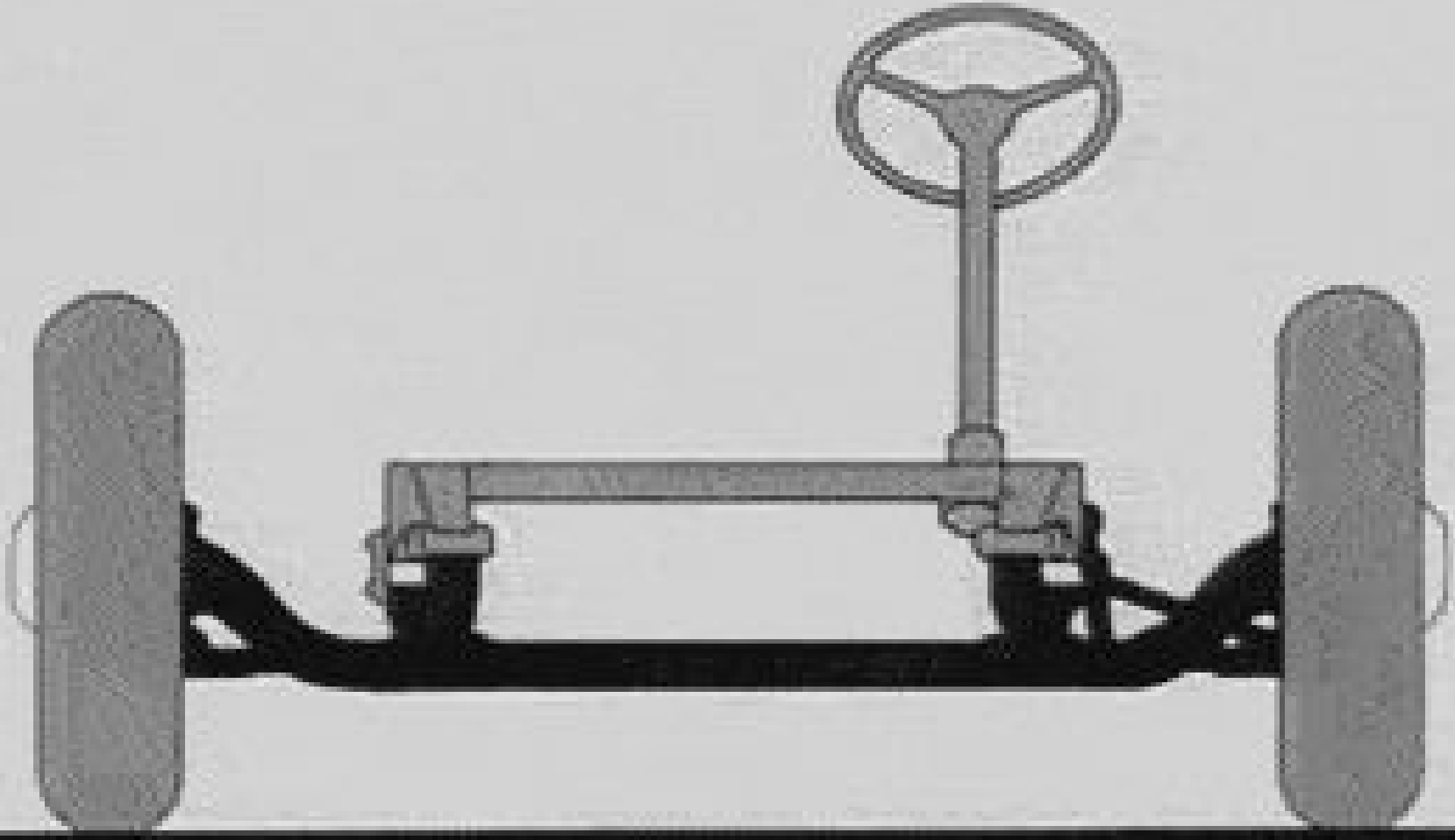
Uncertainty in Online Experiments with Dependent Data: An Evaluation of Bootstrap Methods

Eytan Bakshy and Dean Eckles,
Facebook Core Data Science



Jeremy

Secure https://www.kickstarter.com/recommendations

Explore Start a project For you KICKSTARTER

Projects for you

See all 1000 live projects

Perfect City Fall Launch 2017

by Perfect City

Perfect City is a 20-year art and activism project in NYC about how the people who live in a city can evolve it more inclusively.

