Robust estimation and inference with categorical data

Max Welz

Erasmus School of Economics, Erasmus University Rotterdam
Soon: University of Zurich

August 27, 2024, EEA-ESEM

Who thinks that questionnaire participants always respond accurately and attentively?

I propose a general methodological solution!

…and much more…

Who thinks that questionnaire participants always respond accurately and attentively?

I propose a general methodological solution!

. . . and much more. . .

Motivation: Respondent inattention

Behavior Research Methods (2020) 52:2489–2505
https://doi.org/10.3758/s13428-020-01401-8

A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data

Víctor B. Arias[1] · L. E. Garrido[2] · C. Jenaro[1] · A. Martínez-Molina[3] · B. Arias[4]

Published online: 27 May 2020
© The Psychonomic Society, Inc. 2020

Journal of Applied Psychology
2015, Vol. 100, No. 3, 828–845

© 2014 American Psychological Association
0021-9010/15/$12.00 http://dx.doi.org/10.1037/a0038510

Insufficient Effort Responding: Examining an Insidious Confound in Survey Data

Jason L. Huang and Mengqiao Liu
Wayne State University

Nathan A. Bowling
Wright State University

Motivation: Respondent inattention

Respondent inattention is a big problem in questionnaire studies

- Can lead to biased parameter estimates, invalid inference, deteriorated model fit, errors in hypothesis testing (e.g. Arias et al., 2020; Huang et al., 2015; Meade & Craig, 2012)
- Already prevalence of 5–10% problematic (e.g. Credé, 2010; Woods, 2006)
- Likely present in all questionnaire data (Ward & Meade, 2023)

Motivation: Respondent inattention

Respondent inattention is a big problem in questionnaire studies

- Can lead to biased parameter estimates, invalid inference, deteriorated model fit, errors in hypothesis testing (e.g. Arias et al., 2020; Huang et al., 2015; Meade & Craig, 2012)
- Already prevalence of 5–10% problematic (e.g. Credé, 2010; Woods, 2006)
- Likely present in all questionnaire data (Ward & Meade, 2023)

Contribution (Welz, 2024)

Develop estimator for categorical data that is robust to inattention/misspecification

- Novel categorical analogue to robust M-estimation theory (Huber, 1964)
- No assumption on the type or magnitude of misspecification
- Generalizes MLE, attractive statistical guarantees
- Statistical test to identify cells/responses that cannot be fitted well

Contribution (Welz, 2024)

Develop estimator for categorical data that is robust to inattention/misspecification

- Novel categorical analogue to robust *M*-estimation theory (Huber, 1964)
- No assumption on the type or magnitude of misspecification
- Generalizes MLE, attractive statistical guarantees
- Statistical test to identify cells/responses that cannot be fitted well

General setup

A model $\{p(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for k-dimensional categorical $\boldsymbol{Z}$, parameter $\boldsymbol{\theta} \in \mathbb{R}^d$

- Categorical outcome takes values in finite sample space $\mathcal{Z} = \{z_1, \ldots, z_m\}$
- Model assigns to each event $z \in \mathcal{Z}$ a probability $p_z(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[Z = z]$
- Examples:
  - Factor models/SEMs on latent variables for survey scales
  - Discrete choice
  - Poisson counting process...

General setup

A model $\{p(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for k-dimensional categorical $\boldsymbol{Z}$, parameter $\boldsymbol{\theta} \in \mathbb{R}^d$

- Categorical outcome takes values in finite sample space $\mathcal{Z} = \{z_1, \ldots, z_m\}$
- Model assigns to each event $z \in \mathcal{Z}$ a probability $p_z(\theta) = \mathbb{P}_\theta[Z = z]$
- Examples:
  - Factor models/SEMs on latent variables for survey scales
  - Discrete choice
  - Poisson counting process...

General setup

A model $\{p(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for k-dimensional categorical $\boldsymbol{Z}$, parameter $\boldsymbol{\theta} \in \mathbb{R}^d$

- Categorical outcome takes values in finite sample space $\boldsymbol{\mathcal{Z}} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$
- Model assigns to each event $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ a probability $p_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{Z} = \boldsymbol{z}]$
- Examples:
  - Factor models/SEMs on latent variables for survey scales
  - Discrete choice
  - Poisson counting process...

General setup

A model $\{p(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for k-dimensional categorical $\boldsymbol{Z}$, parameter $\boldsymbol{\theta} \in \mathbb{R}^d$

- Categorical outcome takes values in finite sample space $\boldsymbol{\mathcal{Z}} = \{z_1, \ldots, z_m\}$
- Model assigns to each event $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ a probability $p_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{Z} = \boldsymbol{z}]$
- Examples:
  - Factor models/SEMs on latent variables for survey scales
  - Discrete choice
  - Poisson counting process. . .

Model misspecification (Huber, 1964, AoS)

Misspecification: Instead of true $p(\boldsymbol{\theta}_*)$, sample from corrupted mixture

$$f_\varepsilon = (1-\varepsilon)p(\boldsymbol{\theta}_*) + \varepsilon h$$

- Fraction $\varepsilon \in [0,1]$ is degree of misspecification (unspecified)
- Density h is type of misspecification (unspecified)

No assumptions on misspecification!

Model misspecification (Huber, 1964, AoS)

Misspecification: Instead of true $p(\boldsymbol{\theta}_*)$, sample from corrupted mixture

$$f_\varepsilon = (1 - \varepsilon)p(\boldsymbol{\theta}_*) + \varepsilon h$$

- Fraction $\varepsilon \in [0, 1]$ is degree of misspecification (unspecified)
- Density h is type of misspecification (unspecified)

No assumptions on misspecification!

Model misspecification (Huber, 1964, AoS)

Misspecification: Instead of true $p(\boldsymbol{\theta}_*)$, sample from corrupted mixture

$$f_\varepsilon = (1-\varepsilon)p(\boldsymbol{\theta}_*) + \varepsilon h$$

- Fraction $\varepsilon \in [0, 1]$ is degree of misspecification (unspecified)
- Density h is type of misspecification (unspecified)

No assumptions on misspecification!

