Random projections, reweighting and half-sampling for high-dimensional statistical inference

Art B. Owen

Stanford University

based on joint works with:

Dean Eckles Facebook Inc.

Sarah Emerson Oregon State University

About these slides

These are the slides I presented on February 15 at MCQMC 2012 in Sydney Australia.

I have corrected some typos and extended the presentation of the challenging integral over the Stiefel manifold.

A few of these slides were skipped over in order to allow time for questions.

This talk covers two projects. The bootstrap work with Dean Eckles has now been accepted by the Annals of Applied statistics. The projection work with Sarah Emerson is still in progress.

Monte Carlo methods for statistics

1) Markov chain Monte Carlo

2) Bootstrap resampling

The above are mainstays. We also use:

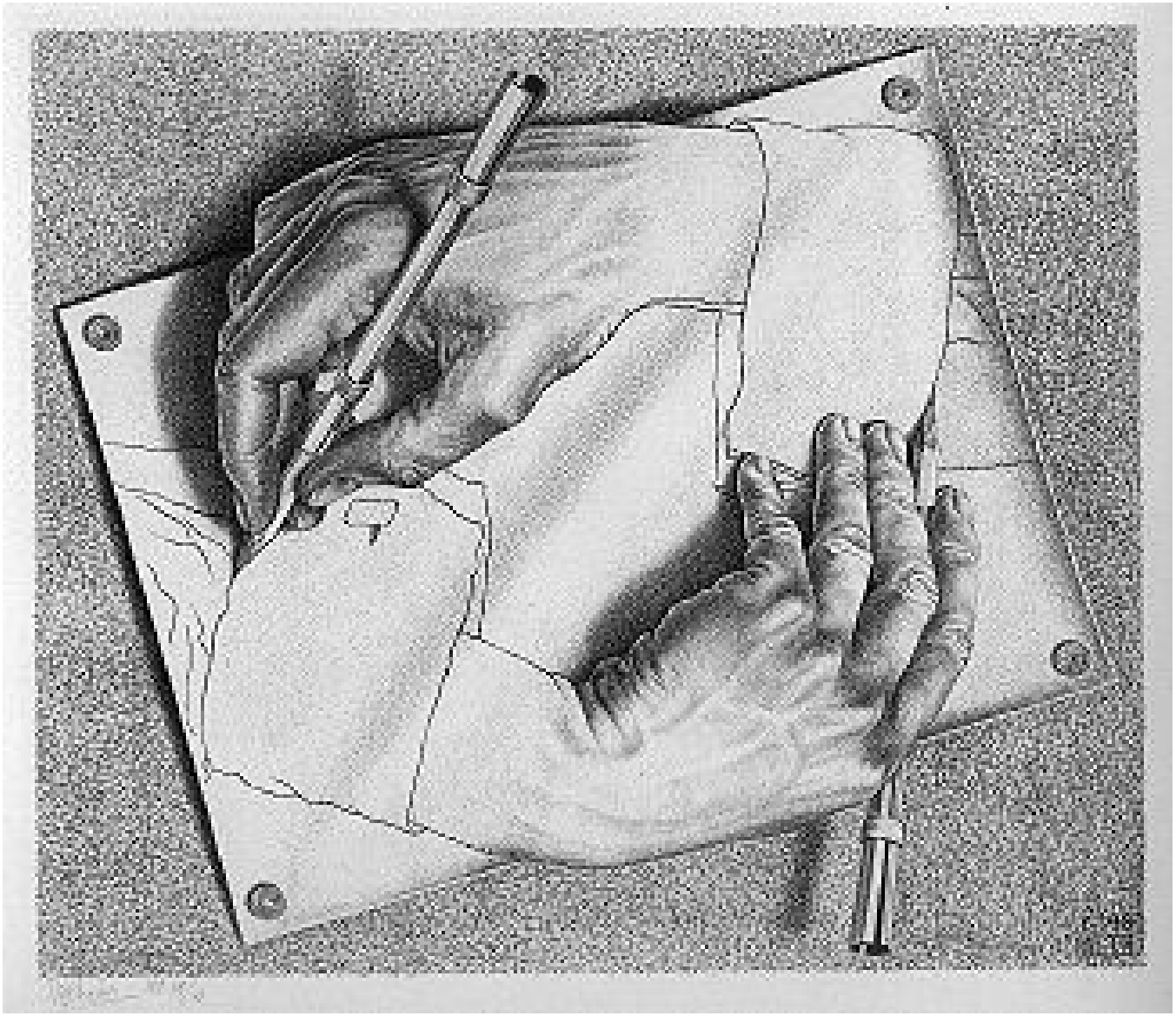
1) Random permutations

2) Random projections

3) Sample splitting

Probability/Statistics and Monte Carlo are closely intertwined.

Statistics and Monte Carlo



M. C. Escher (1948)

This talk will show some uses of MC uses in statistics.

Some statistical notions

$\boldsymbol{X} \sim F$ random vector $\boldsymbol{X}$ has distribution F

$\boldsymbol{X}_i \overset{\text{iid}}{\sim} F$ $\boldsymbol{X}_i$ are statistically **I**ndependent and **I**dentically **D**istributed (IID) from F

$\mathcal{N}_d(\mu, \Sigma)$ The Gaussian distribution with mean $\mu \in \mathbb{R}^d$

and variance covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$

$\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$, means

$$\Pr(\boldsymbol{X} \in A) = \int_A f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{where}$$

$$f(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^\mathsf{T}\Sigma^{-1}(\boldsymbol{x}-\mu)\right)}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}$$

p-values

Observe $T = t$ and compute $p = \Pr(T \geqslant t)$.

If $p < 0.01$ then the observed value t happens 1% or less of the time.

Evidence against the hypothesized distribution of T.

Problem one

We have

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_x} \overset{\text{iid}}{\sim} F \text{ in } \mathbb{R}^d$$
$$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_y} \overset{\text{iid}}{\sim} G \text{ in } \mathbb{R}^d$$

is $F = G$?

We might assume $F = \mathcal{N}(\mu_1, \Sigma)$ and $G = \mathcal{N}(\mu_2, \Sigma)$.

Then we test $\mu_1 = \mu_2$. This is an old problem.

