

Bootstrapping r -Fold Tensor Data

OR:

how to get confidence intervals for click-through rates, despite horrendous correlation structures you'd rather not have to model or even think about*

Art B. Owen

Stanford University

Dean Eckles

Facebook Inc.

*Apologies to Jay Ward

Dean Eckles



Stanford Communications PhD, 2012

Took tons of stat classes

Deep understanding of causality issues

Now at Facebook

The IID bootstrap

- data are IID F
- we resample IID from the empirical distribution \hat{F}
- getting variance estimates and confidence intervals

We like it because

- face value validity (or at least explainability)
- deep theory for \bar{X} vs. $\mathbb{E}(X)$
- extensions to more general statistics

Bootstrap (and cross-validation) let us use very mild assumptions:

- 1) IID data, and
- 2) non-pathological moments.

IID data vectors

| | Variable 1 | ... | Variable C |
|--------|------------|-----|------------|
| Case 1 | | | |
| ⋮ | | | |
| Case R | | | |

- 1) Variables are named entities:
 - E.g. pressure, volume, income ...
 - They persist
- 2) Cases are anonymous replicates
 - Sampled IID from some F
 - Of no inherent interest
 - We'd rather just know F

For IID data ...

... we only care about cases because they show relationships among variables.

Two-way data

| Rating | Viewer 1 | Viewer 2 | Viewer 3 | ... | Viewer C |
|---------|----------|----------|----------|-----|----------|
| Movie 1 | 4 | 4 | 1 | ... | 4 |
| Movie 2 | 5 | 5 | NA | ... | NA |
| Movie 3 | 3 | 3 | NA | ... | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Movie R | NA | 5 | 3 | ... | 4 |

More examples of two-way data:

genes × environments → crop yields

terms × documents → counts

candidate × interviewer → rating

nodes × more nodes → labeled edges

Tensor data

r -way data, i.e. an r -tuple of named entities. For example:

Suppose that customer U

comes from computer (machine) M

enters query Q

reads review R

buys book B

with credit card book C

ships to address A

Then Amazon's logs get (U, M, Q, R, B, C, A) among other variables (such as price paid).

While $r = 2$ is most common, $r > 2$ arises frequently.

Tuples

| | Movie | Viewer | Rating |
|--------|-------|--------|--------|
| Case 1 | 1 | 1 | 4 |
| Case 2 | 1 | 2 | 4 |
| Case 3 | 2 | 1 | 5 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| Case N | R | C | 4 |

- Now cases are anonymous
- We don't store the NAs
- 2 categorical variables with lots of levels
- Not independent:
 - Cases 1 & 2 share a movie
 - Cases 1 & 3 share a viewer

How should we bootstrap and cross-validate data like this?

What about $r > 2$?

Maybe large N means no meaningful uncertainty.

Click through rates

$$\text{CTR} = \frac{\sum_{\text{obs} \in \text{logs}} Z_{\text{obs}} \times \text{Click}(\text{obs})}{\sum_{\text{obs} \in \text{logs}} Z_{\text{obs}} \times \text{Impression}(\text{obs})}$$

Where

$Z_{\text{obs}} = 1$ for campaign of interest 0 else

Observation obs indexed by URL, IP, ad ID

I.e., there may be multiple factors correlating these binary events

Similarly

Other large scale averages over logs may be based on correlated data.

Random effects model

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

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

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

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

Used in agriculture

Studied for decades

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

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

None can exist . . .

. . . McCullagh (2000)