Model misspecification (Huber, 1964, AoS)

Misspecification: Instead of true $p(\boldsymbol{\theta}_*)$, sample from corrupted mixture

$$f_\varepsilon = (1 - \varepsilon)p(\boldsymbol{\theta}_*) + \varepsilon h$$

- Fraction $\varepsilon \in [0, 1]$ is degree of misspecification (unspecified)
- Density h is type of misspecification (unspecified)

No assumptions on misspecification!

Notation

- Let $\{Z_i\}_{i=1}^N$ denote N samples from f_ε $\implies$ model is possibly misspecified!
- $\widehat{f}_N(z) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\{Z_i = z\}$ is empirical probability of event $z \in \mathcal{Z}$
- $p_z(\boldsymbol{\theta})$ is theoretical probability of z at $\boldsymbol{\theta}$ (returned by model)

Proposed estimator

The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ minimizes over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right) p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

The fraction $\widehat{f}_N(z)/p_z\left(\theta\right)$ is called Pearson residual (Lindsay, 1994, AoS)

- Values close to 1 indicate good model fit, far away from 1 poor fit
- Avoid that classes that cannot be fitted well dominate fit
- Idea: Downweight influence of poorly fitted classes via choice of $\rho(\cdot)$

Proposed estimator

The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ minimizes over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{z}}(\boldsymbol{\theta})$$

The fraction $\widehat{f}_N(\boldsymbol{z}) / p_{\boldsymbol{z}}(\boldsymbol{\theta})$ is called Pearson residual (Lindsay, 1994, AoS)

- Values close to 1 indicate good model fit, far away from 1 poor fit
- Avoid that classes that cannot be fitted well dominate fit
- Idea: Downweight influence of poorly fitted classes via choice of $\rho(\cdot)$

Proposed estimator

The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ minimizes over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho \left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \right) p_{\boldsymbol{z}}(\boldsymbol{\theta})$$

The fraction $\widehat{f}_N(\boldsymbol{z}) / p_{\boldsymbol{z}}(\boldsymbol{\theta})$ is called Pearson residual (Lindsay, 1994, AoS)

- Values close to 1 indicate good model fit, far away from 1 poor fit
- Avoid that classes that cannot be fitted well dominate fit
- Idea: Downweight influence of poorly fitted classes via choice of $\rho(\cdot)$

Proposed estimator

The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ minimizes over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right) p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

The fraction $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$ is called Pearson residual (Lindsay, 1994, AoS)

- Values close to 1 indicate good model fit, far away from 1 poor fit
- Avoid that classes that cannot be fitted well dominate fit
- Idea: Downweight influence of poorly fitted classes via choice of $\rho(\cdot)$

Proposed estimator

The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ minimizes over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the loss

$$L\left(\boldsymbol{\theta}, \widehat{f}_N\right) = \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \rho \left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})} \right) p_{\boldsymbol{z}}(\boldsymbol{\theta})$$

The fraction $\widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ is called Pearson residual (Lindsay, 1994, AoS)

- Values close to 1 indicate good model fit, far away from 1 poor fit
- Avoid that classes that cannot be fitted well dominate fit
- Idea: Downweight influence of poorly fitted classes via choice of $\rho(\cdot)$

Choice of discrepancy function $\rho(\cdot)$

Depending on the situation, one may choose from an array of $\rho(\cdot)$ functions

- Theory developed for general $\rho(\cdot)$
- This talk (for simplicity): specific choice of $\rho(\cdot)$

Robust choice of $\rho(\cdot)$(Ruckstuhl & Welsh, 2001, AoS)

For $x = \widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ a Pearson residual at $\boldsymbol{\theta}$, use function

$$\rho(x) = \begin{cases} x \log(x) & \text{if } x \in [0, c], \\ x(\log(c) + 1) - c & \text{if } x > c, \end{cases}$$

where the constant $c \in [1, \infty]$ is prespecified

- If $x \in [0, c]$: Good fit, loss behaves like MLE $\Longrightarrow$ no need to downweight
- If $x > c$: Bad fit $\Longrightarrow$ downweight influence to be linear

Robust choice of $\rho(\cdot)$(Ruckstuhl & Welsh, 2001, AoS)

For $x = \widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ a Pearson residual at $\boldsymbol{\theta}$, use function

$$\rho(x) = \begin{cases} x\log(x) & \text{if } x \in [0, c], \\ x(\log(c) + 1) - c & \text{if } x > c, \end{cases}$$

where the constant $c \in [1, \infty]$ is prespecified

- If $x \in [0, c]$: Good fit, loss behaves like MLE $\Longrightarrow$ no need to downweight
- If $x > c$: Bad fit $\Longrightarrow$ downweight influence to be linear

Robust choice of $\rho(\cdot)$(Ruckstuhl & Welsh, 2001, AoS)

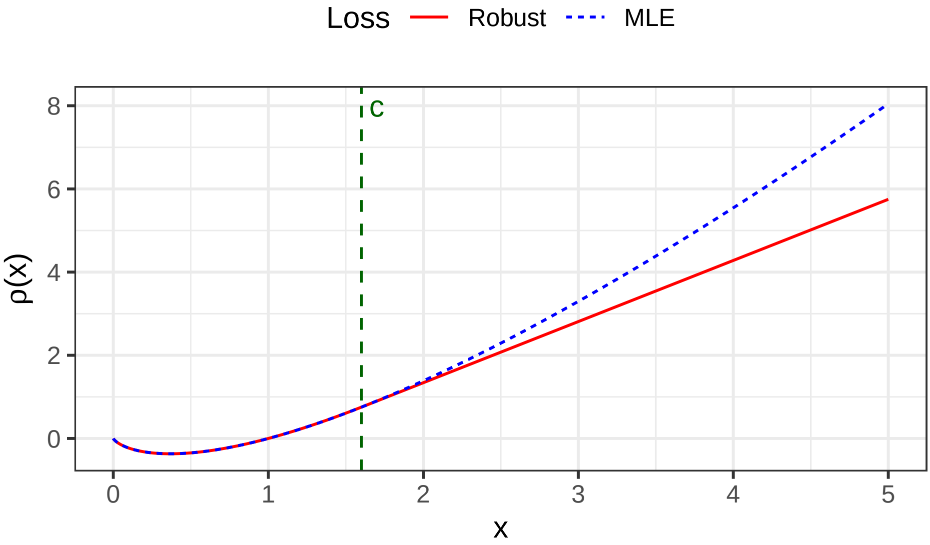
For $x = \widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ a Pearson residual at $\boldsymbol{\theta}$, use function