Lower East Side, Manhattan, NY Conceptual Art

103%	$11,363	216	3
funded	pledged	backers	days to go

Video Games

Space Odyssey - The Video Game: A unique opportunity for gamers and fans to build the game — suggest ideas, ask…

by Neil deGrasse Tyson's Space Odyssey

Graphic Novels

BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by my family'…

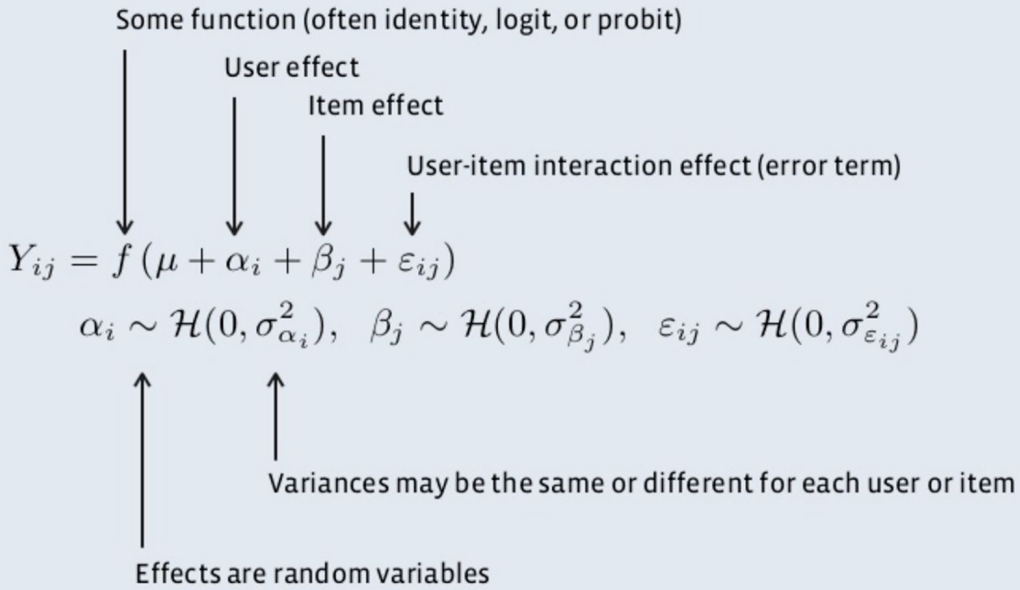
by Jeneeka Perera

Generative models

Slides from
https://www.slideshare.net/EytanBakshy/uncertainty-in-online-experiments-with-dependent-data/

Basic crossed random effects model

Random effects models are a general way of describing a data-generating process



A random effects model for experiments

Model of potential outcomes for thinking about ads, search, and feed experiments

Response of user *i* under treatment *d*

Average effect under *d*

d might affect units in different ways

$$Y_{ij}^{(d)} = \mu^{(d)} + \alpha_i^{(d)} + \beta_j^{(d)} + \varepsilon_{ij}^{(d)}$$

$$\vec{\alpha}_i \sim \mathcal{N}(0, \Sigma_\alpha), \quad \vec{\beta}_j \sim \mathcal{N}(0, \Sigma_\beta), \quad \vec{\varepsilon}_{ij} \sim \mathcal{N}(0, \Sigma_\varepsilon).$$

True difference in means $\delta \equiv \mathbb{E}[Y_{ij}^{(1)} \mid Z_{ij}^{(1)} = 1] - \mathbb{E}[Y_{ij}^{(0)} \mid Z_{ij}^{(0)} = 1].$

