots$$

$$= \sum_i (N_{i\bullet}-1)s_i^2, \quad \text{where}$$

$$s_i^2 = \frac{1}{N_{i\bullet}-1}\sum_j Z_{ij}(Y_{ij}-\bar{Y}_{i\bullet})^2 \qquad \text{taking } 0 \text{ if } N_{i\bullet}=1$$

- $\mathbb{E}(s_i^2) = \sigma_B^2 + \sigma_E^2$ sample variance in row i
- weight roughly inversely to variance
- zero weight for $N_{i\bullet} < 2$

The matrix M

$$M = \begin{pmatrix} 0 & N-R & N-R \\ N-C & 0 & N-C \\ N^2 - \sum_i N_{i\bullet}^2 & N^2 - \sum_j N_{\bullet j}^2 & N^2 - N \end{pmatrix}$$

We need it to be invertible

$$\det(M) = (N-R)(N-C)\Big(N^2 - \sum_i N_{i\bullet}^2 - \sum_j N_{\bullet j}^2 + N\Big)$$

We need more than one column (so $N > R$) more than one row (so $N > C$).

The third factor is positive if no row or column has over half the data.

Variance

Let $\theta = \begin{pmatrix} \sigma_A^2 & \sigma_B^2 & \sigma_E^2 \end{pmatrix}^{\mathsf{T}}$ and $U = \begin{pmatrix} U_a & U_b & U_e \end{pmatrix}^{\mathsf{T}}$

$$\hat{\theta} = M^{-1}U \implies \mathrm{Var}(\hat{\theta}) = M^{-1}\mathrm{Var}(U)(M^{-1})^{\mathsf{T}}$$

We need $\mathrm{Var}(U)$

It depends on kurtoses κ_A, κ_B, κ_E and on Z_{ij}. We will plug-in.

Henderson 0

This is like Henderson I from Searle, Casella & McCulloch (1992)

Except we account for kurtosis (non-normal data)

Also, we will upper bound the variances, to keep cost $O(N)$

and we use U-statistics instead of ANOVA mean squares

Constraints

$O(N)$ time, $O(R + C)$ extra space

Can take more than one pass over data (we need two)

Conservative estimates ok.

Sample size idioms

Rows $i\ i'\ r\ r'$ Cols $j\ j'\ s\ s'$ $\sum_i$ is $\sum_{i:N_{i\bullet}>0}$ $\sum_j$ is $\sum_{j:N_{\bullet j}>0}$

$$(ZZ^{\mathsf{T}})_{ir} = \sum_j Z_{ij}Z_{rj} \leqslant N_{i\bullet}$$ overlap rows i & r

$$(Z^{\mathsf{T}}Z)_{js} = \sum_i Z_{ij}Z_{is} \leqslant N_{\bullet j}$$ overlap cols j & s

$$\sum_{ir}(ZZ^{\mathsf{T}})_{ir} = \sum_{ijr} Z_{ij}Z_{rj} = \sum_j N_{\bullet j}^2$$

$$\sum_{js}(Z^{\mathsf{T}}Z)_{js} = \sum_{ijs} Z_{ij}Z_{is} = \sum_i N_{i\bullet}^2$$

After some algebra

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

Can't always be done in $O(N)$ work

One pass of $O(N)$ work gets N, R, C, $N_{i\bullet}$, $N_{\bullet j}$.

Subsequent $O(R)$ computation gets last 3 terms.

Double sums $\sum_{ir}$ problematic. Hides a $\sum_{ijr}$. Usually not $O(N)$.

$\mathrm{Var}(U_b)$ has the same issue.

The other parts of $\mathrm{Var}(U)$ can be done in two passes of $O(N)$.

$\mathrm{Var}(U_e)$ is the messiest. (9 terms)

Upper bounds

Fourth and smallest term below is the problem:

$$\sum_{ir}(ZZ^{\mathsf{T}})_{ir}(1-N_{i\bullet}^{-1})(1-N_{r\bullet}^{-1}) = \sum_{ir}(ZZ^{\mathsf{T}})_{ir}(1-N_{i\bullet}^{-1}-N_{r\bullet}^{-1}+N_{i\bullet}^{-1}N_{r\bullet}^{-1})$$

First upper bound

$$\begin{aligned}\sum_{ir}(ZZ^{\mathsf{T}})_{ir}(1-N_{i\bullet}^{-1})(1-N_{r\bullet}^{-1}) &\leqslant \sum_{ir}(ZZ^{\mathsf{T}})_{ir}(1-N_{i\bullet}^{-1})\\ &= \sum_{ijr} Z_{ij}Z_{rj}(1-N_{i\bullet}^{-1})\\ &= \sum_{j} N_{\bullet j}^2 - \sum_{ij} Z_{ij}N_{\bullet j}N_{i\bullet}^{-1} \quad \text{costs } O(N)\end{aligned}$$