$$\rho(x) = \begin{cases} x \log(x) & \text{if } x \in [0, c], \\ x(\log(c) + 1) - c & \text{if } x > c, \end{cases}$$

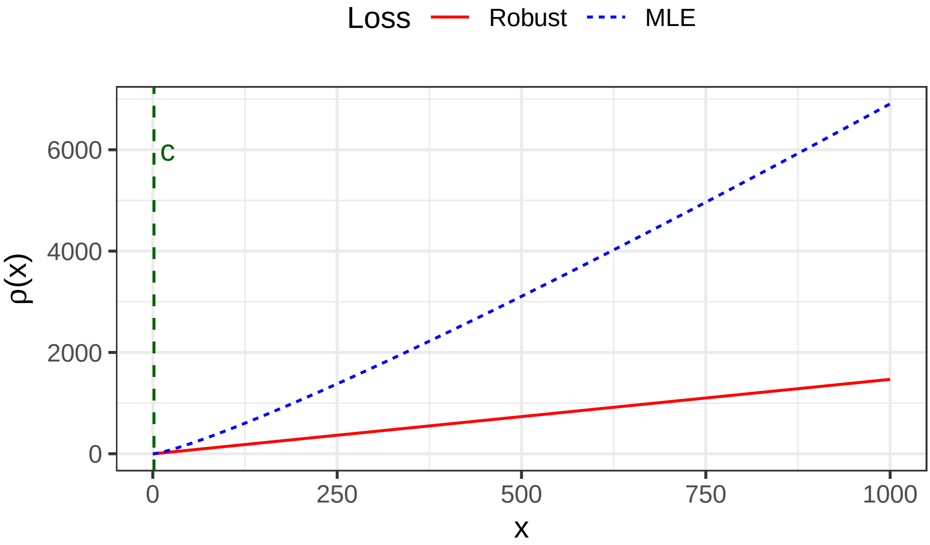
where the constant $c \in [1, \infty]$ is prespecified

- If $x \in [0, c]$: Good fit, loss behaves like MLE $\Longrightarrow$ no need to downweight
- If $x > c$: Bad fit $\Longrightarrow$ downweight influence to be linear

Robust choice of $\rho(\cdot)$, for $c = 1.6$



Robust choice of $\rho(\cdot)$, for $c = 1.6$



Asymptotics

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{z}}(\boldsymbol{\theta}) \qquad \rho(x) = \begin{cases} x\log(x) & \text{if } x \in [0, c], \\ x(\log(c)+1) - c & \text{if } x > c. \end{cases}$$

Estimand: $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L\left(\boldsymbol{\theta}, f_\varepsilon\right)$, equals $\boldsymbol{\theta}_*$ if $\varepsilon = 0$ (Fisher consistent)

Theorem (Consistency & asymptotic normality)

Under standard mild regularity conditions assumptions, it holds that

$$\widehat{\theta}_N \xrightarrow{\text{a.s.}} \theta_0$$

as $N \to \infty$, and

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right) \xrightarrow{\text{d}} \text{N}_d\left(0, \Sigma\left(\theta_0\right)\right),$$

where $\Sigma(\theta) = \boldsymbol{M}(\theta)^{-1}\boldsymbol{U}(\theta)\boldsymbol{M}(\theta)^{-1}$ is MLE variance at true model def more.

Asymptotics

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{z}}(\boldsymbol{\theta}) \qquad \rho(x) = \begin{cases} x\log(x) & \text{if } x \in [0, c], \\ x(\log(c)+1) - c & \text{if } x > c. \end{cases}$$

Estimand: $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L\left(\boldsymbol{\theta}, f_\varepsilon\right)$, equals $\boldsymbol{\theta}_*$ if $\varepsilon = 0$ (Fisher consistent)

Theorem (Consistency & asymptotic normality)

Under standard mild regularity conditions assumptions, it holds that

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$$

as $N \to \infty$, and

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \xrightarrow{\text{d}} \mathrm{N}_d\left(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{M}(\boldsymbol{\theta})^{-1}\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{M}(\boldsymbol{\theta})^{-1}$ is MLE variance at true model def more.

Example: Analysis of survey scales

Standard model to analyze responses to survey scales (e.g. Muthén, 1984)

- Suppose $Z_j \in \{1, \dots, K_j\}$ is a response to j-th survey question, $j \in [k]$
- Discrete Z_j is governed by unobserved discretization of latent ξ_j (e.g., utility)
- Identification: Assume $\boldsymbol{\xi} = (\xi_1, \dots, \xi_k)^\top$ is multivariate standard normal
- $\boldsymbol{\theta}$ holds the correlation parameters and discretization thresholds

Goal: Estimate correlation structure of $\boldsymbol{\xi}$ from discrete $\boldsymbol{Z} = (Z_1, \dots, Z_k)^\top$

Example: Analysis of survey scales

Standard model to analyze responses to survey scales (e.g. Muthén, 1984)

- Suppose $Z_j \in \{1, \ldots, K_j\}$ is a response to j-th survey question, $j \in [k]$
- Discrete Z_j is governed by unobserved discretization of latent ξ_j (e.g., utility)
- Identification: Assume $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$ is multivariate standard normal
- $\boldsymbol{\theta}$ holds the correlation parameters and discretization thresholds

