Automated Bot Detection in Online Surveys with Bayesian Latent Class Models

Latent Class Bot Detection

Zachary Roman Ph.D., Holger Brandt Ph.D., & Jason M. Miller Ph.D.
University of Zurich, University of Tuebingen, University of Kansas
Zachary.Roman@psychologie.uzh.ch

IMPS 2022

Introduction

Roman, Brandt, and Miller (2022)

https://doi.org/10.3389/fpsyg.2022.789223

Introduction

Online survey platforms are useful

Behavioral Scientists can:

- Sample specific/ rare populations
- Many respondents in a short time
- Access to anyone with internet access

Introduction

Online survey platforms are prone to exploitation for profit

Some examples:

- Click farms
- Content Non-Responders (CNR)
- Bots

Introduction

Bot literature is fairly new, but the outcomes are the same with CNR

Data is contaminated

- Add noise to data (Buchanan & Scofield, 2018; A. W. Meade & Craig, 2012)
- Item correlations are drawn towards zero
- Increased error variance
- Type II error rate increase (Marjanovic et al., 2014)

Breif state of literature

There are three categories of bot/CNR detection

Catagories

- External items
  - (e.g.,DeSimone (2015);P. G. A. K. Huang J. L. AND Curran (2012);N. A. A. L. Huang J. L. AND Bowling (2015);Wise (2006))
  - Bogus items
  - Validity scales
  - Response times
- Indices approaches
  - Mismatch of positively and negative worded items (Greene, 1978)
  - Frequency coding of specific responses (Baumgartner, 2006)
  - Person-fit cutoff (Drasgow, 1985; Karabatsos, 2003)
- Model based approaches
  - ...

Breif state of literature

External and indices approaches both calculate an index and some cutoff is utilized

Model based

- Simultaneously estimate inattention with model of interest
- Majority utilize Latent Class (LC) framework
    - (Chen & Wang, 2018; S. B. Meade A. W. AND Craig, 2012; Terzi, 2017)
- Two or more classes
    - One is attentive
    - Rest represent CNR response patterns
        - e.g., uniform response pattern
        - Could include a-priori fit indices

Bayesian Latent Class Models

Goal: simouteaniously estimate a CFA model for cases not flagged as bots,

Latent class model with CFA

$$g(\mu_{y,ij}|_{C_i=1}) = \tau_{j1} + \boldsymbol{\lambda}_j \boldsymbol{\eta}_i \quad (1)$$

$$g(\mu_{y,ij}|_{C_i=2}) = \tau_{j2} \quad (2)$$

$$Y_{ij}|_{C_i=c} \sim F(\mu_{y,ij}|_{C_i=c}, (\sigma^2_{y,jc})) \quad (3)$$

where: g is a link function $F(\mu, (\sigma^2))$ is a distribution function, for continuous items, we use an identity function and a normal distribution C_i is the latent categorical variable $C = 2$ is bots $C = 1$ is attentive

Bayesian Latent Class Models for Bot Identification

for $C = 1$ (attentive class), we assume multivariate normality

Multivariate normality

$$\boldsymbol{\eta}_i|_{C_i=1} \sim MVN(\boldsymbol{\kappa}, \boldsymbol{\Phi}) \tag{4}$$

We assume standard SEM identification process for the latent factors of $C = 1$.

Bayesian Latent Class Models for Bot Identification

To ensure classes are interpreted correctly ($C_i = 1$ is attentive), restrictions are imposed on the other class

Restrictions

- We utilize item level mean and variances
- $(Y_{ij} | C_i = 2) \sim N(\tau_{j2}, \sigma^2_{y,j2})$

Bayesian Latent Class Models for Bot Identification

Previous model implementations have used

$\pi = (P(C_i = 1), \ldots, P(C_i = C_{max}))$

We utilize a sub-model to support in class prediction and consequently interpretation

Class prediction sub-model

$$P(C_i = 1 | \Upsilon_{1i}, \Upsilon_{2i}) = expit(\beta_0 + \beta_1 \Upsilon_{1i} + \beta_2 \Upsilon_{2i}) \tag{5}$$

- with $expit(x) := 1/(1 + exp(-x))$
- where
- Υ are additional predictors for the latent classes

Bayesian Latent Class Models for Bot Identification

For Υ_{1i} we use the person-fit index and Υ_{2i} person level factor variance

Class prediction sub-model

$$P(C_i = 1|\Upsilon_{1i}, \Upsilon_{2i}) = expit(\beta_0 + \beta_1\Upsilon_{1i} + \beta_2\Upsilon_{2i}) \tag{6}$$

- Higher Υ_{1i} suggest greater departures from the factor model
- Higher Υ_{2i} suggest greater within factor variability
- Reverse coded items

Predictors of class membership

Variability Υ_{1i}

$$\Upsilon_{2i} = \frac{1}{m} \sum_{k=1}^{m} Var(\mathbf{y}_{ik}) \tag{7}$$

Responses to items that belong to the same factor should have a rather small variability because persons are more likely to respond in a similar fashion depending on their expression of the construct (e.g., low or high). Bots with a random response style will provide a larger variability in comparison.

Predictors of class membership

In general greater values of the person-fit index suggest greater departures from the factor model, suggesting bot like responses.