Second upper bound

$$\begin{aligned}\sum_{ir} N_{i\bullet}^{-1}N_{r\bullet}^{-1}(ZZ^{\mathsf{T}})_{ir}((ZZ^{\mathsf{T}})_{ir}-1)) &\leqslant \sum_{ir} N_{i\bullet}^{-1}(ZZ^{\mathsf{T}})_{ir}\\ &= \sum_{ij} Z_{ij}N_{\bullet j}N_{i\bullet}^{-1}\end{aligned}$$

Comparisons

$$\sum_j N_{\bullet j}^2 \quad \text{vs} \quad \sum_{ij} Z_{ij} N_{\bullet j} N_{i\bullet}^{-1}$$

Divide both by N

$$\frac{1}{N}\sum_j N_{\bullet j}^2 = \frac{1}{N}\sum_{ij} Z_{ij} N_{\bullet j}$$

$$\text{versus} \quad \frac{1}{N}\sum_{ij} Z_{ij} N_{\bullet j} N_{i\bullet}^{-1}$$

Pick random ij with $Z_{ij} = 1$. Report $N_{\bullet j}$ versus $N_{\bullet j}/N_{i\bullet}$.

Ordinarily

$\sum_{ij} Z_{ij} N_{\bullet j} N_{i\bullet}^{-1} \ll \sum_i N_{i\bullet}^2$ and $\sum_{ij} Z_{ij} N_{i\bullet} N_{\bullet j}^{-1} \ll \sum_j N_{\bullet j}^2$

Kurtoses

$$W_a = \frac{1}{2}\sum_{ijj'} N_{i\bullet}^{-1} Z_{ij} Z_{ij'} (Y_{ij} - Y_{ij'})^4$$

$$W_b = \frac{1}{2}\sum_{iji'} N_{\bullet j}^{-1} Z_{ij} Z_{i'j} (Y_{ij} - Y_{i'j})^4, \quad \text{and}$$

$$W_e = \frac{1}{2}\sum_{iji'j'} Z_{ij} Z_{i'j'} (Y_{ij} - Y_{i'j'})^4.$$

$$\mathbb{E}(W_a) = (\mu_{B,4} + 3\sigma_B^4 + 12\sigma_B^2\sigma_E^2 + \mu_{E,4} + 3\sigma_E^4)(N-R), \qquad \mu_{B,4} = \mathbb{E}(b^4)$$

Plug in $\hat\sigma_A^2, \hat\sigma_B^2, \hat\sigma_E^2$

Solve for $\mu_{A,4}, \mu_{B,4}, \mu_{E,4}$ then for $\kappa_A, \kappa_B, \kappa_E$

Plug those into $\mathrm{Var}(U)$.

Kurtoses can be hard to estimate, but big data helps here

(unless Y_{ij} is heavy tailed)

Theorem

If for some small $\epsilon > 0$ the following hold

$$N_{i\bullet} \leqslant \epsilon N, \quad N_{\bullet j} \leqslant \epsilon N, \quad R \leqslant \epsilon N, \quad C \leqslant \epsilon N, \quad N \leqslant \epsilon \sum_i N_{i\bullet}^2, \quad N \leqslant \epsilon \sum_j N_{\bullet j}^2$$

$$\sum_{ij} Z_{ij} N_{i\bullet}^{-1} N_{\bullet j} \leqslant \epsilon \sum_i N_{i\bullet}^2, \quad \text{and} \quad \sum_{ij} Z_{ij} N_{i\bullet} N_{\bullet j}^{-1} \leqslant \epsilon \sum_j N_{\bullet j}^2$$

$$0 < \underline{m} \leqslant \kappa_A + 2, \kappa_B + 2, \kappa_E + 2, \sigma_A^4, \sigma_B^4, \sigma_E^4 \leqslant \overline{m} < \infty.$$

Then $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\sigma}_E^2$ are asymptotically uncorrelated as $\epsilon \to 0$ with

$$\mathrm{Var}(\hat{\sigma}_A^2) = \sigma_A^4(\kappa_A + 2)\frac{1}{N^2}\sum_j N_{i\bullet}^2(1 + O(\epsilon))$$

$$\mathrm{Var}(\hat{\sigma}_B^2) = \sigma_B^4(\kappa_B + 2)\frac{1}{N^2}\sum_j N_{\bullet j}^2(1 + O(\epsilon)), \quad \text{and}$$

$$\mathrm{Var}(\hat{\sigma}_E^2) = \sigma_E^4(\kappa_E + 2)\frac{1}{N}(1 + O(\epsilon)).$$

Yahoo! Webscope movie data*

Movies rated on a 13 point scale $A+ \quad A \quad A- \quad \cdots \quad D- \quad F$