Goal: Estimate correlation structure of $\boldsymbol{\xi}$ from discrete $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\top$

Example: Analysis of survey scales

Standard model to analyze responses to survey scales (e.g. Muthén, 1984)

- Suppose $Z_j \in \{1, \ldots, K_j\}$ is a response to j-th survey question, $j \in [k]$
- Discrete Z_j is governed by unobserved discretization of latent ξ_j (e.g., utility)
- Identification: Assume $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$ is multivariate standard normal
- $\boldsymbol{\theta}$ holds the correlation parameters and discretization thresholds

Goal: Estimate correlation structure of ξ from discrete $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\top$

Example: Analysis of survey scales

Standard model to analyze responses to survey scales (e.g. Muthén, 1984)

- Suppose $Z_j \in \{1, \ldots, K_j\}$ is a response to j-th survey question, $j \in [k]$
- Discrete Z_j is governed by unobserved discretization of latent ξ_j (e.g., utility)
- Identification: Assume $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$ is multivariate standard normal
- $\boldsymbol{\theta}$ holds the correlation parameters and discretization thresholds

Goal: Estimate correlation structure of $\boldsymbol{\xi}$ from discrete $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\top$

Example: Analysis of survey scales

Standard model to analyze responses to survey scales (e.g. Muthén, 1984)

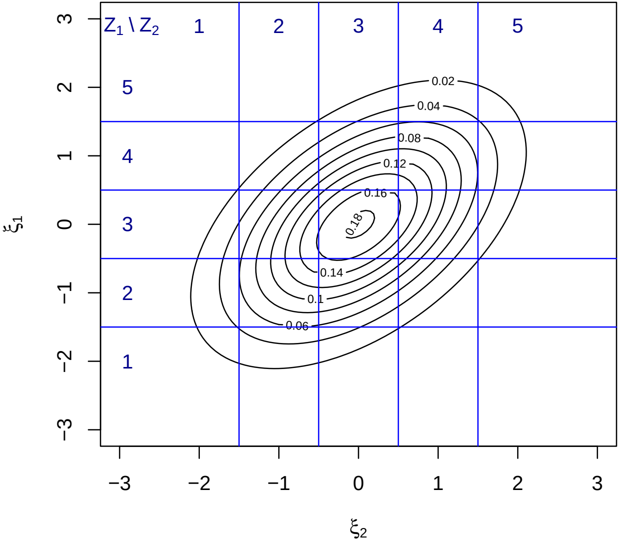
- Suppose $Z_j \in \{1, \ldots, K_j\}$ is a response to j-th survey question, $j \in [k]$
- Discrete Z_j is governed by unobserved discretization of latent ξ_j (e.g., utility)
- Identification: Assume $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$ is multivariate standard normal
- $\boldsymbol{\theta}$ holds the correlation parameters and discretization thresholds

Goal: Estimate correlation structure of $\boldsymbol{\xi}$ from discrete $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\top$

Example: Analysis of survey scales

The τ-thresholds discretize $\boldsymbol{\xi}$ to $\boldsymbol{Z}$ via

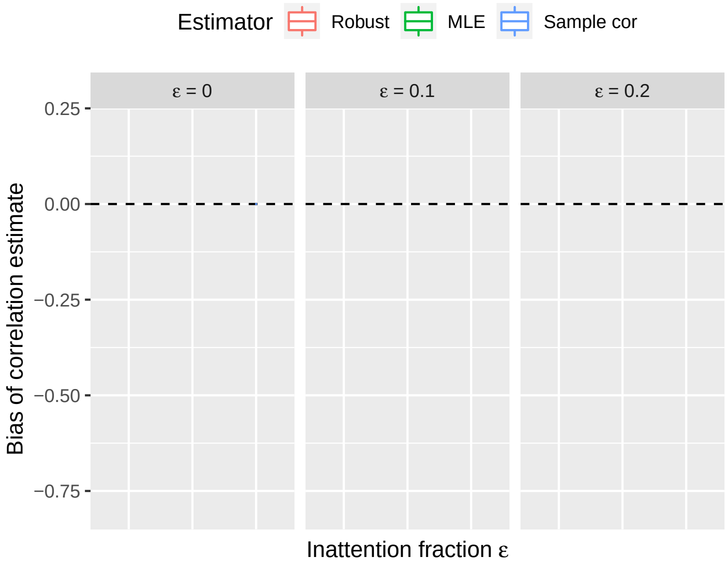
$$Z_j = \begin{cases} 1 & \text{if } \xi_j < \tau_{j,1}, \\ 2 & \text{if } \tau_{j,1} \leq \xi_j < \tau_{j,2}, \\ 3 & \text{if } \tau_{j,2} \leq \xi_j < \tau_{j,3}, \\ \vdots & \\ K_j & \text{if } \tau_{j,K_j-1} \leq \xi_j, \end{cases}$$



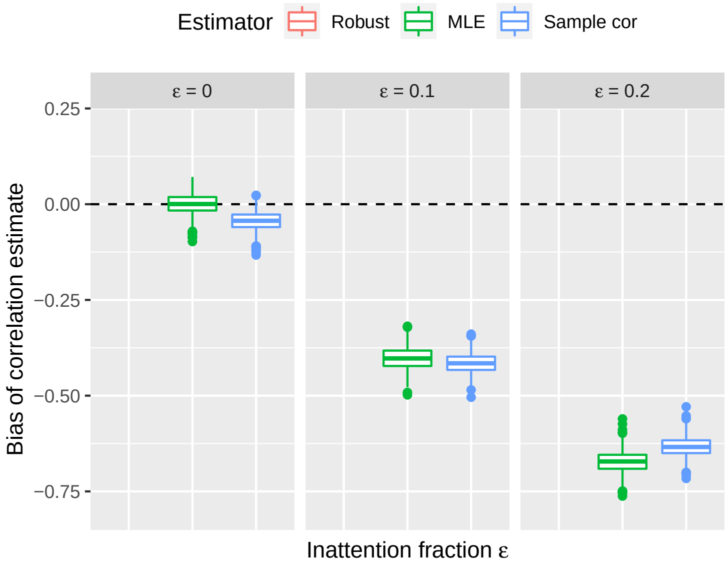
Simulation: Correlation in bivariate 5-point-scale more