Revived interest, $d \gg n_x + n_y$

DNA microarrays

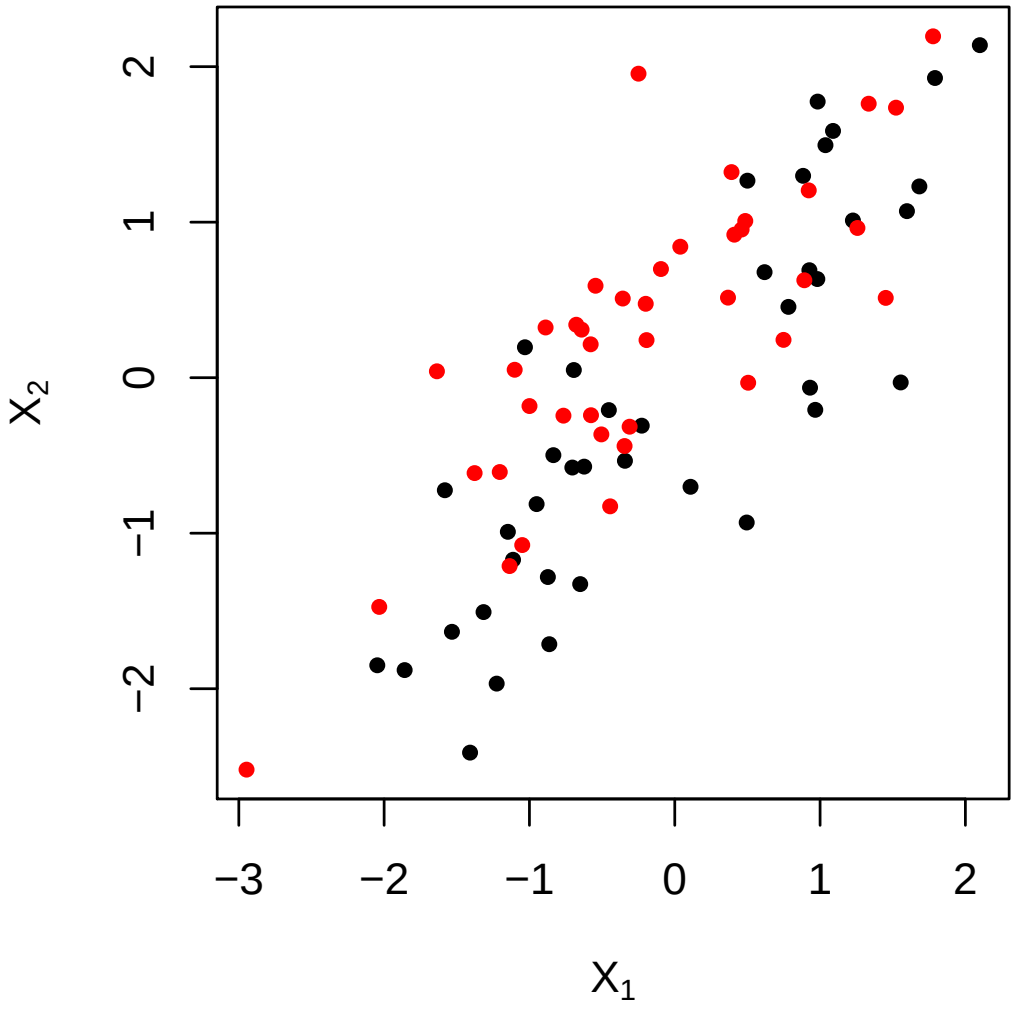
- expression levels of $d \approx 30{,}000$ genes
- on n_x healthy and n_y diseased individuals
- n_x, n_y tens or hundreds

Genome wide association studies

- $d \approx 2{,}000{,}000$ markers
- with n_x, n_y thousands or more

Also: fMRI, finance

Illustration



50 red and 50 black points in $\mathbb{R}^2$

Black points normally distributed

Red points shifted Northwest wrt black

X_1 not significantly different,

$p = 0.47$

X_2 not significantly different,

$p = 0.09$

$X_1 + X_2$ not significantly different,

$p = 0.60$

$X_1 - X_2$ **very** significantly different,

$p = 1.7 \times 10^{-4}$

So: how to find the interesting projection?

Hotelling's T^2

Find $\theta \in \mathbb{R}^d$ with $\theta^{\mathsf{T}}\theta = 1$ to maximize the apparent separation between

$\widetilde{X}_i = \theta^{\mathsf{T}} \boldsymbol{X}_i \in \mathbb{R}$ and $\widetilde{Y}_i = \theta^{\mathsf{T}} \boldsymbol{Y}_i \in \mathbb{R}$

Answer depends on

$$\bar{\boldsymbol{X}} = \frac{1}{n_x}\sum_{i=1}^{n_x} \boldsymbol{X}_i \qquad S_x = \sum_{i=1}^{n_x} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\mathsf{T}}$$

$$\bar{\boldsymbol{Y}} = \frac{1}{n_y}\sum_{i=1}^{n_y} \boldsymbol{Y}_i \qquad S_y = \sum_{i=1}^{n_y} (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\mathsf{T}}$$

Algebraicly we get

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^{\mathsf{T}} S^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) \quad \text{where}$$

$$S = \frac{S_x + S_y}{n_x + n_y - 2}$$

Hotelling (1931)

Get $T^2 = 18.58$. Also $\Pr(T^2 \geqslant 18.58) = 2.6 \times 10^{-4}$ (p value)

In high dimensions

When $d \gg n_x + n_y$ the covariance matrix S is not invertible. Can't use

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^{\mathsf{T}} S^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$$

Geometrically

Some projection $\theta \in \mathbb{R}^d$ has

$\theta^{\mathsf{T}} \boldsymbol{X}_i =$ constant for $i = 1, \ldots, n_x$ and

$\theta^{\mathsf{T}} \boldsymbol{Y}_i =$ different constant.

IE: we will get perfect separation, even if $F = G$.

A classic remedy by Dempster (1958) takes

$$T^2_{\text{Dempster}} = \frac{n_x n_y}{n_x + n_y} \frac{\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\|^2}{\text{tr}(S)} = \frac{n_x n_y}{n_x + n_y} \frac{\sum_{j=1}^d (\bar{X}_j - \bar{Y}_j)^2}{\sum_{j=1}^d S_{jj}}$$

but this makes no use of correlations

Recent improvements by: Bai, Saradanasa, Hall, Fan, Chen, Srivastava

also don't use correlations

Random projections

Lopes, Jacob, Wainwright (2011)

Choose random $\Theta \in \mathbb{R}^{d\times k}$ with $\Theta^{\mathsf{T}}\Theta = I_k$.

Put $\widetilde{\boldsymbol{X}}_i = \Theta^{\mathsf{T}}\boldsymbol{X}_i$ and $\widetilde{\boldsymbol{Y}}_i = \Theta^{\mathsf{T}}\boldsymbol{Y}_i$

Then use

$$\widetilde{T}^2_\Theta = \frac{n_x n_y}{n_x + n_y}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^{\mathsf{T}}\Theta(\Theta^{\mathsf{T}}S\Theta)^{-1}\Theta^{\mathsf{T}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$$

Exists if $k < n_x + n_y - 2$

That is

project data into a random k dimensional subspace

test means of projected data

this retains some of the correlations

Uniform random projections

From $\mathbb{R}^d$ to $\mathbb{R}$: normalize a Gaussian vector

$$\theta = \frac{Z}{\|Z\|}, \quad Z \sim \mathcal{N}(0, I_k)$$

To project from $\mathbb{R}^d$ to $\mathbb{R}^k$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1k} \\ Z_{21} & Z_{22} & \cdots & Z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{d1} & Z_{d2} & \cdots & Z_{dk} \end{pmatrix} \in \mathbb{R}^{d \times k} \qquad Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

Gram-Schmidt yields $Z = QR$

deliver $\quad \Theta = Q \in \mathbb{R}^{d \times k}$

project $\quad \widetilde{\boldsymbol{X}}_i = \Theta^{\mathsf{T}} \boldsymbol{X}_i$

Any QR decomposition with positive R_{ii} will do.