Bayesian Estimatation and priors

We specify the model in Jags

Model for $C_i = 1$

- We specify our model of interest as a classic CFA model

where

$$\boldsymbol{\eta}_i \sim MVN(\boldsymbol{\kappa}, \boldsymbol{\Phi}), \qquad i = 1 \dots N \tag{8}$$

The latent class variable follows a Bernoulli distribution

$$C_i \sim Bern(\pi_i), \qquad i = 1 \dots N \tag{9}$$

with $\pi_i = expit(\beta_0 + \beta_1 \Upsilon_{1i} + \beta_2 \Upsilon_{2i})$.

Bayesian Estimatation and priors

General priors

$$
\begin{aligned}
\tau_{jc} &\sim N(\mu_{\tau 0c}, \sigma^2_{\tau 0c}), & j = 1 \dots p, c = 1, 2 && (10)\\
\lambda_{jk} &\sim N(\mu_{\lambda 0j}, \sigma^2_{\lambda 0j}), & j = 1 \dots p, k = 1 \dots m && (11)\\
\kappa_k &\sim N(\mu_{\kappa 0k}, \sigma^2_{\kappa 0k}), & k = 1 \dots m && (12)\\
\boldsymbol{\Psi}^{-1} &\sim Wish(\boldsymbol{\Psi}_0, df_{\Psi}) & && (13)\\
\beta_r &\sim N(\mu_{\beta 0r}, \sigma^2_{\beta 0r}), & r = 0 \dots 2 && (14)\\
\sigma^{-2}_{jc} &\sim Ga(a_{\sigma jc}, b_{\sigma jc}), & j = 1 \dots p, c = 1, 2. && (15)
\end{aligned}
$$

Emperical Example

Data were collected via Mturk, in an unrelated political psychology study, prior to Mturk's implementation of more stringent screening criteria.

Therefor the meta data reveals known bots in the survey (duplicated IP addresses, browser, and geolocation data)

Data

- n = 395
- 40% bots based on meta-data
- Three factors were measured
  - Social Dominance Orientation (SDO)
  - Nationalism (NAT)
  - Right Wing Authoritarianism (RWA)

We can say with relative certainty, that those flagged as bots are bots, but not the opposite.

Methods

Goals

- Model the data with the aforementioned LC-CFA and investigate bot identification accuracy with real bots
- Model the data with a traditional CFA and compare parameter estimates to imitate ignoring bots

Results

JAGS version 4.2 (Plummer, 2003) and deployed in R version 3.6.2 (R Core Team, 2019), 3 chains were specified with 12,000 iterations each, half of which was burn-in, with a thinning parameter of 2. For 18,000 total post burn-in draws.

Diagnostics

Typical CFA parameters exhibited acceptable $\hat{R}$ statistics, all < 1.001, and the lowest ESS $= 490$, mean ESS $= 5057$

Results

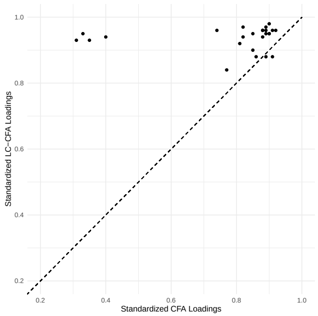
Diagnostic Accuracy

Table 1: Sensitivity = 71.07%, Specificity = 95.34%

		Estimated Class	
		Bot	Non-bot
True Class	Bot	113	46
	Non-bot	11	225

- Variance function β CE [-0.423;-0.127]
- Person-fit index β CE [-0.017;-0.001]
- Higher scores on both indicate higher probability of being a bot

Results



	Factor(s)	Θ_{CFA}	$\Theta_{\text{LC-CFA}}$	$\hat{R}_{CFA}$	ESS_{CFA}	$\hat{R}_{LC-CFA}$	$\text{ESS}_{\text{LC-CFA}}$
Correlations							
	RWA & SDO	.77	.89	1.00	1500	1.00	1200
	RWA & NAT	.76	.88	1.00	18000	1.00	1400
	SDO & NAT	.90	.95	1.00	2500	1.00	18000
Variances							
	RWA	14.44	12.01	1.00	15000	1.00	4100
	SDO	12.96	11.10	1.00	4900	1.00	890
	NAT	11.61	14.40	1.00	4100	1.00	3300

Discussion I

- The LC-CFA successfully removes the majority of known bots from the data
- When this happens, the CFA parameters move away from zero
- In a comparable simulation setting ($N = 400$ and 50% bots, specificity = 98%) the example model exhibited similar specificity (95.34%)
- In the same simulation setting (N = 400 and 50% bots, sensitivity = 99%) the example model exhibited much lower sensitivity (71.07%)

Discussion II

Online survey services have adapted and now implement more stringent screening criteria

Discussion II

Online survey services have since adapted and now implement more stringent screening criteria

But, as a famous bot once said . . .

Discussion II

I'll be back

Thank You

Thank you IMPS 2022

Appendix

Person-Fit Index

$$ll_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\left(p \cdot \ln(2\pi) + \ln|\boldsymbol{\Sigma}| + D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) \tag{16}$$

with a Mahalanobis distances D_i^2 based on these model-implied mean vector and covariance matrix

$$D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \tag{17}$$

Calculate

$$\Upsilon_{1i} = -2 \cdot (ll_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - ll_i(\bar{\mathbf{y}}, \mathbf{S})) \tag{18}$$