- Sample $N = 1,000$ responses to $k = 2$ questions with 5 answer categories
- True $\rho_* = \text{Cor}[\xi_1, \xi_2] = 0.5$, thresholds $\boldsymbol{\tau}_* = (-1.5, -0.5, 0.5, 1.5)^\top$
- Estimate $\rho_* = 0.5$ with robust estimator, MLE, and sample correlation
- What happens to estimates if a fraction ε is inattentive? Repeat 1,000 times details

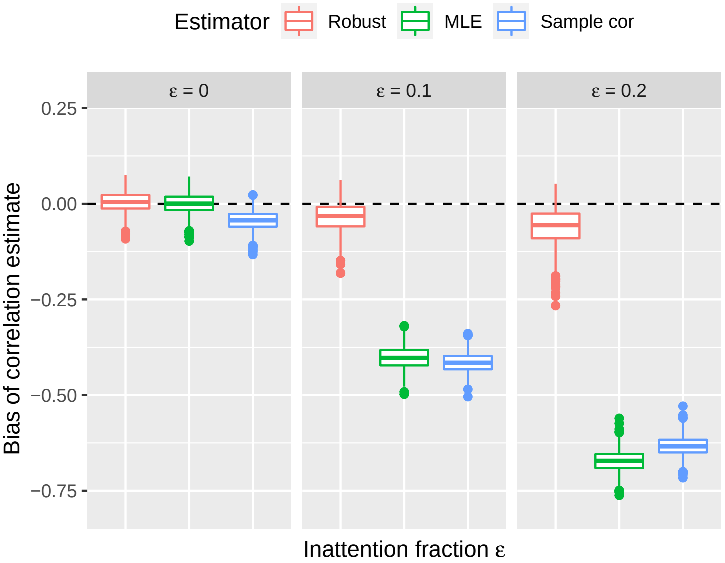
Simulation: Correlation in bivariate 5-point-scale more



Simulation: Correlation in bivariate 5-point-scale more



Simulation: Correlation in bivariate 5-point-scale more



Empirical application: Inattention in Big Five

Use data in Arias et al. (2020, BRM) of 100 unipolar markers of Big 5

- Traits measured by pairs of opposite adjectives (e.g. “talkative" vs. “silent")
- 5-point Likert scale on agreement with each adjective item
- Theory expects strong negative correlation between opposite adjectives
- $N = 725$, but some are probably inattentive (Arias et al., 2020)

Empirical application: Inattention in Big Five

Use data in Arias et al. (2020, BRM) of 100 unipolar markers of Big 5

- Traits measured by pairs of opposite adjectives (e.g. "talkative" vs. "silent")
- 5-point Likert scale on agreement with each adjective item
- Theory expects strong negative correlation between opposite adjectives
- $N = 725$, but some are probably inattentive (Arias et al., 2020)

Empirical application: Inattention in Big Five

Use data in Arias et al. (2020, BRM) of 100 unipolar markers of Big 5

- Traits measured by pairs of opposite adjectives (e.g. “talkative" vs. “silent")
- 5-point Likert scale on agreement with each adjective item
- Theory expects strong negative correlation between opposite adjectives
- $N = 725$, but some are probably inattentive (Arias et al., 2020)

Empirical application: Inattention in Big Five

Use data in Arias et al. (2020, BRM) of 100 unipolar markers of Big 5

- Traits measured by pairs of opposite adjectives (e.g. “talkative" vs. “silent")
- 5-point Likert scale on agreement with each adjective item
- Theory expects strong negative correlation between opposite adjectives
- $N = 725$, but some are probably inattentive (Arias et al., 2020)

Empirical application: Inattention in Big Five

Use data in Arias et al. (2020, BRM) of 100 unipolar markers of Big 5

- Traits measured by pairs of opposite adjectives (e.g. “talkative" vs. “silent")
- 5-point Likert scale on agreement with each adjective item
- Theory expects strong negative correlation between opposite adjectives
- $N = 725$, but some are probably inattentive (Arias et al., 2020)

Data of Arias et al. (2020, BRM): unipolar markers; $N = 725$

Correlation btw. *Neuroticism* adjective pair "envious" vs. "not envious":

	Sample cor		MLE		Robust	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
ρ	−0.562	0.031	−0.618	0.025	−0.925	0.062
$\tau_{1,1}$			−1.370	0.061	−1.570	0.276
$\tau_{1,2}$			−0.476	0.043	−0.560	0.203
$\tau_{1,3}$			0.121	0.042	0.109	0.187
$\tau_{1,4}$			1.060	0.054	1.080	0.105
$\tau_{2,1}$			−0.857	0.049	−0.905	0.073
$\tau_{2,2}$			−0.004	0.041	−0.040	0.091
$\tau_{2,3}$			0.608	0.045	0.640	0.364
$\tau_{2,4}$			1.580	0.071	1.171	0.811

Data of Arias et al. (2020, BRM): unipolar markers; $N = 725$

Correlation btw. *Neuroticism* adjective pair "envious" vs. "not envious":

	Sample cor		MLE		Robust	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
ρ	−0.562	0.031	−0.618	0.025	−0.925	0.062
$\tau_{1,1}$			−1.370	0.061	−1.570	0.276
$\tau_{1,2}$			−0.476	0.043	−0.560	0.203
$\tau_{1,3}$			0.121	0.042	0.109	0.187
$\tau_{1,4}$			1.060	0.054	1.080	0.105
$\tau_{2,1}$			−0.857	0.049	−0.905	0.073
$\tau_{2,2}$			−0.004	0.041	−0.040	0.091
$\tau_{2,3}$			0.608	0.045	0.640	0.364
$\tau_{2,4}$			1.580	0.071	1.171	0.811