Lopes et al. ctd.

They make just **one** random projection of the data

They find that $k \approx (n_x + n_y - 2)/2$ performs well

Why just one?

If your one projection is 'unlucky' then you might miss the pattern.

But with just one projection the distribution of $\widetilde{T}^2$ is known.

Multiple projections

$$\bar{T}^2 = \frac{1}{M}\sum_{i=1}^{M} \widetilde{T}_i^2$$

T_i^2 based on $\Theta_i \in \mathbb{R}^{d\times k}$

Get some kind of 'average' luck.

But distribution of $\bar{T}^2$ not known.

Multiple projections

Work with S. Emerson: average over M independent random $\Theta_i \in \mathbb{R}^{d\times k}$

$$\bar{T}^2 = \frac{1}{M}\sum_{i=1}^{M} \widetilde{T}_i^2, \qquad \text{where}$$

$$\widetilde{T}_i^2 = \frac{n_x n_y}{n_x + n_y}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^{\mathsf{T}}\Theta_i(\Theta_i^{\mathsf{T}} S \Theta_i)^{-1}\Theta_i^{\mathsf{T}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$$

Easily

1) $\mathbb{E}(\bar{T}^2) = \mathbb{E}(\widetilde{T}_i^2)$
2) $\mathrm{Var}(\bar{T}^2) < \mathrm{Var}(\widetilde{T}_i^2)$, unless both are infinite! (averaging reduces variance)

Less easily

1) Finite variance requires $k \leqslant n_x + n_y - 6$
2) Finite mean requires $k \leqslant n_x + n_y - 4$

Unfortunately: the distribution of $\bar{T}^2$ is not known.

Separation

Simulate 2000 data sets

$$\boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma), \quad \boldsymbol{Y}_i \overset{\text{iid}}{\sim} \mathcal{N}(\delta, \Sigma), \quad \delta \in \mathbb{R}^d$$

Of these:

1000 **null** cases	$\|\delta\| = 0$
1000 **non-null** cases	$\|\delta\| > 0$

Rank 2000 $\bar{T}^2$ scores

See if nulls get smaller $\bar{T}^2$ values.

The ROC* curve

Shown later, shows how well the test separates the two cases

***R**eceiver **O**perating **C**haracteristic (don't ask)

Simulated case

$\boldsymbol{X}_i, \boldsymbol{Y}_i \in \mathbb{R}^{200} \qquad n_x = n_y = 50$

Pick $\|\delta\| = 3$ uniform on 200 dimensional sphere

Pick $\Sigma = I_d \times 50/\sqrt{d}$

Why these

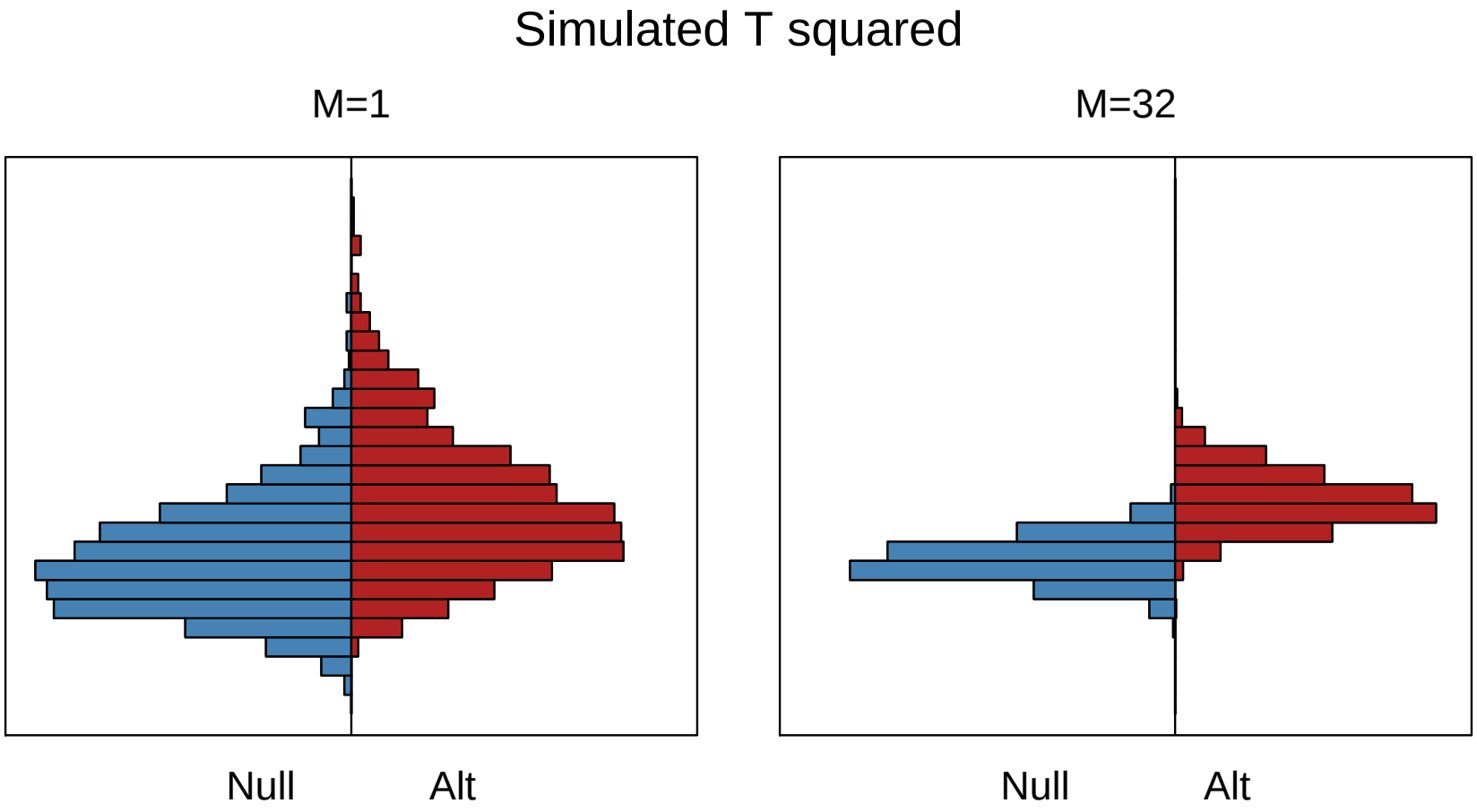
Uniform δ means that the group separation is unrelated to the covariance structure.

Debatable. We follow Lopes et al in making this assumption.

