Bootstrapping r-Fold Tensor Data

Art B. Owen — Stanford University

Dean Eckles — Facebook Inc.

The IID bootstrap

- data are IID F
- we resample IID from the empirical distribution $\widehat{F}$
- 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	$\cdots$	Variable C
Case 1			
$\vdots$			
Case R			

1) Variables are named entities:
   - E.g. pressure, volume, income $\cdots$
   - 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 $\cdots$

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

Two-way 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

More examples of two-way data:

genes	$\times$	environments	$\rightarrow$	crop yields
terms	$\times$	documents	$\rightarrow$	counts
candidate	$\times$	interviewer	$\rightarrow$	rating
nodes	$\times$	more nodes	$\rightarrow$	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.

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. plants

$b_j \sim \mathcal{N}(0, \sigma_B^2)$ 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 $V(\hat{\mu})$

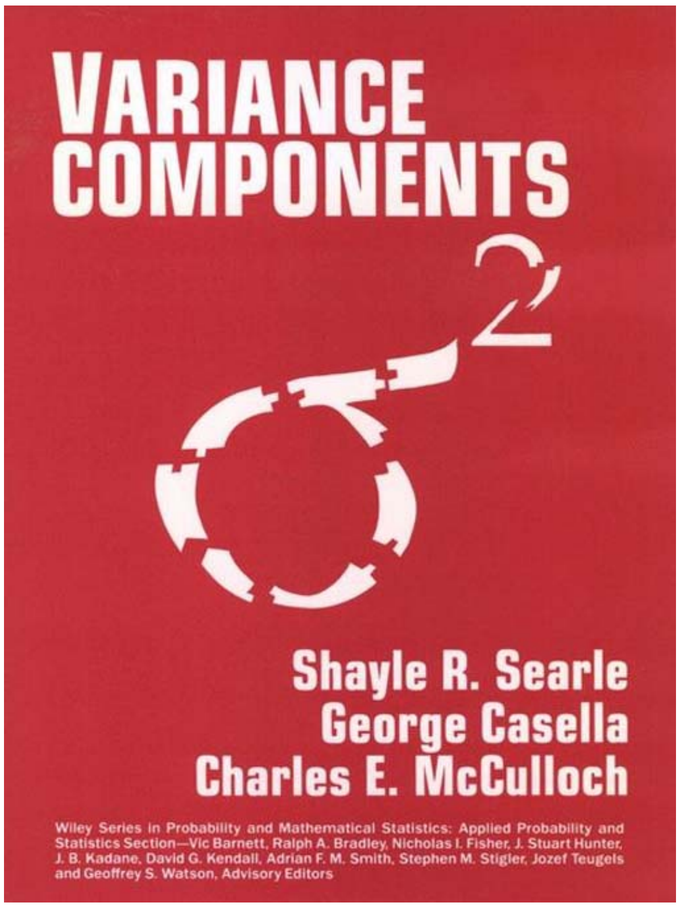
None can exist $\cdots$

$\cdots$ McCullagh (2000)

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

What about classical approaches?

prime reference:



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

McCullagh (2000)

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)

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

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

$$\mathbb{E}(\widehat{V}_{\mathrm{II}}(\hat{\mu})) \doteq \frac{\sigma_A^2}{R} + \frac{\sigma_B^2}{C} + \frac{3\sigma_E^2}{RC}$$ 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 $\cdots$ but conditions on pattern of observed data

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

Still get $\mathbb{E}(\widehat{V}_{\mathrm{B}}(\hat{\mu})) \doteq V(\hat{\mu})$, 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 $r \geqslant 2$.