Data of Arias et al. (2020, BRM): unipolar markers; $N = 725$

Correlation btw. *Neuroticism* adjective pair "envious" vs. "not envious":

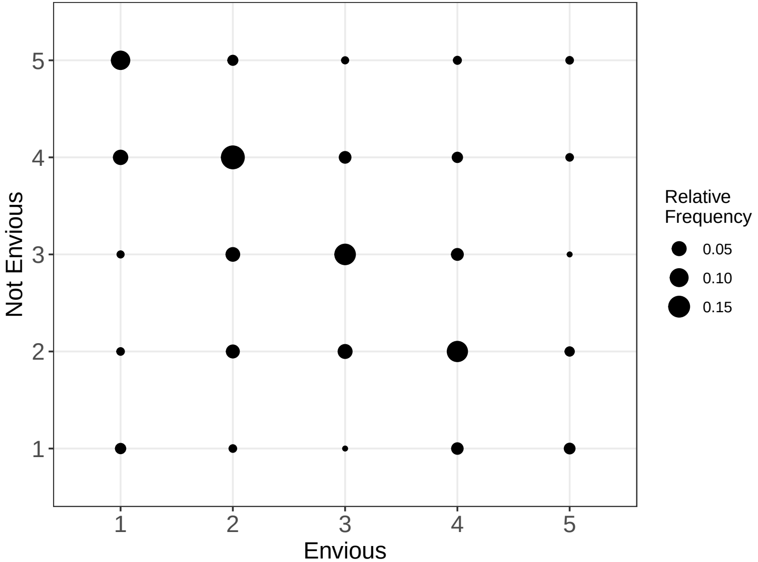
	Sample cor		MLE		Robust	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
ρ	−0.562	0.031	−0.618	0.025	−0.925	0.062
$\tau_{1,1}$			−1.370	0.061	−1.570	0.276
$\tau_{1,2}$			−0.476	0.043	−0.560	0.203
$\tau_{1,3}$			0.121	0.042	0.109	0.187
$\tau_{1,4}$			1.060	0.054	1.080	0.105
$\tau_{2,1}$			−0.857	0.049	−0.905	0.073
$\tau_{2,2}$			−0.004	0.041	−0.040	0.091
$\tau_{2,3}$			0.608	0.045	0.640	0.364
$\tau_{2,4}$			1.580	0.071	1.171	0.811

Data of Arias et al. (2020, BRM): unipolar markers; $N = 725$

Correlation btw. *Neuroticism* adjective pair "envious" vs. "not envious":

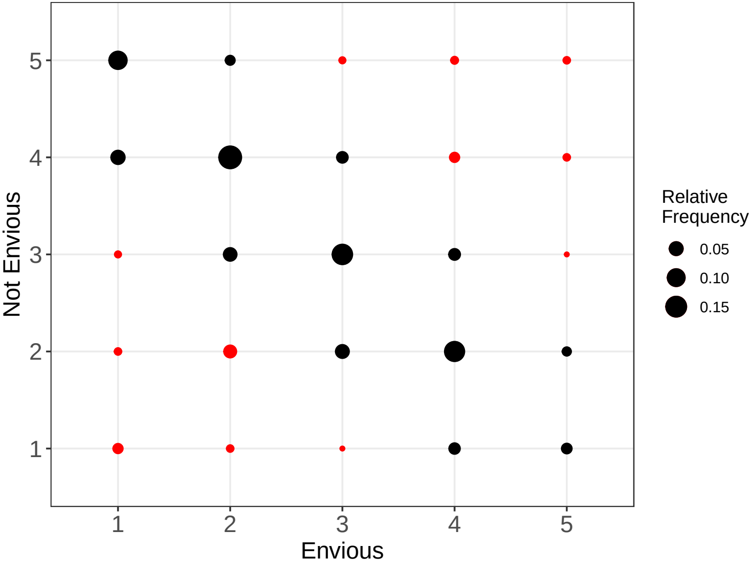
	Sample cor		MLE		Robust	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
ρ	−0.562	0.031	−0.618	0.025	−0.925	0.062
$\tau_{1,1}$			−1.370	0.061	−1.570	0.276
$\tau_{1,2}$			−0.476	0.043	−0.560	0.203
$\tau_{1,3}$			0.121	0.042	0.109	0.187
$\tau_{1,4}$			1.060	0.054	1.080	0.105
$\tau_{2,1}$			−0.857	0.049	−0.905	0.073
$\tau_{2,2}$			−0.004	0.041	−0.040	0.091
$\tau_{2,3}$			0.608	0.045	0.640	0.364
$\tau_{2,4}$			1.580	0.071	1.171	0.811

Frequency of “envious” vs. “not envious” in Arias et al. (2020)



details

Frequency of “envious” vs. “not envious” in Arias et al. (2020)



details

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to *M*-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes... $\Longrightarrow$ possibly new research line!

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to M-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes... $\Longrightarrow$ possibly new research line!

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to M-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes... $\Longrightarrow$ possibly new research line!

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to M-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes. . . $\Longrightarrow$ possibly new research line!

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to M-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes... $\Longrightarrow$ possibly new research line!

Discussion and conclusion

- Developed robust estimator for categorical data
- Generalizes MLE, categorical analogue to M-estimation
- Proposed diagnostic test to identify "outlying" cells (omitted) details
- R package `robcat` will be on CRAN soon: https://github.com/mwelz/robcat
- Relevant special cases: SEMs, reliability coefficients, counting processes. . . $\Longrightarrow$ possibly new research line!