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

Simulated dialog



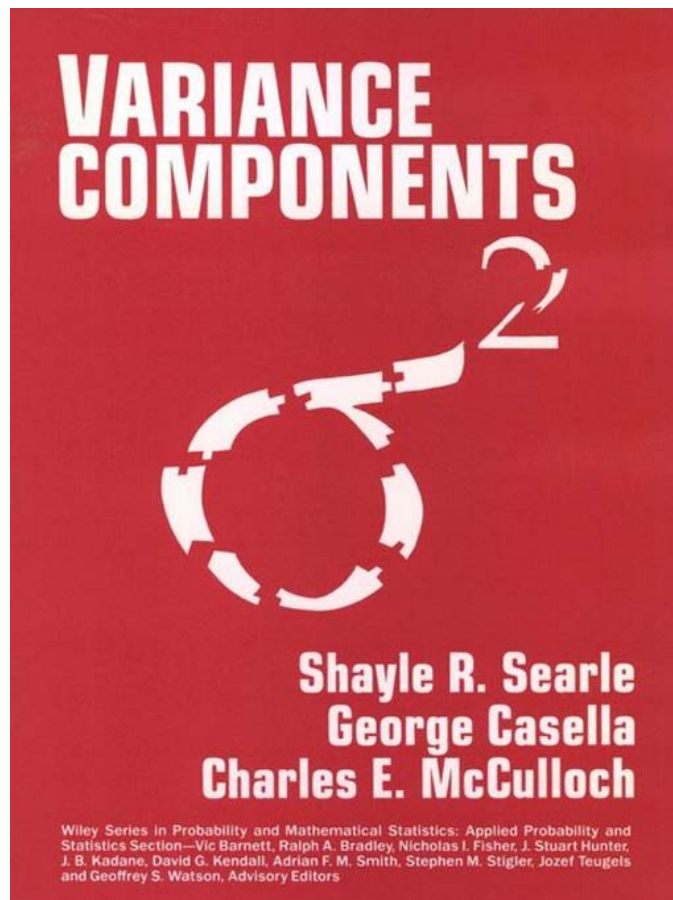
Nobody can bootstrap the crossed random effects model.



Not even me. I'm comfortable with that.

What about classical approaches?

prime reference:



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

McCullagh (2000)

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

Boot-I Resample from $N = RC$ values

Boot-II Resample R rows and resample C columns (indep)

$$V(\hat{\mu}) = \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{\sigma_E^2}{RC} \quad \text{true var}$$

$$\mathbb{E}(\hat{V}_I(\hat{\mu})) \doteq \left(\sigma_A^2 + \sigma_B^2 + \sigma_E^2 \right) \frac{1}{RC} \quad \text{way too small}$$

$$\mathbb{E}(\hat{V}_{II}(\hat{\mu})) \doteq \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{3\sigma_E^2}{RC} \quad \text{not so bad}$$

Boot-I is seriously flawed, Boot-II is close

The case $r = 2$

O (2007)

Independent bootstrap of rows and columns

Allows for missing data \dots but conditions on pattern of observed data

Allows non-homogeneous $V(a_i)$, $V(b_j)$ and $V(\varepsilon_{ij})$

Still get $\mathbb{E}(\widehat{V}_B(\hat{\mu})) \doteq V(\hat{\mu})$, i.e.

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

$\approx 3 \times$ the interaction contribution

On Netflix data ... naive bootstrap can under-estimate variance by 56,200 fold

Sunday vs. Tuesday edge of 0.02 stars is real

mimics pigeonhole model of Cornfield & Tukey (1956)

Fine print:

uniform bounds on variances, and

no row/column has more than ϵ of the data

Goals

We would like to get an approximate bootstrap for arbitrary data patterns with $r \geq 2$.

We focus on getting the variance approximately right.

| Challenge | Today |
|--|-------|
| What happens to that 3 for $r > 2$? | ● |
| There are many missing data values. | ● |
| Missingness might be informative. | ● |
| The entities might have unequal variances. | ● |
| We might want a little more than \bar{X} . | ● |
| We might want a lot more than \bar{X} . | ● |

Illustrative data sets

Netflix

$N = 100,480,507$ ratings,
by 480,189 customers,
on 17,770 movies
 X is 1 to 5 stars
used in famous contest

Facebook

18,134,419 comments
by 8,078,531 commenters
on 2,085,639 URLs
shared by 3,904,715 sharers
 X is $\log(\# \text{ chars in comment})$

Example

Alice (shares a URL) “Hey, check out `http://stat.stanford.edu`”

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

Data `url = http://stat.stanford.edu`

`sharer = Alice`

`commenter = Bob`

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

Random effects: r -way case

| | |
|-----------|---|
| Index | $\mathbf{i} = (i_1, i_2, \dots, i_r) \in \{1, 2, 3, \dots\}^r$ |
| Sub-index | $\mathbf{i}_u = (i_{j_1}, \dots, i_{j_L}) \quad u = \{j_1, \dots, j_L\} \subseteq \{1, 2, \dots, r\}$ |
| Data | $X_{\mathbf{i}} \in \mathbb{R}^d$ short for X_{i_1, i_2, \dots, i_r} use $d = 1$ |
| Presence | $Z_{\mathbf{i}} \in \{0, 1\}$ |