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 $\log($# 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 $\boldsymbol{i} = (i_1, i_2, \dots, i_r) \in \{1, 2, 3, \cdots\}^r$

Sub-index $\boldsymbol{i}_u = (i_{j_1}, \dots, i_{j_L}) \quad u = \{j_1, \dots, j_L\} \subseteq \{1, 2, \dots, r\}$

Data $X_{\boldsymbol{i}} \in \mathbb{R}^d$ short for $X_{i_1, i_2, \dots, i_r}$ use $d = 1$

Presence $Z_{\boldsymbol{i}} \in \{0, 1\}$

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

$$X_{\boldsymbol{i}} = \mu + \sum_{u \neq \varnothing} \varepsilon_{\boldsymbol{i},u}$$

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

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

Homogeneous special case

$$\sigma^2_{\boldsymbol{i},u} \equiv \sigma^2_u \quad \forall \boldsymbol{i} \in \mathbb{N}^r \quad \forall u \subseteq \{1, \dots, r\}$$

The product reweighted bootstrap

$$\hat{\mu} = \frac{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} X_{\boldsymbol{i}}}{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}}} \quad \text{and} \quad \hat{\mu}^* = \frac{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} W_{\boldsymbol{i}} X_{\boldsymbol{i}}}{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} W_{\boldsymbol{i}}}$$

Our reweighting

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

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

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

Resampling vs. reweighing

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

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

Joys of half-sampling

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

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

$$X_{\boldsymbol{i}} = \mu + \sum_{u \neq \varnothing} \varepsilon_{\boldsymbol{i},u}$$

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

$$N \equiv \sum_{\boldsymbol{i}} Z_{\boldsymbol{i}}.$$

Then

$$V_{\text{RE}}(\hat{\mu}) = \frac{1}{N^2} \sum_{u \neq \varnothing} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{i}'} 1_{\boldsymbol{i}_u = \boldsymbol{i}'_u} \sigma_u^2 \equiv \frac{1}{N} \sum_{u \neq \varnothing} \nu_u \sigma_u^2$$

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

Our examples

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

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

For Facebook

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

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

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

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

Naive bootstrap (homog. case)

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

$$\mathbb{E}_{\text{RE}}(V_{\text{NB}}(\hat{\mu}^*)) = \frac{1}{N} \sum_{u \neq \varnothing} \Bigl(1 - \frac{\nu_u}{N}\Bigr) \sigma_u^2$$

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 $B \to \infty$ resamplings.

Product bootstrap

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

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

The delta method is reliable for large data

(Chamandy, Muralidharan, Najmi (2011))

Main result

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

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

otherwise small $\gamma_u/\nu_u > 1$

Exact formula depends on

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

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

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

Fine print from article

- $\nu_{k,u}$ depends on the number of $\boldsymbol{i}$, $\boldsymbol{i}'$ pairs that match in precisely k indices, including those in u.
- $\widetilde{\nu}_{k,j}$ depends on the number of triples $\boldsymbol{i}$, $\boldsymbol{i}'$, $\boldsymbol{i}''$ where $\boldsymbol{i}$ matches $\boldsymbol{i}'$ in the set u and matches $\boldsymbol{i}''$ 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 $\not\Rightarrow$ match occupation

Duplication indices

(level dup) $$\epsilon = \max_{\boldsymbol{i}} \max_{u \neq \varnothing} \frac{N_{\boldsymbol{i},u}}{N} = \max_{\boldsymbol{i}} \max_{1 \leqslant j \leqslant r} \frac{N_{\boldsymbol{i},\{j\}}}{N}$$

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

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

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

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

Facebook loquacity

For each commenter, url and sharer, we obtain:

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

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

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

Now let

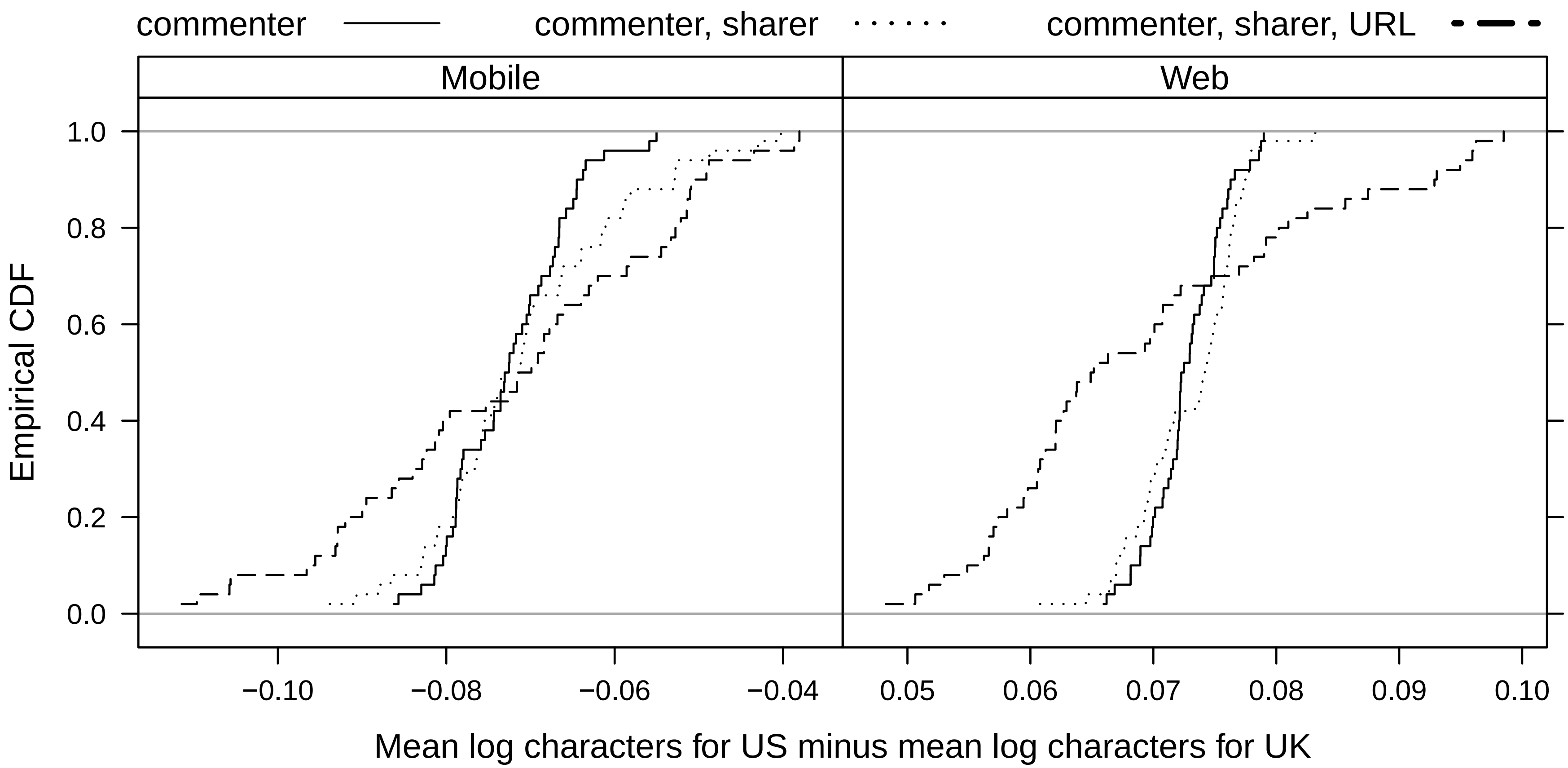
$$\hat{\mu}_{cm} = \frac{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} X_{\boldsymbol{i}} 1_{\text{country}=c} 1_{\text{mode}=m}}{\sum_{\boldsymbol{i}} Z_{\boldsymbol{i}} 1_{\text{country}=c} 1_{\text{mode}=m}}$$