WLOG, under uniformity $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_d)$ $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$

Interesting cases are equal λ_j and rapidly decreasing λ_j

Multiple projections

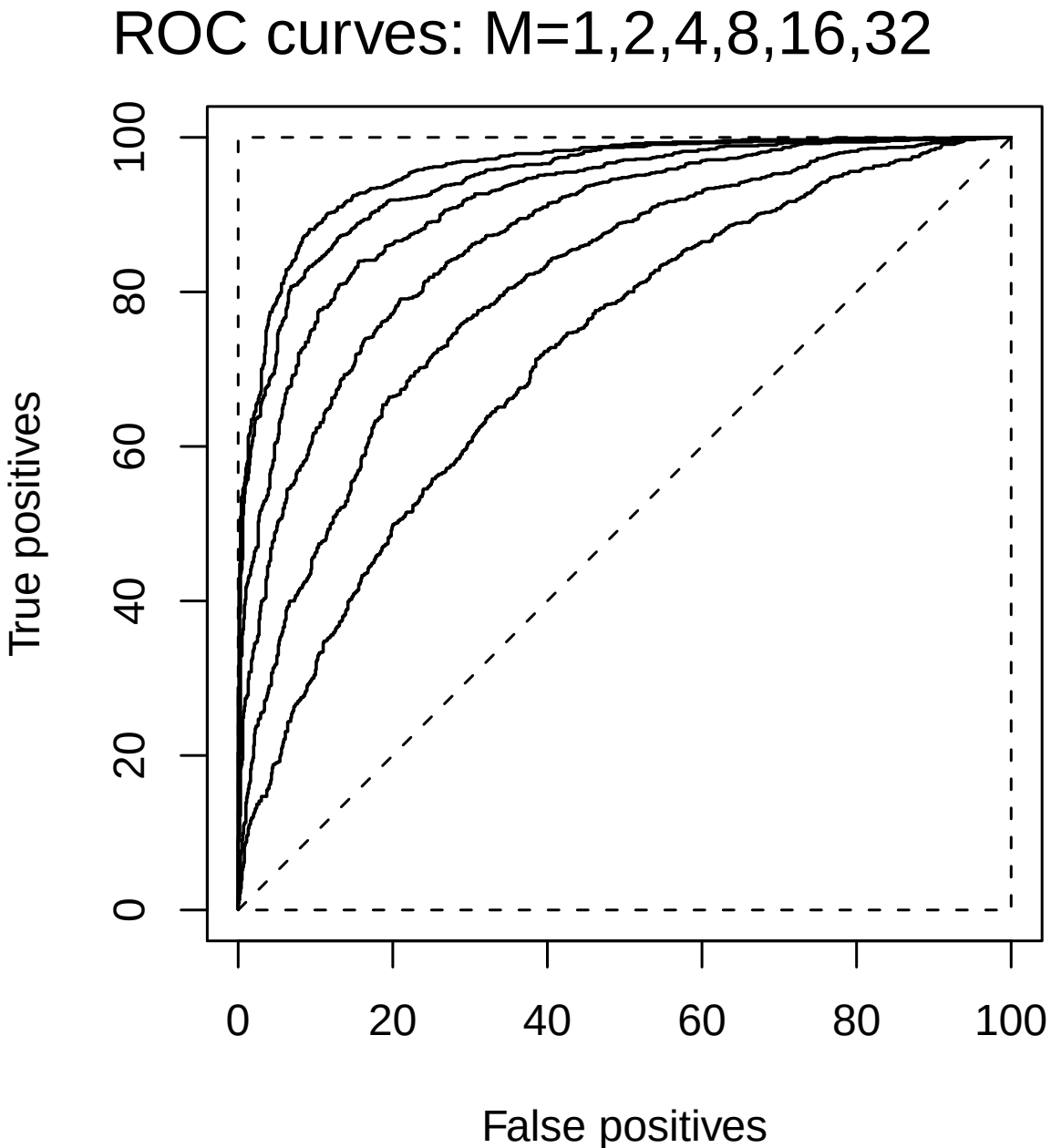


$n_x = n_y = 50, \quad d = 200, \quad k = 49,$

Null: $\|\delta\| = 0$

Alt: $\|\delta\| = 3$

The ROCs



Larger M has greater area under the curve:

M	AUC
1	71.9
2	80.6
4	87.1
8	91.4
16	94.3
32	95.7

Varying k

Lopes et al. prefer $k \approx (n_x + n_y - 2)/2$

That is not always optimal. But may be a good default.

For the previous scenario: small k do relatively poorly.

$32 \leqslant k \leqslant 56$ all gave AUC ≈ 0.95 with $M = 32$

Other scenarios

S. Emerson: advantage of averaging persists in other decay rates for eigenvalues of Σ

Using $\bar{T}^2$

The usual p-value is $\Pr(\bar{T}^2 \geqslant t^2)$ where t^2 is the observed value on our data.

We have no good approximation for this.

Even the moments of $\bar{T}^2$ involve difficult integrals over $\Theta \in V_{d,k}$, the Stiefel manifold, e.g.

$$\int_{\Theta \in V_{d,k}} \Theta(\Theta^{\mathsf{T}} S \Theta)^{-1} \Theta^{\mathsf{T}} \, \mathrm{d}\mathbf{U}(\Theta)$$

$$= (2\pi)^{-dk/2} \int_{Z \in \mathbb{R}^{d \times k}} Z(Z^{\mathsf{T}} S Z)^{-1} Z^{\mathsf{T}} \, e^{-\mathrm{tr}(Z^{\mathsf{T}} Z)/2} \, \mathrm{d}Z$$

Non-negative diagonal $S \in \mathbb{R}^{d \times d}$ with $n_x + n_y - 2$ positive entries

$\mathbf{U}(\Theta)$ is the uniform (Haar) measure.

Above is the first moment. Closed forms for first and second moments could lead to useful test statistics.

Permutation tests

There are $\binom{n_x+n_y}{n_x}$ ways to allocate n_x of the pooled observations

$$(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_x}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_y})$$

to the first sample (the $\boldsymbol{X}$'s). The re-allocated data:

$$(\boldsymbol{X}^*_1, \ldots, \boldsymbol{X}^*_{n_x}, \boldsymbol{Y}^*_1, \ldots, \boldsymbol{Y}^*_{n_y})$$

have statistic $\bar{T}^{*2}$. Then

$$p = \frac{\#\{\bar{T}^{2,*} \mid \bar{T}^{2,*} \geqslant \bar{T}^2\}}{\binom{n_x+n_y}{n_x}}$$

For $n_x = n_y = 50$ there are $\approx 10^{29}$ allocations (permutations).

Use a Monte Carlo sample of random permutations.

Justified in text by Lehmann & Romano (2005), "Testing Statistical Hypotheses"

Combination test might be a better name

Summary of problem one

If $d > n_x + n_y - 2$, then T^2 is not defined

Dempster (almost) used coordinate projections

Lopes et al. use one random projection from d to k dimensions

We find benefits from multiple projections

But have to use permutation tests for significance

Monte Carlo enters to compute the test statistic and then to judge its significance

Problem two

- How can we judge uncertainty in two-way and higher way data?
- We would like to use the bootstrap.
- But McCullagh (2000) proved this is impossible.

Solution

We will get a Monte Carlo method which mildly overestimates the sampling uncertainty.