We model a random effect for each non-empty $u \subseteq \{1, 2, \dots, r\}$.

$$X_{\mathbf{i}} = \mu + \sum_{u \neq \emptyset} \varepsilon_{\mathbf{i}, u}$$

$$\mathbb{E}(\varepsilon_{\mathbf{i}, u}) = 0$$

$$\text{Cov}(\varepsilon_{\mathbf{i}, u}, \varepsilon_{\mathbf{i}', u'}) = \sigma_{\mathbf{i}, u}^2 \mathbf{1}_{u=u'} \mathbf{1}_{\mathbf{i}_u = \mathbf{i}'_u}$$

Homogeneous special case

$$\sigma_{\mathbf{i}, u}^2 \equiv \sigma_u^2 \quad \forall \mathbf{i} \in \mathbb{N}^r \quad \forall u \subseteq \{1, \dots, r\}$$

The product reweighted bootstrap

$$\hat{\mu} = \frac{\sum_i Z_i X_i}{\sum_i Z_i} \quad \text{and} \quad \hat{\mu}^* = \frac{\sum_i Z_i W_i X_i}{\sum_i Z_i W_i}$$

Our reweighting

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

$$\mathbb{E}(W_{j,i_j}) = 1 \quad \text{all indep.}$$

$$V(W_{j,i_j}) = \tau^2 \quad \text{usually } \tau^2 = 1$$

Resampling vs. reweighing

| Bootstrap | Distribution of W_{j,i_j} | Reference |
|---------------|--|----------------------------------|
| Original | Multinomial($N_j; 1/N_j, \dots, 1/N_j$) | Efron (1979) |
| Bayesian | $W_{j,i_j} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ | Rubin (1981) |
| Poisson | $W_{j,i_j} \stackrel{\text{iid}}{\sim} \text{Poi}(1)$ | Oza (2001) Lee & Clyde (2004) |
| Half sampling | $W_{j,i_j} \stackrel{\text{iid}}{\sim} \mathbf{U}\{0, 2\}$ | McCarthy (1969) |

Independent weights are much simpler to analyze and implement:
data may be spread over servers, countries, continents.

Joys of half-sampling

$$W_i = \prod_{j=1}^d W_{j,i_j} \quad \text{where} \quad W_{j,i_j} \stackrel{\text{iid}}{\sim} \mathbf{U}\{0, 2\}$$

Original context was stratified sampling, $n = 2$ per stratum.

As a bootstrap

- All data get integer weights
- All nonzero weights are equal
- Has minimal kurtosis subject to mean = variance = 1.

Each bootstrap computation is the same as the original one but with about $2^{-r} N$ observations.

True variance (homog. case)

Recall

$$X_i = \mu + \sum_{u \neq \emptyset} \varepsilon_{i,u}$$

$$V(\varepsilon_{i,u}) = \sigma_u^2, \quad \text{and let}$$

$$N \equiv \sum_i Z_i.$$

Then

$$\begin{aligned} V_{\text{RE}}(\hat{\mu}) &= \frac{1}{N^2} \sum_{u \neq \emptyset} \sum_i \sum_{i'} 1_{i_u=i'_u} \sigma_u^2 \equiv \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma_u^2 \\ &\doteq \frac{1}{N} \left(56,200 \sigma_{\text{movies}}^2 + 646 \sigma_{\text{viewers}}^2 + \sigma_{\text{interaction}}^2 \right) \quad (\text{for Netflix}) \end{aligned}$$

Our examples

$$\begin{aligned}
 V_{\text{RE}}(\hat{\mu}) &\equiv \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma_u^2 \\
 &\doteq \frac{1}{N} \left(56,200 \sigma_{\text{movies}}^2 + 646 \sigma_{\text{viewers}}^2 + \sigma_{\text{interaction}}^2 \right) \quad (\text{for Netflix})
 \end{aligned}$$

For Facebook

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

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

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

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

Naive bootstrap (homog. case)

$$V_{\text{RE}}(\hat{\mu}) = \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma_u^2$$

$$\mathbb{E}_{\text{RE}}(V_{\text{NB}}(\hat{\mu}^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \left(1 - \frac{\nu_u}{N}\right) \sigma_u^2 \quad \text{O and Eckles (2011)}$$

Typically $1 \ll \nu_u \ll N$ for $u \neq \{1, \dots, r\}$

Note: $V_{\text{NB}}(\hat{\mu}^*)$ is what the bootstrap settles down to in $B \rightarrow \infty$ resamplings.