Thank you! QR code to the paper, Welz (2024):

References I

Arias, V. B., Garrido, L., Jenaro, C., Martinez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505. https://doi.org/https://doi.org/10.3758/s13428-020-01401-8

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, *22*(2), 1081–1114. https://doi.org/10.1214/aos/1176325512

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://psycnet.apa.org/doi/10.1037/a0028085

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

References II

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Ruckstuhl, A. F., & Welsh, A. H. (2001). Robust fitting of the binomial model. *Annals of Statistics*, *29*(4), 1117–1136. https://doi.org/10.1214/aos/1013699996

Ward, M., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*(1), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Welz, M. (2024). Robust estimation and inference in categorical data [arXiv:2403.11954]. https://doi.org/10.48550/arXiv.2403.11954

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://psycnet.apa.org/doi/10.1007/s10862-005-9004-7

Robust choice of $\rho(\cdot)$ (Ruckstuhl & Welsh, 2001, AoS)

For $x = \widehat{f}_N(\boldsymbol{z})/p_{\boldsymbol{z}}(\boldsymbol{\theta})$ a Pearson residual at $\boldsymbol{\theta}$, use function

$$\rho(x) = \begin{cases} x \log(x) & \text{if } x \in [0, c], \\ x(\log(c) + 1) - c & \text{if } x > c, \end{cases}$$

where the constant $c \in [1, \infty]$ is prespecified

- If $x \in [0, c]$: Good fit, loss behaves like MLE $\Longrightarrow$ no need to downweight
- If $x > c$: Bad fit $\Longrightarrow$ downweight influence to be linear
- Similar idea as Huber loss: linear loss in tails, quadratic in center

Assumptions asymptotics

Assumption Set 1

A1. $c \in [1, +\infty]$,

A2. $\boldsymbol{\Theta} \subset \mathbb{R}^d$ *is compact,*

A3. $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}, f_\varepsilon)$ *is a unique global minimum, and* $\boldsymbol{\theta}_0 \in \operatorname{int} \boldsymbol{\Theta}$*, where* $f_\varepsilon(z) = (1 - \varepsilon)p_{\mathbf{z}}(\boldsymbol{\theta}_*) + \varepsilon h(z)$ *is the sampling distribution,*

A4. $p_{\mathbf{z}}(\boldsymbol{\theta})$ *is continuously differentiable with respect to* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and twice differentiable at* $\boldsymbol{\theta}_0$*, for all cells* $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$*,*

A5. $\left\| \frac{\partial p_{\mathbf{z}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| < \infty$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}, \mathbf{z} \in \boldsymbol{\mathcal{Z}}$*,*

A6. $p_{\mathbf{z}}(\boldsymbol{\theta}) > 0$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$*.*

Assumptions (cont'd) asymptotics

Assumption Set 1

A7. $\#\{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}} : f_\varepsilon(\boldsymbol{z}) > 0\} > d$,

A8. $L(\boldsymbol{\theta}, f_\varepsilon)$ *is convex in a neighborhood of* $\boldsymbol{\theta}_0$,

A9. $\frac{f_\varepsilon(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta}_0)} \neq c$ *for any* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$.

Consistency

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} \rho\left(\frac{\widehat{f}_N(\boldsymbol{z})}{p_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) p_{\boldsymbol{z}}(\boldsymbol{\theta}) \quad \rho(x) = \begin{cases} x\log(x) & \text{if } x \in [0, c], \\ x(\log(c)+1) - c & \text{if } x > c. \end{cases}$$

Theorem (Consistency)

Under Assumptions A1–A6, it holds true that

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0,$$

as $N \to \infty$.

Asymptotic normality

Theorem (Asymptotic normality)

Under Assumption Set 1, it holds that

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}_d\Big(0, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\Big),$$

as $N \to \infty$, where

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \boldsymbol{M}(\boldsymbol{\theta})^{-1}\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{M}(\boldsymbol{\theta})^{-1}.$$

$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ can be consistently estimated by plug-in principle.

Diagnostic test for identifying outlying cells

Is the model misspecified for an *individual* event $\boldsymbol{z} \in \mathcal{Z}$? Test formulation:

$$\mathrm{H}_0 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) = f_{\varepsilon}(\boldsymbol{z}) \qquad \text{vs.} \qquad \mathrm{H}_1 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) < f_{\varepsilon}(\boldsymbol{z})$$

Corollary (Limit distribution of test statistic)

Under $\mathrm{H}_0 : p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) = f_{\varepsilon}(\boldsymbol{z})$ and the assumptions of Theorem 2, the test statistic

$$T_N(\boldsymbol{z}) = \frac{p_{\boldsymbol{z}}\left(\widehat{\boldsymbol{\theta}}_N\right) - f_{\varepsilon}(\boldsymbol{z})}{\sqrt{\sigma^2_{\boldsymbol{z}}\left(\boldsymbol{\theta}_0\right) / N}}$$

converges to $\mathrm{N}(0,1)$ as $N \to \infty$, where $\sigma^2_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \left(\frac{\partial p_{\boldsymbol{z}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) \left(\frac{\partial p_{\boldsymbol{z}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$.

That is, the test rejects if the model is misspecified for $\boldsymbol{z}$, and otherwise does not

Definitions used in theorems asymptotics

$$\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{f}_\varepsilon) - \boldsymbol{f}_\varepsilon \boldsymbol{f}_\varepsilon^\top \qquad \text{[estimable via } \widehat{f}_N\text{]}$$