$N = 211{,}231 \quad R = 7642$ (users) $\quad C = 11{,}916$ (movies) $\quad \frac{RC}{N} \doteq 431.1$

Row and column sizes

$$\max_i \frac{N_{i\bullet}}{N} \doteq 0.0077 \qquad \max_j \frac{N_{\bullet j}}{N} \doteq 0.020$$

$$\frac{1}{N}\sum_i N_{i\bullet}^2 \doteq 124.4 \qquad \frac{1}{N}\sum_j N_{\bullet j}^2 \doteq 650.6$$

$$\frac{1}{N}\sum_{ij} Z_{ij} N_{\bullet j}/N_{i\bullet} \doteq 33.1 \qquad \frac{1}{N}\sum_{ij} Z_{ij} N_{i\bullet}/N_{\bullet j} \doteq 13.1$$

Estimates

$$\hat\sigma_A^2 = 2.56 \qquad \hat\sigma_B^2 = 2.86 \qquad \hat\sigma_E^2 = 7.68$$

* Yahoo! Webscope dataset ydata-ymovies-user-movie-ratings-train-v1_0

`http://research.yahoo.com/Academic_Relations`

Yahoo! Webscope song data*

Songs rated on a 100 point scale. First million ratings.

$$N = 1{,}000{,}000 \quad R = 16{,}685 \text{ (users)} \quad C = 13{,}930 \text{ (songs)} \quad \frac{RC}{N} \doteq 232.4$$

Row and column sizes

$$\max_i \frac{N_{i\bullet}}{N} \doteq 0.0080 \qquad \max_j \frac{N_{\bullet j}}{N} \doteq 0.0068$$

$$\frac{1}{N}\sum_i N_{i\bullet}^2 \doteq 409.6 \qquad \frac{1}{N}\sum_j N_{\bullet j}^2 \doteq 1339.1$$

$$\frac{1}{N}\sum_{ij} Z_{ij} N_{\bullet j}/N_{i\bullet} \doteq 32.6 \qquad \frac{1}{N}\sum_{ij} Z_{ij} N_{i\bullet}/N_{\bullet j} \doteq 22.4$$

Estimates

$$\hat\sigma_A^2 = 819.0 \qquad \hat\sigma_B^2 = 257.8 \qquad \hat\sigma_E^2 = 603.8$$

* Yahoo! Webscope dataset ydata-ymusic-user-artist-ratings-v1_0

http://research.yahoo.com/Academic_Relations

Shrinkage

Predict/smooth Y_{ij} by

$$\lambda_0 \sum_{ij} Z_{ij} Y_{ij} + \lambda_a \sum_{j} Z_{ij} Y_{ij} + \lambda_b \sum_{i} Z_{ij} Y_{ij}$$

New row and new column

shrinks to $\hat{\mu}$

New row and old column

quickly approaches column mean as $N_{\bullet j}$ increases

Old row and old column

$$1/\epsilon \leqslant N_{i\bullet}, N_{\bullet j} \leqslant \epsilon N \implies \hat{Y}_{ij} = (\bar{Y}_{i\bullet} + \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})(1 + O(\epsilon)).$$

assuming σ_A^2 etc bounded away from 0 and ∞

Regression

$$Y_{ij} = X_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}, \quad i, j \in \mathbb{N}, \quad \text{observed iff } Z_{ij} = 1$$

$X_{ij} \in \mathbb{R}^p$ key variables, $p \ll \min(R, C)$

$\beta \in \mathbb{R}^p$ parameter of interest

$a_i, b_j =$ nuisance random effects

Correlation structure

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

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

Generalized least squares

$$\hat{\beta}_{GLS} = (X^\mathsf{T} V^{-1} X)^{-1} X^\mathsf{T} V^{-1} Y \qquad N = 8, \quad R = 3, \quad C = 2$$

Row 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

Col correlations, $C = 2$

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

GLS continued

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

We would need $V^{-1}X$

Costs $O(R^3 + C^3)$ to solve. Sherman-Morrison-Woodbury does not help further.

Colin Fox shows linear solvers converge at the same rate MCMC does.

OLS

We will use $\hat{\beta} = \hat{\beta}_{OLS} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$

But $\widehat{\mathrm{Var}}(\hat{\beta}_{OLS})$ will account for a_i and b_j

It is also possible to account for row correlations (or column correlations) but not both. $\hat{\beta}_{ROW}$ or $\hat{\beta}_{COL}$

Preliminary results

Worst case efficiency losses via Kantorovich

'Average' cases less conservative.

Further steps

- complete regression
- try logistic regression
- higher way tables
- interactions
- heteroscedastic effects

Eckles & O (2012) handle last three items for bootstrap sampling of means (should apply to smooth fns of means)

Thanks

- Co-author Katelyn Gao
- Brad Klingenberg (StitchFix) 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.