Product bootstrap

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

$$V_{PW}(\hat{\mu}^*) \approx \tilde{V}_{PW}(\hat{\mu}^*) \equiv \frac{1}{N^2} \mathbb{E}_{PW}((T^* - \hat{\mu} N^*)^2) \quad (\text{as } B \rightarrow \infty)$$

The delta method is reliable for large data
(Chamandy, Muralidharan, Najmi (2011))

Main result

$$\mathbb{E}_{RE}(\tilde{V}_{PW}(\hat{\mu}^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \gamma_u \sigma_u^2$$

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

Exact formula depends on

| Notation | Definition | Meaning |
|-----------|---------------------------------|----------------------|
| $N_{i,u}$ | $\sum_{i'} Z_{i'} 1_{i_u=i'_u}$ | Match i in u |
| ν_u | $N^{-1} \sum_i Z_i N_{i,u}$ | Avg # matches on u |

Exact result $\gamma_u = \sum_{k=0}^r (1 + \tau^2)^k (\nu_{k,u} - 2\tilde{\nu}_{k,u} + \rho_k \nu_u)$ non-asymptotic

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

Fine print from article

- $\nu_{k,u}$ depends on the number of i, i' pairs that match in precisely k indices, including those in u .
- $\tilde{\nu}_{k,j}$ depends on the number of triples i, i', i'' where i matches i' in the set u and matches i'' in precisely k indices.

Approximations

The exact formula captures some bad cases. We can often simplify them.

Extreme level duplication

e.g. $N/2$ obs in row 1 and $N/2$ obs in col 1.

effective sample size is about one or two

Formulas simplify if level duplication is not extreme.

Variable duplication

Almost every record that matches on some variables matches on a superset of those variables

e.g. match name and phone number \Rightarrow usually match fax number

match age and zip code \nRightarrow match occupation

Duplication indices

$$\text{(level dup)} \quad \epsilon = \max_i \max_{u \neq \emptyset} \frac{N_{i,u}}{N} = \max_i \max_{1 \leq j \leq r} \frac{N_{i,\{j\}}}{N}$$

$$\text{(variable dup)} \quad \eta = \max_{\emptyset \subsetneq u \subsetneq v} \frac{\nu_v}{\nu_u} = \max_{\emptyset \subsetneq u \subsetneq v} \frac{\nu_v}{\nu_u}$$

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

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 $r = 2$

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

For $r = 3$

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

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

$$\frac{\mathbb{E}_{\text{RE}}(\tilde{V}_{\text{PW}}(\hat{\mu}^*))}{V_{\text{RE}}(\hat{\mu})} = 1 + O(\eta + \epsilon).$$

Facebook loquacity

For each commenter, url and sharer, we obtain:

$X = \log(\#char \text{ in comment})$ as well as,
country $c \in \{US, UK\}$ of commenter, and
mode $m \in \{web, mobile\}$ of commenter.