We see small differences

	US	UK
web	3.62	3.55
mobile	3.50	3.57

but they're larger than sample fluctuations

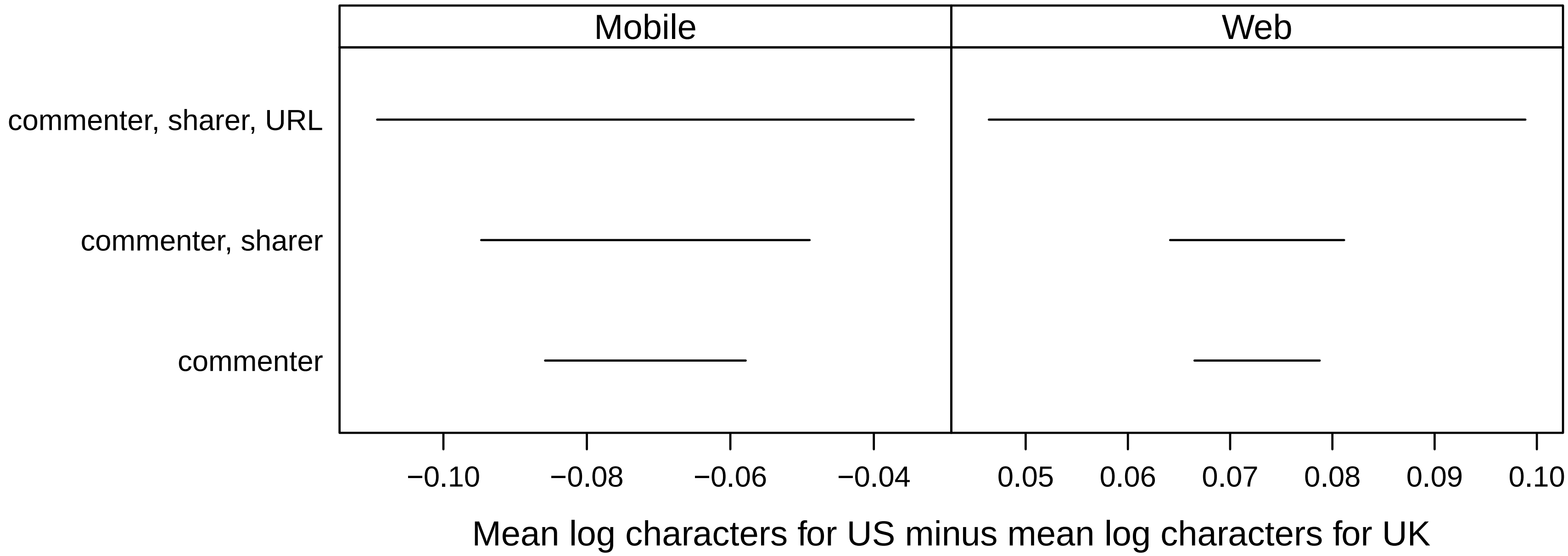
Loquacity ECDFs



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

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

Heterosceadastic random effects

Every $u \subseteq \{1, 2, \ldots, r\}$ and every $\boldsymbol{i}_u \in \mathbb{N}^{|u|}$ has it's own variance

$$\sigma^2_{\boldsymbol{i},u} \equiv \sigma^2_{\boldsymbol{i}_u,u}$$

We cannot estimate them all.

There may be association between $\sigma^2_{\boldsymbol{i},u}$ and $N_{\boldsymbol{i},u}$.

The analysis now has

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

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

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

$$0 < m \leqslant \min_{\boldsymbol{i},u} \sigma^2_{\boldsymbol{i},u} \leqslant \max_{\boldsymbol{i},u} \sigma^2_{\boldsymbol{i},u} \leqslant M < \infty.$$

Whence such heteroscedasticity?

Fixed factor F and random mean zero loading L

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

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

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

More generally

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

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

contributes $\prod_{j \in u} F^2_{h,i_j} \prod_{j \in v} V(L_{j,i_j})$ to $\sigma^2_{\boldsymbol{i},v}$ 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 $\sigma^2_{\boldsymbol{i},\{4,5\}}$

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

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

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
- Deepak Agarwal, David Banks, Diane Lambert
- Sue McDonald and Karem Jackson
- NSF DMS-0906056 for support

The unistrap

Definition $\qquad \widetilde{V}_{\mathsf{PW}}(\hat{\mu}^*) \equiv \frac{1}{N^2}\mathbb{E}_{\mathsf{PW}}\bigl((T^* - \hat{\mu}N^*)^2\bigr)$

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

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 $\boldsymbol{i}$

We are studying sums for each $\boldsymbol{i}$. 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

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

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

Exact formula depends on

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

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

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

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).