If $Z = Z^{(0)} = Z^{(1)}$, δ is an ATE

$$\begin{aligned}\delta &= \mathbb{E}[Y_{ij}^{(1)} - Y_{ij}^{(0)} \mid Z_{ij} = 1]. \\ &= \mu^{(1)} - \mu^{(0)}\end{aligned}$$

Variance of ATE for user-item experiments*

Estimates of the average treatment effect (ATE) include noise that depends on users and items

$$\mathrm{V}[\hat{\delta}] = \frac{1}{N}\Bigg[\left(\nu_A^{(1)}\sigma^2_{\alpha(1)} + \nu_A^{(0)}\sigma^2_{\alpha(0)}\right) + \left(\nu_B^{(1)}\sigma^2_{\beta(1)} + \nu_B^{(0)}\sigma^2_{\beta(0)} - 2\omega_\beta\sigma^2_{\beta(0),\beta(1)}\right) + \sigma^2_{\varepsilon(0)} + \sigma^2_{\varepsilon(1)}\Bigg].$$

Correlation between item-level effects under treatment and control

Duplication coefficients

$$\nu_A^{(d)} \equiv \frac{1}{N}\sum_i \left(n_{i\bullet}^{(d)}\right)^2 \qquad \nu_B^{(d)} \equiv \frac{1}{N}\sum_j \left(n_{\bullet j}^{(d)}\right)^2, \qquad \omega_B \equiv \frac{1}{N}\sum_j n_{\bullet j}^{(0)} n_{\bullet j}^{(1)}.$$

Average number of observations sharing the same user

Balance of items across conditions

*under the the linear homogeneous random effects model with an equal number of observations in each condition, randomizing over users with $Z^{(0)} = Z^{(1)}$

Bootstrapping dependent data

Jeremy

Secure https://www.kickstarter.com/recommendations

Explore Start a project For you KICKSTARTER

Projects for you

See all 1000 live projects

Perfect City Fall Launch 2017

by Perfect City

Perfect City is a 20-year art and activism project in NYC about how the people who live in a city can evolve it more inclusively.

Lower East Side, Manhattan, NY Conceptual Art

103%	$11,363	216	3
funded	pledged	backers	days to go

Video Games

Space Odyssey - The Video Game: A unique opportunity for gamers and fans to build the game — suggest ideas, ask…

by Neil deGrasse Tyson's Space Odyssey

Graphic Novels

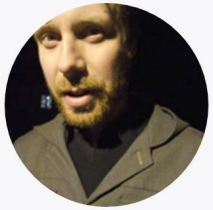
BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

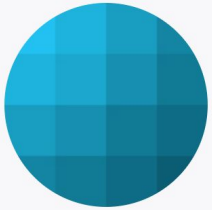
Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by my family'…

by Jeneeka Perera

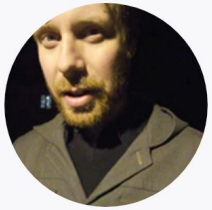


Jeremy Salfen



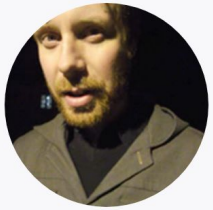
Jean Salfen

Tieg Zaharia

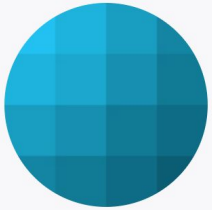


Jeremy Salfen

Tieg Zaharia

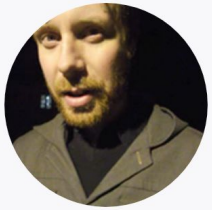


Jeremy Salfen



Jean Salfen

Tieg Zaharia



Jeremy Salfen

Tieg Zaharia

Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by…

by Jeneeka Perera

$5,387 pledged
15% funded
11 days to go

Graphic Novels

BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

$9,308 pledged
186% funded
6 hours to go

Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by…

by Jeneeka Perera

$5,387 pledged
15% funded
11 days to go

Video Games

Space Odyssey - The Video Game: A unique opportunity for gamers and fans to build the game — suggest ideas, as…

by Neil deGrasse Tyson's Space Odyssey

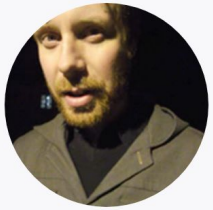
$237,236 pledged
75% funded
8 days to go