Now let

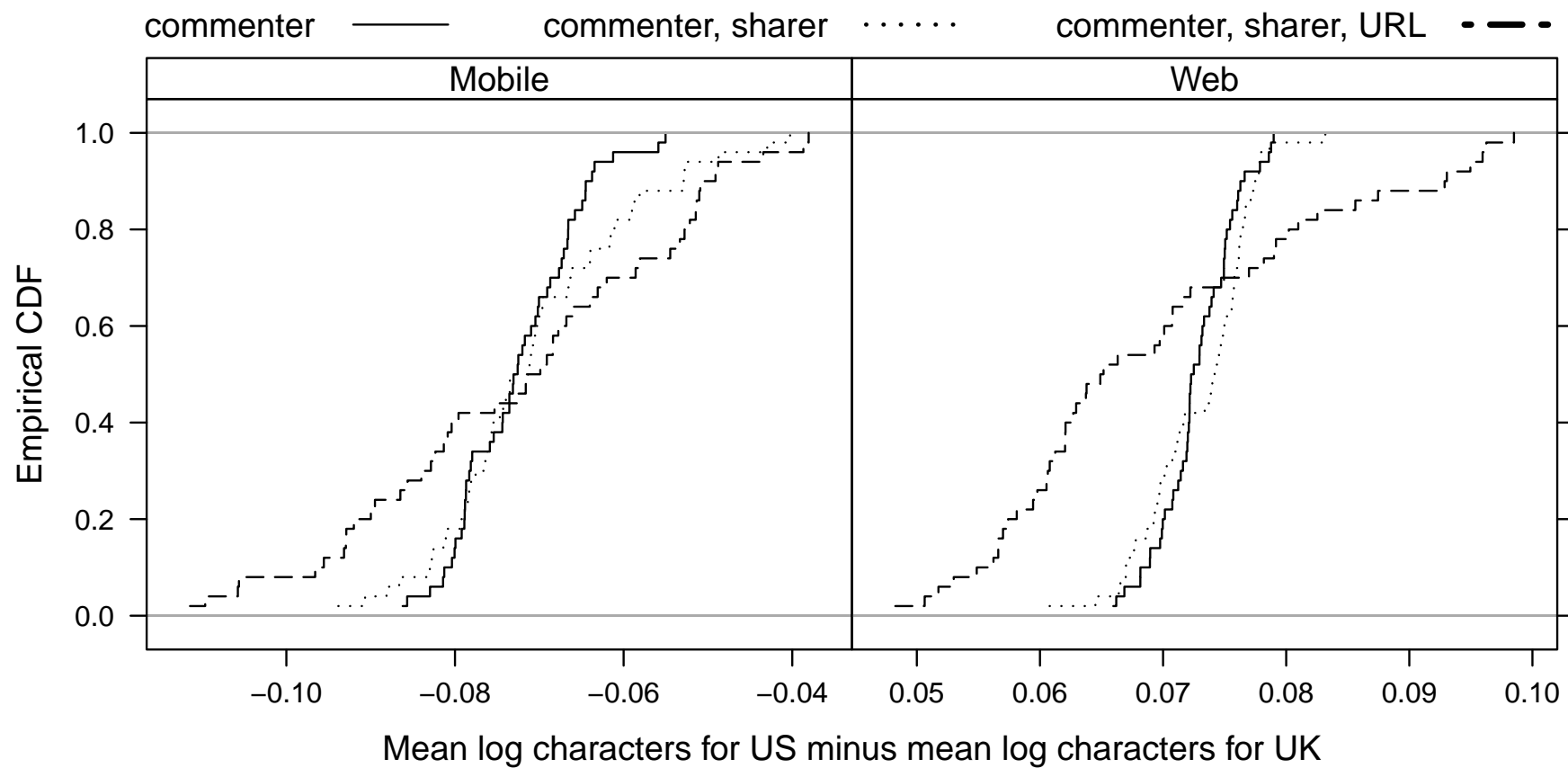
$$\hat{\mu}_{cm} = \frac{\sum_i Z_i X_i 1_{\text{country}=c} 1_{\text{mode}=m}}{\sum_i Z_i 1_{\text{country}=c} 1_{\text{mode}=m}}$$

We see small differences

| | US | UK |
|--------|------|------|
| web | 3.62 | 3.55 |
| mobile | 3.50 | 3.57 |