But first

it is necessary to describe the bootstrap, as well as multi-way data.

Bootstrap sampling

Data are $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$

We compute $\widehat{T} = T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$

What is sampling uncertainty in $\widehat{T}$, e.g. $\mathrm{Var}(\widehat{T}) = \mathrm{Var}(\widehat{T} \mid F)$?

Combine two ideas

Monte Carlo	Sample from F to estimate $\mathrm{Var}(\widehat{T} \mid F)$ (but we don't know F).
Plug in	Use $\mathrm{Var}(\widehat{T} \mid \widehat{F})$, as if $F = \widehat{F}$, the empirical distribution*.

From Efron (1979).

$\exists$ extensive variations on the idea.

*Empirical distribution

$\widehat{F} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}$ puts probability $1/n$ on each sample observation.

$\delta_{\boldsymbol{x}}$ is a 'point mass' at $\boldsymbol{x}$

To sample $\boldsymbol{X} \sim \widehat{F}$, pick one of the original data points

Bootstrap pseudocode

Given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$

Resample the data B times

For $b = 1, \ldots, B$

  For $i = 1, \ldots, n$

    $i^* \sim \mathbf{U}\{1, \ldots, n\}$

    $\boldsymbol{X}_i^* = \boldsymbol{X}_{i^*}$

  $T_b^* = T(\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$

Compute summary

$$\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T_b^*$$

$$\widehat{\mathrm{Var}}(\widehat{T}) = \frac{1}{B-1}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2$$

First reactions

1) How could that ever work?

2) How could that ever fail?

The bootstrap works by a continuity argument

Plug-in:

Under conditions,

$\mathrm{Var}(\widehat{T} \mid F)$ is continuous in F, and $\widehat{F} \to F$.

So $\mathrm{Var}(\widehat{T} \mid \widehat{F}) \to \mathrm{Var}(\widehat{T} \mid F)$

Monte Carlo:

As $B \to \infty$, $\widehat{\mathrm{Var}}(\widehat{T} \mid \widehat{F}) \to \mathrm{Var}(\widehat{T} \mid \widehat{F})$

Fails when

Continuity conditions fail or when $\widehat{F}$ does not mimick F well enough

Bootstrapping the mean

The base case is for

$$T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \bar{\boldsymbol{X}} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

We have non-bootstrap methods for $\widehat{\mathrm{Var}}(\bar{\boldsymbol{X}})$ (!)

Correctness for the mean extends to more complicated statistics via Taylor approximations

Why we like it

No need to assume the Gaussian or any other distributional form

It is explainable to scientific colleagues

Some variants

The Bayesian bootstrap, Rubin (1981)

$$\widehat{F} = \frac{\sum_{i=1}^n W_i \, \delta_{\boldsymbol{X}_i}}{\sum_{i=1}^n W_i}$$

$$W_i \overset{\text{iid}}{\sim} \text{Exp}(1)$$

That is: place an independent random weight on each observation.

If $T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is the sample mean, then under bootstrapping

$$T^* = \frac{\sum_{i=1}^n W_i \, \boldsymbol{X}_i}{\sum_{i=1}^n W_i}$$

is a random ratio.

Weight condition

We need $\mathbb{E}(W_i) = 1$ and $\mathrm{Var}(W_i) = 1$

Then the Bayesian bootstrap is comparable to ordinary bootstrap

Can also use Poi(1) weights (Poisson distribution)

Used in machine learning Oza (2001)

Half sampling:

$$W_i = \begin{cases} 0 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/2 \end{cases}$$

Next

That was the bootstrap

We'll bootstrap 2-way data

One-way data

Patient	Height	Weight	Age	$\cdots$	Blood pressure
$\vdots$					
42	1.7	72	32	$\cdots$	141
43	2.1	97	42	$\cdots$	109
$\vdots$					

A usual data matrix has $\boldsymbol{X}_i$ in row i.

Data within a row are dependent,

e.g. height and weight of same person are correlated

as are blood pressure and age

Data in two different rows are independent.

patient 42 and patient 43

Two way data

patients	$\times$	medical interns	$\to$	blood pressure measurements
genes	$\times$	environments	$\to$	crop yields
students	$\times$	exam questions	$\to$	points

Features

Measurements on the same patient are correlated

Measurements by the same interns are also correlated

(hopefully that effect is smaller)

Neither rows nor columns are IID

Netflix data

Rating	Viewer 1	Viewer 2	Viewer 3	$\cdots$	Viewer C
Movie 1	4	4	1	$\cdots$	4
Movie 2	5	5	NA	$\cdots$	NA
Movie 3	3	3	NA	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Movie R	NA	5	3	$\cdots$	4

Ratings on same movie are correlated as are ratings by same person

Danny Deckchair (2003)

Most common rating was 4 stars

Netflix continued

17,770 movies, 480,189 customers, 100,000,000+ ratings,

Sample uncertainty

Ratings made on Tuesday came out a little lower than Sunday ratings.

Is it real or a sampling artifact?

Further problems

1) Missing data make the matrix very imbalanced

2) Unequal variances, e.g. some customers just give 1s or 5s

Facebook data

Alice (shares a URL) "Hey, check out `http://www.mcqmc2012.unsw.edu.au/`"

Bob (comments on it) "Thanks for sharing that, I learned a lot."

Data `url` $=$ `http://www.mcqmc2012.unsw.edu.au/`

`sharer` $=$ Alice

`commenter` $=$ Bob

`log length` $X = \log(41) \doteq 3.71$

Data size

$18{,}134{,}419$ comments by $8{,}078{,}531$ commenters on $2{,}085{,}639$ URLs

This is 3 way data: url $\times$ sharer $\times$ commenter

Of interest:

users' sharing and commenting behaviour,

e.g. who makes longer comments, U.S. or U.K. users?

Probably greater interest for ad clicks, linking and liking activity

Random effects model

$$X_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \qquad i = 1, \ldots, R \qquad j = 1, \ldots, C$$

$a_i \sim \mathcal{N}(0, \sigma_A^2)$ e.g. patients

$b_j \sim \mathcal{N}(0, \sigma_B^2)$ e.g. interns

$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2)$

simplest model for two-way data

Used in agriculture

Studied for decades

$\hat{\mu}$ is $\bar{X}_{\bullet\bullet}$

No bootstrap exists for $V(\hat{\mu})$

None can exist $\cdots$

$\cdots$ McCullagh (2000)

We can't even bootstrap a balanced $\bar{X}$!

He rules out resampling, permuting or any "monoid" operations on rows and columns.