Graphic Novels

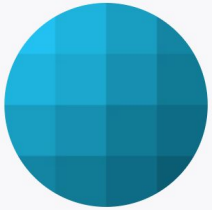
BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

$9,308 pledged
186% funded
6 hours to go

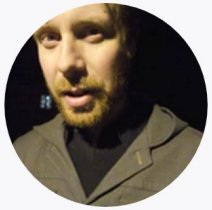


Jeremy Salfen



Jean Salfen

Tieg Zaharia



Jeremy Salfen

Tieg Zaharia

Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by…

by Jeneeka Perera

$5,387 pledged
15% funded
11 days to go

Graphic Novels

BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

$9,308 pledged
186% funded
6 hours to go

Cookbooks

Amma's Kitchen: A Sri Lankan Heritage Cookbook: Amma's Kitchen is a Sri Lankan heritage cookbook inspired by…

by Jeneeka Perera

$5,387 pledged
15% funded
11 days to go

Video Games

Space Odyssey - The Video Game: A unique opportunity for gamers and fans to build the game — suggest ideas, as…

by Neil deGrasse Tyson's Space Odyssey

$237,236 pledged
75% funded
8 days to go

Graphic Novels

BLACK: Remastered: A graphic novel set in a world where only black people have super powers.

by Kwanza Osajyefo

$9,308 pledged
186% funded
6 hours to go

1

0

1

0

1

Empirical evaluations

Basic A/A test

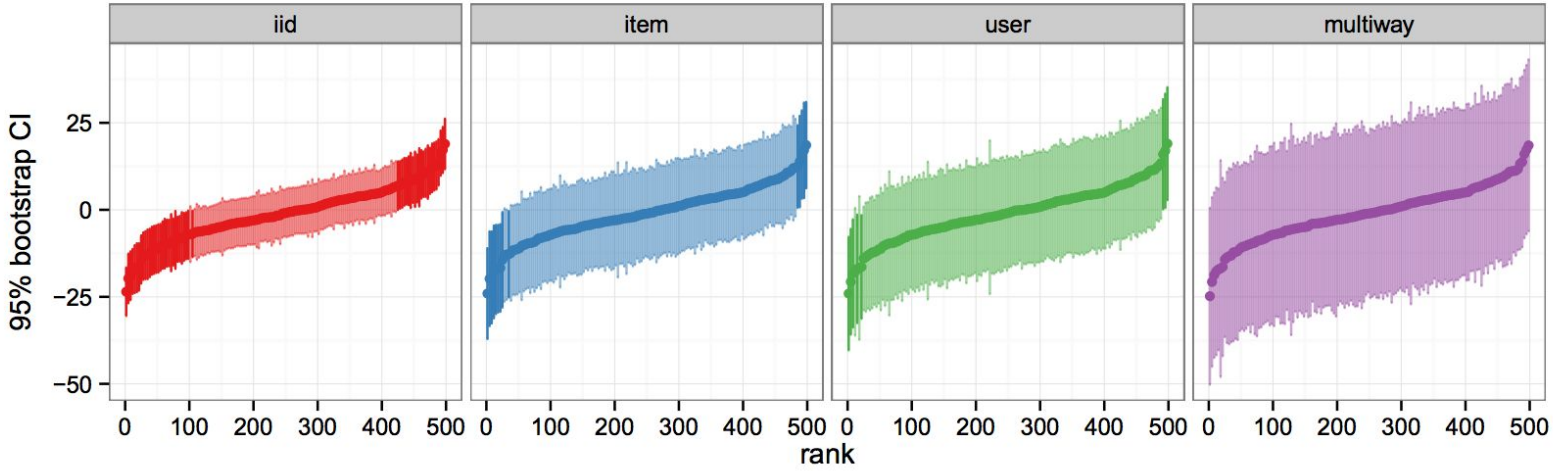


Figure 1: An illustration of our method for computing true coverage rates for the bootstrap methods with the Search dataset. We compute null experiments to obtain nominal "95% confidence intervals" for the difference in means $\hat{\delta}_{kr}$, and count the fraction of tests that