but they're larger than sample fluctuations

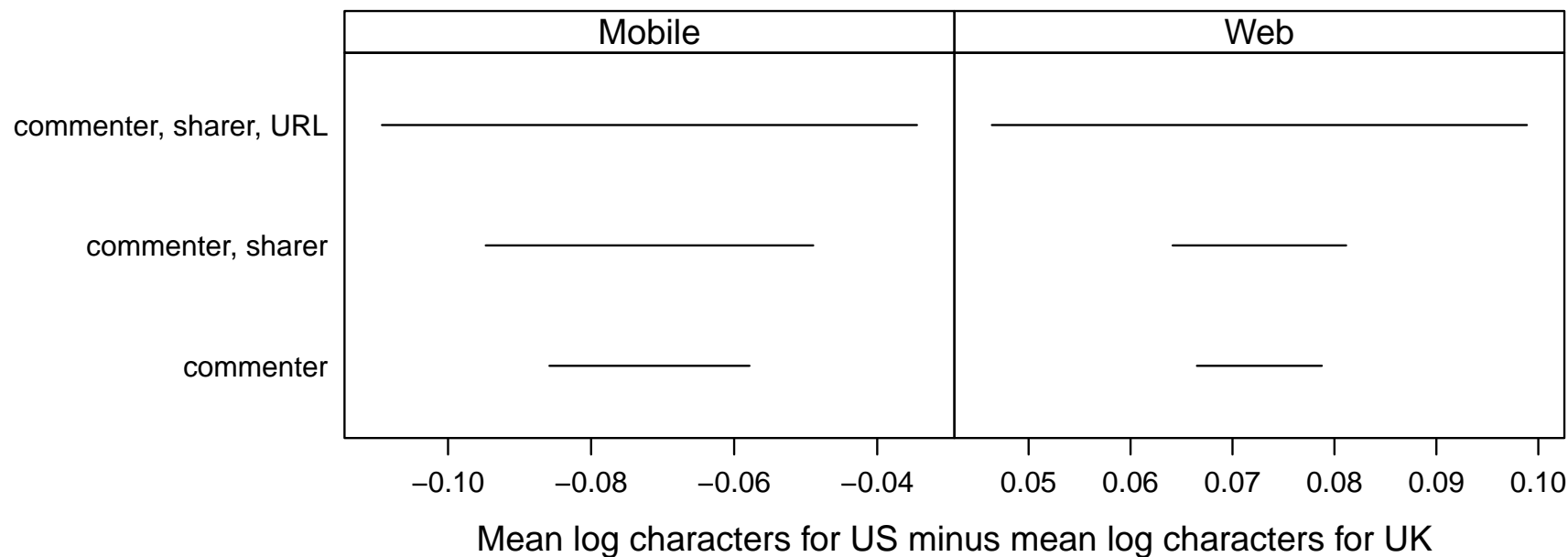
Loquacity ECDFs



ECDF over 50 bootstraps of $\hat{\mu}_{USm} - \hat{\mu}_{UKm}$

Reweighting one, two, or three ways

Loquacity confidence intervals



Central 95% confidence intervals from 50 bootstraps of $\hat{\mu}_{USm} - \hat{\mu}_{UKm}$

Reweighting one, two, or three ways

Heteroscedastic random effects

Every $u \subseteq \{1, 2, \dots, r\}$ and every $\mathbf{i}_u \in \mathbb{N}^{|u|}$ has its own variance

$$\sigma_{\mathbf{i},u}^2 \equiv \sigma_{\mathbf{i}_u,u}^2$$

We cannot estimate them all.

There may be association between $\sigma_{\mathbf{i},u}^2$ and $N_{\mathbf{i},u}$.

The analysis now has

$$V_{\text{RE}}(\hat{\mu}) = \frac{1}{N} \sum_u \sum_{\mathbf{i}} \nu_{\mathbf{i},u} \sigma_{\mathbf{i},u}^2, \quad \text{and}$$

$$\mathbb{E}_{\text{RE}}(\tilde{V}_{\text{PW}}(\hat{\mu}^*)) = \frac{1}{N} \sum_u \sum_{\mathbf{i}} \gamma_{\mathbf{i},u} \sigma_{\mathbf{i},u}^2$$

Product weights still give a mildly conservative variance, with relative error $1 + O(\eta + \epsilon)$ assuming uniform bounds:

$$0 < m \leq \min_{\mathbf{i},u} \sigma_{\mathbf{i},u}^2 \leq \max_{\mathbf{i},u} \sigma_{\mathbf{i},u}^2 \leq M < \infty.$$

Whence such heteroscedasticity?

Fixed factor F and random mean zero loading L

$$X_i = \mu + \cdots + F_{i_1} L_{i_2} + \cdots + \varepsilon_{i, \{1, \dots, r\}}$$

contributes $F_{i_1}^2 V(L_{i_2})$ to $\sigma_{i, \{i_2\}}^2$.

We could have both fixed $i_1 \times$ random i_2 and vice versa

More generally

For $v \neq \emptyset$ and $u \cap v = \emptyset$

$$\prod_{j \in u} F_{j, i_j} \times \prod_{j \in v} L_{j, i_j}$$

contributes $\prod_{j \in u} F_{h, i_j}^2 \prod_{j \in v} V(L_{j, i_j})$ to $\sigma_{i, v}^2$ when L_{j, i_j} are independent.

Factors and loadings don't have to be products

e.g. $F = \Phi(i_1, i_2, i_3)$ fixed & $L = \Lambda(i_4, i_5)$ indep mean 0

$F \times L$ contributes to $\sigma_{i, \{4, 5\}}^2$

So the model allows for generalized SVD contributions.