What about classical approaches?

prime reference:

"Variance Components" by

Searle, Casella, McCulloch (1992)

- Excellent for balanced Gaussian data
- Unbalance $\Longrightarrow$ invert large matrices
- Emphasis on homogeneous variances

McCullagh (2000)

For $\hat{\mu} = \bar{X}_{\bullet\bullet} = \dfrac{1}{R}\dfrac{1}{C}\displaystyle\sum_{i=1}^{R}\sum_{j=1}^{C} X_{ij}$

Naive Resample from $N = RC$ values

Product Resample R rows and resample C columns (indep)

$$V_{\mathrm{RE}}(\hat{\mu}) = \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{\sigma_E^2}{RC} \qquad \text{true var}$$

$$\mathbb{E}_{\mathrm{RE}}(\widehat{V}_{\mathsf{Naiv}}(\hat{\mu})) \doteq \left(\sigma_A^2 + \sigma_B^2 + \sigma_E^2\right)\frac{1}{RC} \qquad \text{way too small}$$

$$\mathbb{E}_{\mathrm{RE}}(\widehat{V}_{\mathsf{Prod}}(\hat{\mu})) \doteq \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{3\sigma_E^2}{RC} \qquad \text{not so bad}$$

Naive resampling seriously flawed, product resampling is close

Notation

Index j takes values $i_j = 1, 2, 3, \ldots$

Observation $\boldsymbol{i}$ has multi-index $\boldsymbol{i} = (i_1, \ldots, i_r) \in \{1, 2, \ldots\}^r$

Random $X_{\boldsymbol{i}} \in \mathbb{R}^d$ with $Z_{\boldsymbol{i}} = \begin{cases} 1 & X_{\boldsymbol{i}} \text{ known} \\ 0 & X_{\boldsymbol{i}} \text{ missing.} \end{cases}$

Sample size

$$0 < N \equiv \sum_{\boldsymbol{i} \in \mathbb{N}^r} Z_{\boldsymbol{i}} < \infty$$

Sample mean

$$\bar{X} = \frac{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} X_{\boldsymbol{i}}}{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}}}$$

We want to estimate $\mathrm{Var}(\bar{X})$

treating $Z_{\boldsymbol{i}}$ as fixed

r-fold product bootstrap

$$\bar{X}^* = \frac{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} W_{\boldsymbol{i}} X_{\boldsymbol{i}}}{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} W_{\boldsymbol{i}}} \quad \text{where}$$

$$W_{\boldsymbol{i}} = \prod_{j=1}^{r} W_{j,i_j}$$

$$\mathbb{E}(W_{j,i_j}) = 1$$

$$\mathrm{Var}(W_{j,i_j}) = 1$$

From replications $\bar{X}^{*1}, \ldots, \bar{X}^{*B}$

$$\widehat{\mathrm{Var}}(\bar{X}) = \frac{1}{B-1}\sum_{b=1}^{B}(X^{*b} - \bar{X}^*)^2.$$

It is operationally much easier to have independent W_{j,i_j}

(as in the Bayesian bootstrap).

Data for URLs might be scattered over several continents.

Then keeping $\sum_i W_{j,i_j} = N$ is awkward.

r-fold random effects

We study $\widehat{\mathrm{Var}}(\bar{X})$ under this model:

$$X_i = \mu + \sum_{\substack{u \subseteq \{1,\dots,r\} \\ u \neq \varnothing}} \varepsilon_{\boldsymbol{i}_u, u}$$

If $u = (j_1, \dots, j_k)$ then $\boldsymbol{i}_u = (i_{j_1}, \dots, i_{j_k})$

Moments

$$\mathbb{E}(\varepsilon_{\boldsymbol{i}_u, u}) = 0 \quad \mathrm{Var}(\varepsilon_{\boldsymbol{i}_u, u}) = \sigma_u^2$$

$$\mathrm{Cov}(\varepsilon_{\boldsymbol{i}_u, u}, \varepsilon_{\boldsymbol{i}'_{u'}, u'}) = 0 \quad \text{if } u \neq u' \quad \text{or} \quad \boldsymbol{i}_u \neq \boldsymbol{i}'_{u'}$$

σ_u^2 can be allowed to depend on $\boldsymbol{i}_u$

Variance

$$\begin{aligned}
\mathrm{Var}_{\mathrm{RE}}(\hat{\mu}) &= \frac{1}{N^2} \sum_{u \neq \varnothing} \sum_{u' \neq \varnothing} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{i}'} Z_{\boldsymbol{i}} Z_{\boldsymbol{i}'} \mathrm{Cov}(\varepsilon_{\boldsymbol{i},u}, \varepsilon_{\boldsymbol{i}',u'}) \\
&= \frac{1}{N^2} \sum_{u \neq \varnothing} \left(\sum_{\boldsymbol{i}} \sum_{\boldsymbol{i}'} Z_{\boldsymbol{i}} Z_{\boldsymbol{i}'} 1_{\boldsymbol{i}_u = \boldsymbol{i}'_u} \right) \sigma_u^2 \\
&= \frac{1}{N} \sum_{u \neq \varnothing} \nu_u \sigma_u^2,
\end{aligned}$$

for gain coefficients,

$$\nu_u = \frac{1}{N} \sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} N_{\boldsymbol{i},u}, \quad \text{where}$$

$$N_{\boldsymbol{i},u} = \sum_{\boldsymbol{i}'} Z_{\boldsymbol{i}'} 1_{\boldsymbol{i}_u = \boldsymbol{i}'_u}.$$

Running examples

$$V_{\mathrm{RE}}(\hat{\mu}) \equiv \frac{1}{N} \sum_{u \neq \varnothing} \nu_u \sigma_u^2$$

$$\doteq \frac{1}{N} \left(56{,}200\, \sigma^2_{\text{movies}} + 646\, \sigma^2_{\text{viewers}} + \sigma^2_{\text{interaction}} \right) \quad \text{(for Netflix)}$$

For Facebook

$$\nu_{\text{sh}} \doteq 17.71, \quad \nu_{\text{com}} \doteq 7.71, \quad \nu_{\text{url}} \doteq 26{,}854.92 \quad !$$

$$\nu_{\text{sh,com}} \doteq 5.92, \quad \nu_{\text{sh,url}} \doteq 12.91, \quad \nu_{\text{com,url}} \doteq 5.19, \quad \text{and}$$

$$\nu_{\text{sh,com,url}} \doteq 4.88.$$

$$\nu_{\text{url}} \geqslant 26{,}000$$

Variances

For gain coefficients ν_u

$$\mathrm{Var}_{\mathrm{RE}}(\bar{X}) = \frac{1}{N} \sum_{\substack{u \subseteq \{1,\dots,r\} \\ u \neq \varnothing}} \nu_u \sigma_u^2$$

Similarly for γ_u (defined later):

$$\mathbb{E}_{\mathrm{RE}}(\widehat{\mathrm{Var}}_{\mathsf{Prod}}(\bar{X}^*)) \doteq \frac{1}{N} \sum_{\substack{u \subseteq \{1,\dots,r\} \\ u \neq \varnothing}} \gamma_u \sigma_u^2$$

And:

$$\mathbb{E}_{\mathrm{RE}}(\widehat{\mathrm{Var}}_{\mathsf{Naiv}}(\bar{X}^*)) \doteq \frac{1}{N} \sum_{\substack{u \subseteq \{1,\dots,r\} \\ u \neq \varnothing}} \sigma_u^2$$

Typically $1 \ll \nu_u \ll N$ for $u \neq \{1,\dots,r\} \implies$ Naive bootstrap unreliable

We want $\gamma_u = \nu_u$. We'll get $\gamma_u \geqslant \nu_u$.