accept the null hypothesis (e.g., indicate there is no significant difference in means). To show how results can vary between comparisons, we sort the results by $\mathbb{E}_r[\hat{\delta}_{kr}]$, and darken results that (incorrectly) reject the null. Anti-conservative tests – in this case, the iid and item-clustered bootstrap – reject in more than 5% of the experiments. Differences in the figure are shown relative to the grand mean.

Duplication

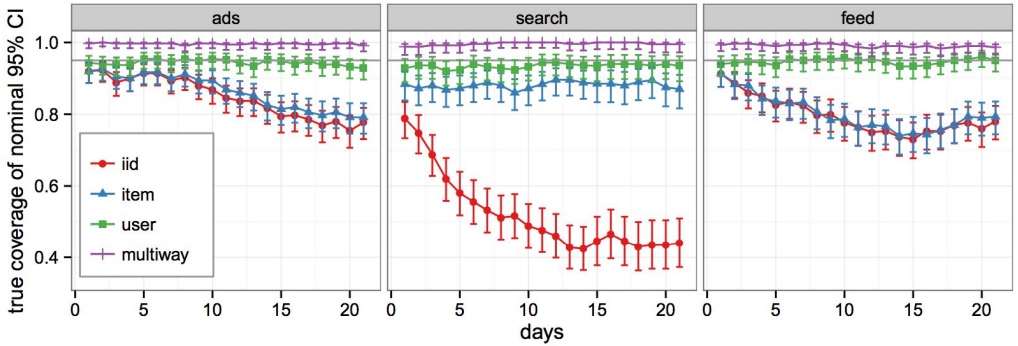


Figure 2: True coverage for nominal 95% confidence intervals produced by the iid, one-way, and multiway bootstrap for A/A tests segmented by user id as a function of time. Uncertainty estimates for the iid and item-level bootstrap become increasingly inaccurate over time, while the user-level and multiway bootstrap have the advertised or conservative Type I error rate.

	Ads	Search	Feed
users	4,515,816	908,339	545,218
items	317,159	1,362,061	326,831
user–item pairs	24,081,939	4,263,769	2,882,452
ν_{users}	18.5	35.5	20.3
ν_{items}	6,625.9	543.6	1,333.0

Table 1: The amount of duplication present in our datasets for a single 1% segment of users.

K bootstrap tests indicate a significant difference in means at $\alpha = 0.05$. We treat each of the K comparisons as independent, and use the Wilson score interval for binomial

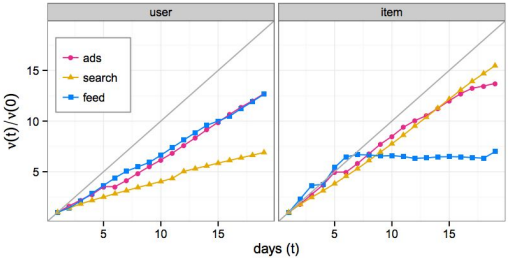


Figure 3: Duplication (ν) for users and items over time relative to the first day.