Gaps and potential next steps

- 1) The resampler does not imitate the generative model
- 2) Handling informative missing data
- 3) Inference for marginal means

$$\bar{X}_{i,u} = \frac{\sum_{i'} Z_{i'} 1_{i_u=i'_u} X_{i'}}{\sum_{i'} Z_{i'} 1_{i_u=i'_u}}$$

- 4) Defining, estimating, and inferring variance components
- 5) Inference for estimated factor models
- 6) What about $B = 1$, $B < 1$?

Thanks

- Dean Eckles for co-authoring
- Netflix and Facebook for data
- Steve Scott and Hal Varian for invitation
- Peter and Chuck for helping out
- NSF DMS-0906056 for support

The unistrap

Definition $\tilde{V}_{\text{PW}}(\hat{\mu}^*) \equiv \frac{1}{N^2} \mathbb{E}_{\text{PW}}((T^* - \hat{\mu}N^*)^2)$

Estimate $\widehat{\tilde{V}}_{\text{PW}}(\hat{\mu}^*) = \frac{1}{N^2} \frac{1}{B} \sum_{b=1}^B (T^{*b} - \hat{\mu}N^{*b})^2$

The b 'th independent bootstrap produces (T^{*b}, N^{*b}) for $b = 1, \dots, B$

Because we're using the ratio estimation formula the estimate exists for $B = 1$.

(and maybe for fractional sampling $B < 1$)

Modelling Z_i

- We do not model the missingness
- Analysis is conditional on Z_i
- Make no use/estimate of X_i for $Z_i = 0$

Can/should we do that?

- Missingness is very important
- Less so if you're predicting ratings that were actually made
- Modelling X_i for $Z_i = 0$ requires untestable assumptions (from outside the data)
- Later: use preferred imputation. Resample the result. MC based variance with expert's view of bias.

Repeated measures

Formally, the model has no duplicate indices

In practice we may get multiple observations at any i

We are studying sums for each i . This is heteroscedastic (for unequal sample sizes).

Alternative

We can adjoin an $r + 1^{\text{st}}$ index

This index describes a random effect nested within the first r effects

Best to have extra index be a unique data point identifier to avoid large ϵ

We could have s crossed random effects nested within each level of the first r effects

It fits into the model with

$$r' = r + s \quad \text{and} \quad \sigma_u^2 = 0 \quad \text{whenever}$$

$$u \cap \{r + 1, \dots, r + s\} \neq \emptyset \quad \text{and} \quad u \cap \{1, 2, \dots, r + s\} \neq \{1, 2, \dots, r + s\}$$

Exact formula depends on

| Notation | Definition | Meaning |
|---------------------|---|--|
| $N_{i,u}$ | $\sum_{i'} Z_{i'} 1_{i_u=i'_u}$ | Match i in u |
| ν_u | $N^{-1} \sum_i Z_i N_{i,u}$ | Avg # matches on u |
| $M_{ii'}$ | $\{j \mid i_j = i'_j\}$ | Match set for i & i' |
| $N_{i,k}$ | $\sum_{i'} Z_{i'} 1_{ M_{ii'} =k}$ | Match i in exactly k places |
| ρ_k | $N^{-1} \sum_i Z_i N_{i,k}$ | Avg # k -matches |
| $\nu_{k,u}$ | $N^{-2} \sum_i \sum_{i'} Z_i Z_{i'} 1_{ M_{ii'} =k} 1_{i_u=i'_u}$ | Match k places including u |
| $\tilde{\nu}_{k,u}$ | $N^{-3} \sum_i \sum_{i'} \sum_{i''} Z_i Z_{i'} Z_{i''} 1_{ M_{ii'} =k} 1_{i_u=i''_u}$ | Hmmm |
| " | $N^{-1} \sum_i N_{i,u} N_{i,k}$ | |

Exact result $\gamma_u = \sum_{k=0}^r (1 + \tau^2)^k (\nu_{k,u} - 2\tilde{\nu}_{k,u} + \rho_k \nu_u)$ non-asymptotic

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

Some history

Boot-II was called Boot-p,i by [Brennan, Harris Hanson \(1987\)](#)

p,i stands for person, item

They wanted to bootstrap variance component estimates in educational testing (students \times questions).

[McCullagh \(2000\)](#) showed it was impossible

[McCullagh \(2000\)](#) has two different Boot-II algorithms, one for nested data

See also [Wiley \(2001\)](#).