The case $r = 2$

O (2007)

Independent bootstrap of rows and columns

Still get $\mathbb{E}_{\mathrm{RE}}(\widehat{V}_{\mathsf{Prod}}(\hat{\mu})) \doteq V_{\mathrm{RE}}(\hat{\mu})$, i.e.

Still get $\approx 1 \times$ the main effect contribution

$\approx 3 \times$ the interaction contribution

Sunday vs. Tuesday edge of 0.02 stars is real (about 8 standard errors)

New here

1) Arbitrary order $r \geqslant 2$

2) Independent product weights (vs resampling)

Product bootstrap

$$\hat{\mu}^* = \frac{\sum_i Z_i W_i X_i}{\sum_i Z_i W_i} \equiv \frac{T^*}{N^*} \qquad \text{(ratio estimator)}$$

$$V_{\text{Prod}}(\hat{\mu}^*) \approx \widetilde{V}_{\text{Prod}}(\hat{\mu}^*) \equiv \frac{1}{N^2}\mathbb{E}_{\text{Prod}}\big((T^* - \hat{\mu}N^*)^2\big)$$

Main result

$$\mathbb{E}_{\text{RE}}(\widetilde{V}_{\text{Prod}}(\hat{\mu}^*)) = \frac{1}{N}\sum_{u\neq\varnothing} \gamma_u \sigma_u^2$$

where $\gamma_u \approx \nu_u$ if $|u| = 1$, (i.e. cardinality 1)

otherwise small $\gamma_u/\nu_u > 1$

Exact formula depends on

index pairs $\boldsymbol{i}$, $\boldsymbol{i}'$ that match in set u:

$$\sum_{\boldsymbol{i}\in\mathbb{N}^r}\sum_{\boldsymbol{i}'\in\mathbb{N}^r} Z_{\boldsymbol{i}} Z_{\boldsymbol{i}'} 1_{\boldsymbol{i}_u=\boldsymbol{i}'_u}$$

index pairs $\boldsymbol{i}$, $\boldsymbol{i}'$ that match in **precisely** k places

index triples $\boldsymbol{i}$, $\boldsymbol{i}'$, $\boldsymbol{i}''$ where $\boldsymbol{i}$ matches $\boldsymbol{i}'$ in set u and matches $\boldsymbol{i}''$ in precisely k places

Duplication indices

(level dup) $\epsilon = \dfrac{\text{greatest item popularity}}{N}$

(variable dup) $\eta = \max_{\varnothing \subsetneq u \subsetneq v} \dfrac{\nu_v}{\nu_u}$

Examples

	ϵ	η
Netflix	$\frac{232{,}944}{100{,}480{,}507} \doteq 0.00232$	$\frac{1}{646} \doteq 0.00155$
	Miss Congeniality	$\nu_{\text{interaction}}/\nu_{\text{movies}}$
Facebook	$\frac{686{,}990}{18{,}134{,}419} \doteq 0.0379$	$\frac{4.88}{5.19} \doteq 0.94$
	a popular URL	$\nu_{\text{sh,com,url}}/\nu_{\text{com,url}}$

η is not small for the Facebook data

bootstrap variances will be somewhat more conservative

Approximations

Theorem 1. *In the homogeneous random effects model, the product weight bootstrap with* $V(W_{j,i_j}) = \tau^2 = 1$, *satisfies*

$$\gamma_u = \nu_u[2^{|u|} - 1 + \Theta_u\epsilon] + \sum_{v \supsetneq u} 2^{|v|}\nu_v,$$

where $|\Theta_u| \leqslant 2^{r+1} - 2$.

Proof. O & Eckles (2011), who consider general τ^2. □

For small ϵ and r (i.e. $2^r\epsilon \ll 1$)

$$\gamma_u \approx (2^{|u|} - 1)\nu_u + \sum_{v \supsetneq u} 2^{|v|}\nu_v$$

If also $\eta \ll 1$

$$\gamma_u \approx (2^{|u|} - 1)\nu_u$$

Some specific approximations

For $r = 2$

$$\gamma_{\{j\}} = \nu_{\{j\}}(1 + \Theta_j \epsilon) + 2 \quad j = 1, 2$$
$$\gamma_{\{1,2\}} = \nu_{\{1,2\}}(3 + \Theta_{\{1,2\}} \epsilon), \quad \text{where}$$
$$|\Theta_u| \leqslant 6.$$

For $r = 3$

$$\gamma_{\{1\}} \approx \nu_{\{1\}} + 4\nu_{\{1,2\}} + 4\nu_{\{1,3\}} + 8$$
$$\gamma_{\{1,2\}} \approx 3\nu_{\{1,2\}} + 8$$
$$\gamma_{\{1,2,3\}} \approx 7.$$

If $0 < m \leqslant \min_u \sigma_u^2 \leqslant \max_u \sigma_u^2 \leqslant M < \infty$ then

$$\frac{\mathbb{E}_{\mathrm{RE}}(\widetilde{V}_{\mathsf{Prod}}(\hat{\mu}^*))}{V_{\mathrm{RE}}(\hat{\mu})} = 1 + O(\eta + \epsilon).$$

Facebook loquacity

For each commenter, url and sharer, we obtain:

$X = \log(\#\text{char in comment})$ as well as,

country $c \in \{\text{US}, \text{UK}\}$ of commenter, and

mode $m \in \{\text{web}, \text{mobile}\}$ of commenter.