Imbalance in items

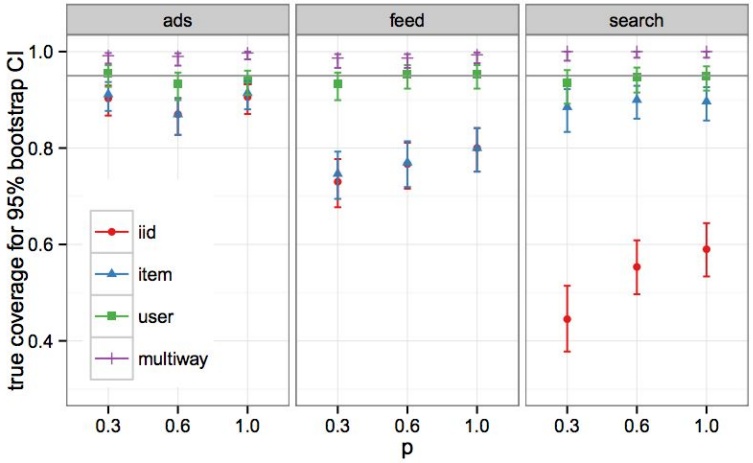


Figure 4: True coverage for nominal 95% confidence intervals for each bootstrap method applied to data with varying levels of synthetic imbalance of items across conditions for 2 weeks of data. Violations to the sharp null via imbalance do not appear to affect the accuracy of the true coverage for the multiway and user-level bootstrap, while the iid and item level bootstrap become more conservative when imbalance is greatest.

Simulations with effects

Item-treatment interactions

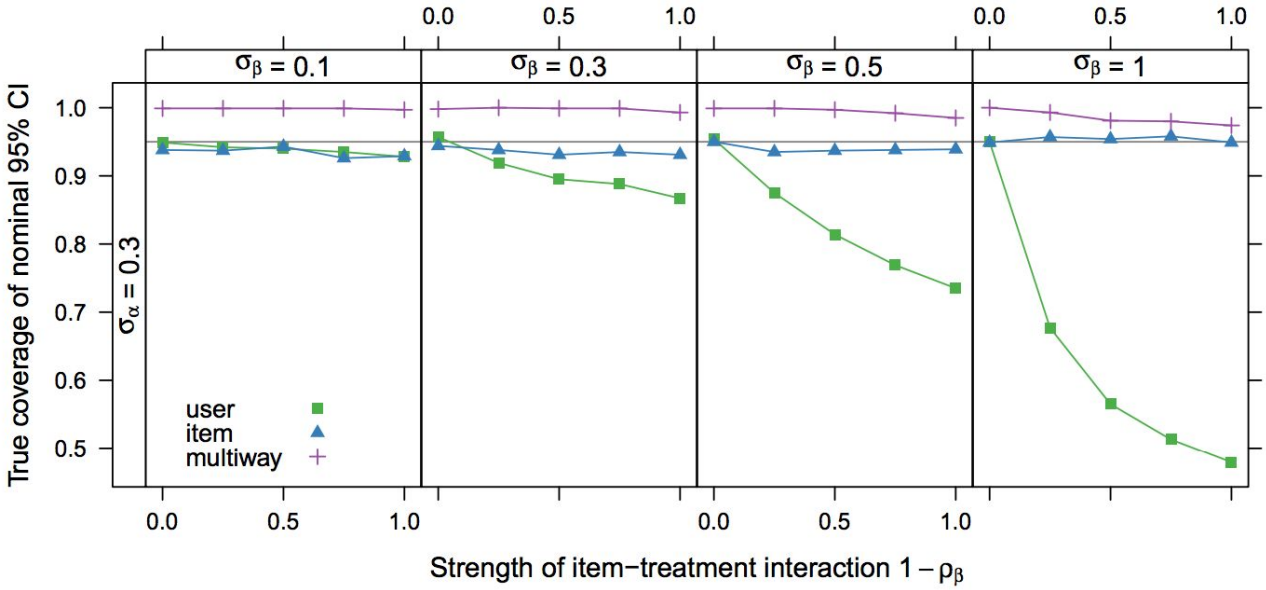


Figure 5: Effects of item–treatment interaction effects on true coverage for nominal 95% confidence intervals. Decreasing ρ_β, which makes the random item effects less correlated between treatment and control, reduces the coverage of user bootstrap confidence intervals. This effect is moderated by the magnitude of the item-level random effects.

KICKSTARTER

Other resources

Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions
http://alexdeng.github.io/public/files/WSDM2017draft.pdf

Mind Your Units
http://www.unofficialgoogledatascience.com/2016/08/mind-your-units.html

R implementation of the multiway bootstrap
https://github.com/deaneckles/multiway_bootstrap