Now let

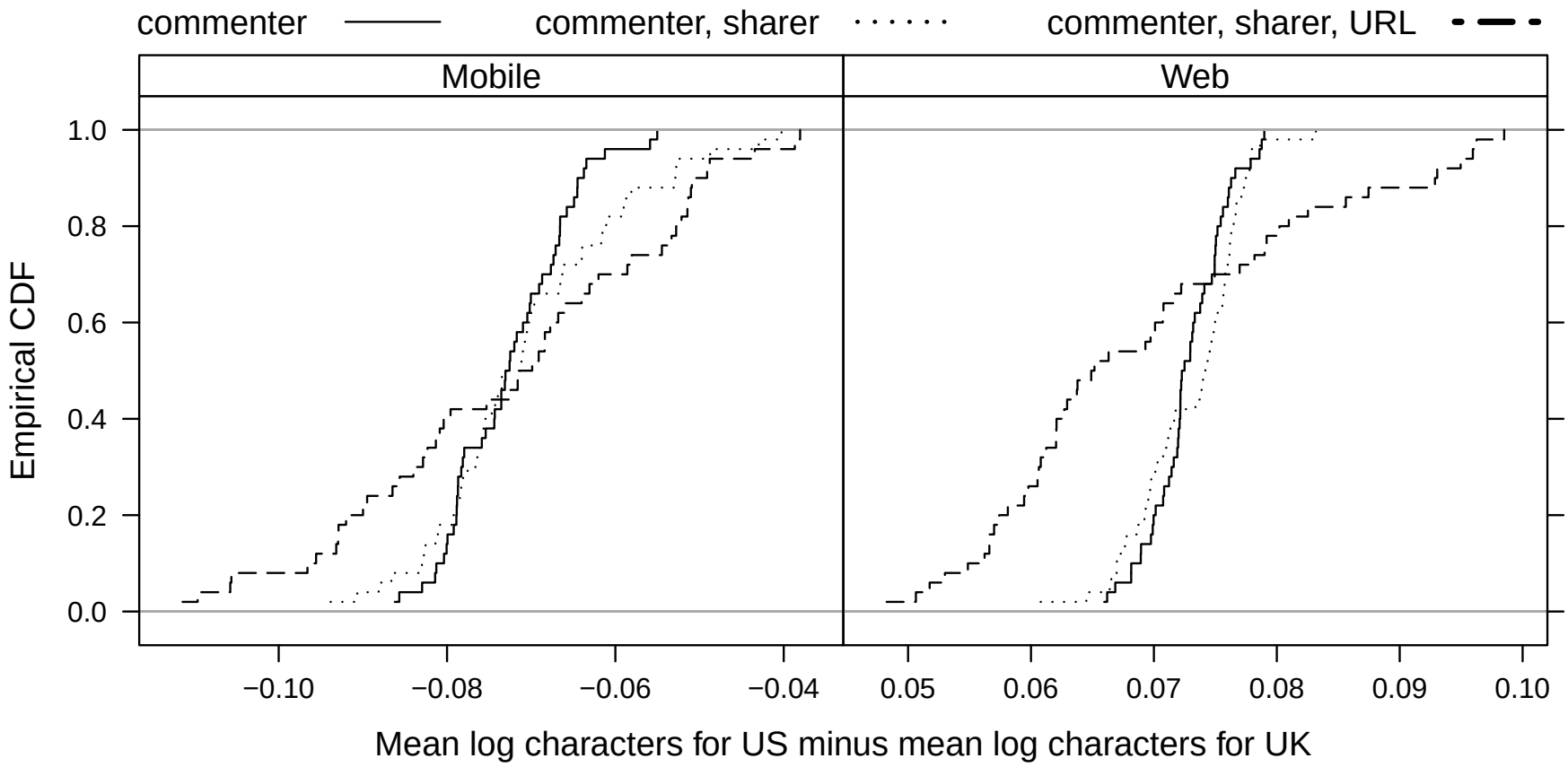
$$\hat{\mu}_{cm} = \frac{\sum_i Z_i X_i 1_{\text{country}=c} 1_{\text{mode}=m}}{\sum_i Z_i 1_{\text{country}=c} 1_{\text{mode}=m}}$$

We see small differences

	US	UK
web	3.62	3.55
mobile	3.50	3.57

but they're larger than sample fluctuations

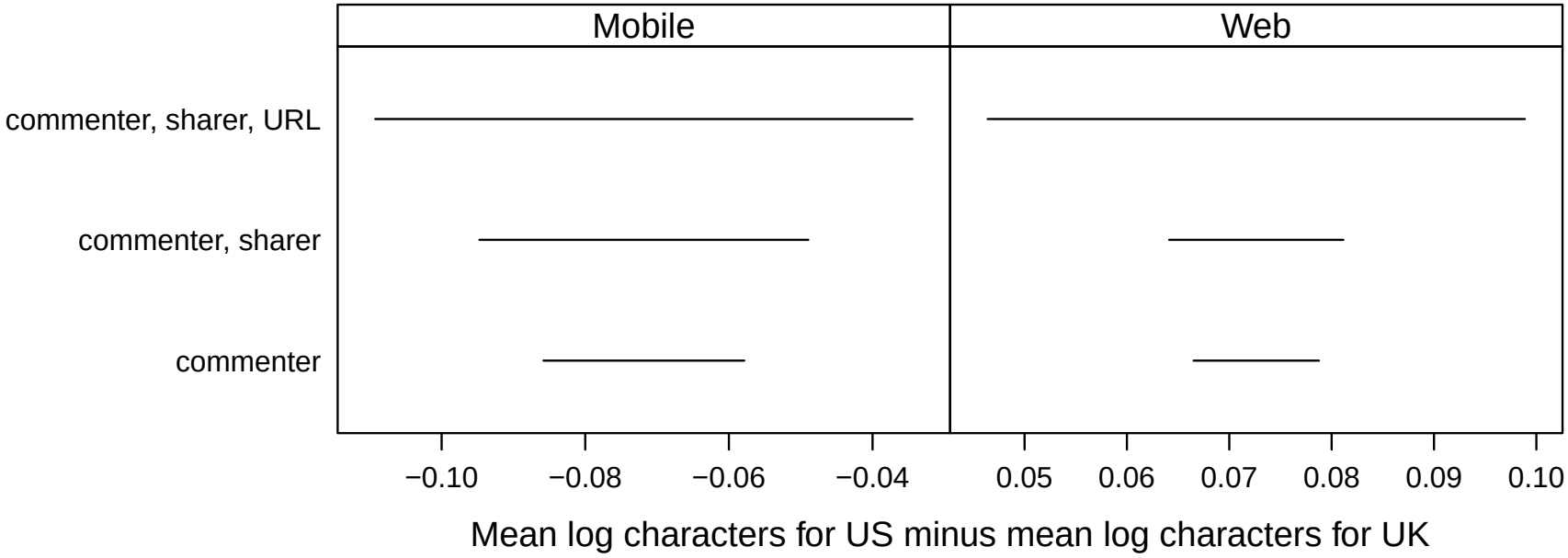
Loquacity ECDFs



ECDF over 50 bootstraps of $\hat{\mu}_{\text{US}m} - \hat{\mu}_{\text{UK}m}$

Reweighting one, two, or three ways

Loquacity confidence intervals



Central 95% confidence intervals from 50 bootstraps of $\hat{\mu}_{\text{US}m} - \hat{\mu}_{\text{UK}m}$

Reweighting one, two, or three ways

Summary of problem two

Crossed random effects data require special care

No correct bootstrap exists

Product sampling is at least conservative, mildly over-estimating the variance

It works also with unequal variances

what remains to do

better seeding

extensions to more complicated analyses

Conclusion

Statistics and machine learning are still finding new ways to consume Monte Carlo ideas

In addition to MCMC there are:

- permutations
- rotations
- projections
- subsampling
- reweighting

Thanks

- Collaborators: Dean Eckles and Sarah Emerson
- NSF DMS-0906056
- Data: Netflix and Facebook
- UNSW
- Organizers: Ian Sloan, Frances Kuo, Josef Dick and Gareth Peters

Missingness

The $\approx 100{,}000{,}000$ movie ratings we see are only 1% of $R \times C$.

Those we see are probably biased towards higher values. How do we adjust?

Ans: we can't, because nobody knows the bias function.

It would not be reasonable to expect a bootstrap method to fix missing data bias. For instance how could one sampling method fix both positive bias and negative bias?

Given an adjustment algorithm (based on assumptions or knowledge from outside the given data) we might be able to bootstrap its predictions.