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{W}(\boldsymbol{\theta}) \boldsymbol{\Omega} \boldsymbol{W}(\boldsymbol{\theta})^\top,$$

$$\boldsymbol{M}(\boldsymbol{\theta}) = \sum_{z \in \mathcal{Z}} f_\varepsilon(z) \left(w' \left(\frac{f_\varepsilon(z)}{p_z(\boldsymbol{\theta}_0)} \right) \frac{f_\varepsilon(k)}{p_z(\boldsymbol{\theta})} \boldsymbol{s}_z(\boldsymbol{\theta}) \boldsymbol{s}_z(\boldsymbol{\theta})^\top - w \left(\frac{f_\varepsilon(z)}{p_z(\boldsymbol{\theta})} \right) \boldsymbol{Q}_z(\boldsymbol{\theta}) \right),$$

$$\boldsymbol{s}_z(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p_z(\boldsymbol{\theta}) = \frac{1}{p_z(\boldsymbol{\theta})} \left(\frac{\partial}{\partial \boldsymbol{\theta}} p_z(\boldsymbol{\theta}) \right),$$

$$\boldsymbol{W}(\boldsymbol{\theta}) = \left(\boldsymbol{s}_{z_1}(\boldsymbol{\theta}) \mathbb{1} \left\{ \frac{f_\varepsilon(z_1)}{p_{z_1}(\boldsymbol{\theta})} \in [0, c] \right\}, \cdots, \boldsymbol{s}_{z_m}(\boldsymbol{\theta}) \mathbb{1} \left\{ \frac{f_\varepsilon(z_m)}{p_{z_m}(\boldsymbol{\theta})} \in [0, c] \right\} \right),$$

$$w(x) = \mathbb{1}\{x \in [0, c]\} + c \mathbb{1}\{x > c\} / x,$$

$$\boldsymbol{Q}_z(\boldsymbol{\theta}) = \frac{1}{p_z(\boldsymbol{\theta})} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} p_z(\boldsymbol{\theta}) \right) - \boldsymbol{s}_z(\boldsymbol{\theta}) \boldsymbol{s}_z(\boldsymbol{\theta})^\top.$$

2nd order equivalence with MLE at true model back

Lemma

Under Assumption Set 1, it holds true that

$$\lim_{\varepsilon \downarrow 0} \boldsymbol{U}\left(\boldsymbol{\theta}_0(\varepsilon)\right) = \boldsymbol{J}\left(\boldsymbol{\theta}_*\right) \qquad \textit{and}$$

$$\lim_{\varepsilon \downarrow 0} \boldsymbol{M}\left(\boldsymbol{\theta}_0(\varepsilon)\right) = \begin{cases} \boldsymbol{J}\left(\boldsymbol{\theta}_*\right) & \textit{if } c > 1 \\ \boldsymbol{J}\left(\boldsymbol{\theta}_*\right) - \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{1}\left\{h(\boldsymbol{z}) > p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right)\right\} p_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right) \boldsymbol{s}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right) \boldsymbol{s}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_*\right)^\top & \textit{if } c = 1 \end{cases}$$

This lemma implies that at the true model, the asymptotic covariances of $\widehat{\boldsymbol{\theta}}_N$ and $\widehat{\boldsymbol{\theta}}_N^{\mathrm{MLE}}$ coincide

Influence function (1/3) back

Definition (Influence function)

Let $\widehat{\boldsymbol{\theta}}_N$ be an estimator that estimates a model $\{\boldsymbol{p}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with finite support $\boldsymbol{\mathcal{Z}}$. Evaluated at model density $p_{\boldsymbol{z}}(\boldsymbol{\theta}), \theta \in \boldsymbol{\theta}$, the estimator's *influence function* at a data point $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ is given by

$$\mathrm{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N, \boldsymbol{p}(\boldsymbol{\theta})\right) = \lim_{\varepsilon \downarrow 0} \frac{\widehat{\boldsymbol{\theta}}_N\big((1-\varepsilon)p_{\boldsymbol{z}}(\boldsymbol{\theta}) + \varepsilon \Delta_{\boldsymbol{z}}\big)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} \widehat{\boldsymbol{\theta}}_N\big((1-\varepsilon)p_{\boldsymbol{z}}(\boldsymbol{\theta}) + \varepsilon \Delta_{\boldsymbol{z}}\big)\Big|_{\varepsilon=0},$$

where $\boldsymbol{\mathcal{Z}} \ni \boldsymbol{y} \mapsto \Delta_{\boldsymbol{z}}(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{z} = \boldsymbol{y}\}$ is the point mass density at point $\boldsymbol{z}$.

Influence function (2/3) back

Theorem (Influence function)

Grant Assumption Set 1. Then, the influence function of estimator $\widehat{\boldsymbol{\theta}}_N$ at cell $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ and true density $\boldsymbol{p}(\boldsymbol{\theta}_*)$ is given by

$$\mathrm{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N, \boldsymbol{p}(\boldsymbol{\theta}_*)\right) = \begin{cases} \mathrm{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N^{\mathrm{MLE}}, \boldsymbol{p}(\boldsymbol{\theta}_*)\right) & \text{if } c > 1, \\ \left[\boldsymbol{J}(\boldsymbol{\theta}_*) - p_{\boldsymbol{z}}(\boldsymbol{\theta}_*)\, \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_*)\, \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_*)^\top\right]^{-1} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}_*)\, p_{\boldsymbol{z}}(\boldsymbol{\theta}_*) & \text{if } c = 1, \end{cases}$$

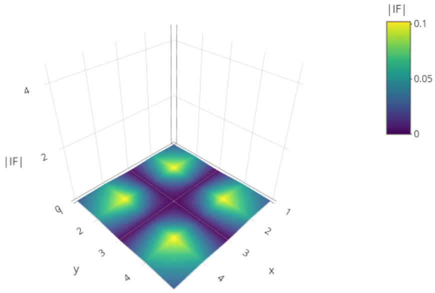
where

$$\mathrm{IF}\left(\boldsymbol{z}, \widehat{\boldsymbol{\theta}}_N^{\mathrm{MLE}}, \boldsymbol{p}(\boldsymbol{\theta})\right) = \boldsymbol{J}(\boldsymbol{\theta})^{-1} \boldsymbol{s}_{\boldsymbol{z}}(\boldsymbol{\theta}),$$

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{\theta}) p_{\boldsymbol{z}}(\boldsymbol{\theta})$$

Influence function (3/3) back

(a) $c > 1$



(b) $c = 1$

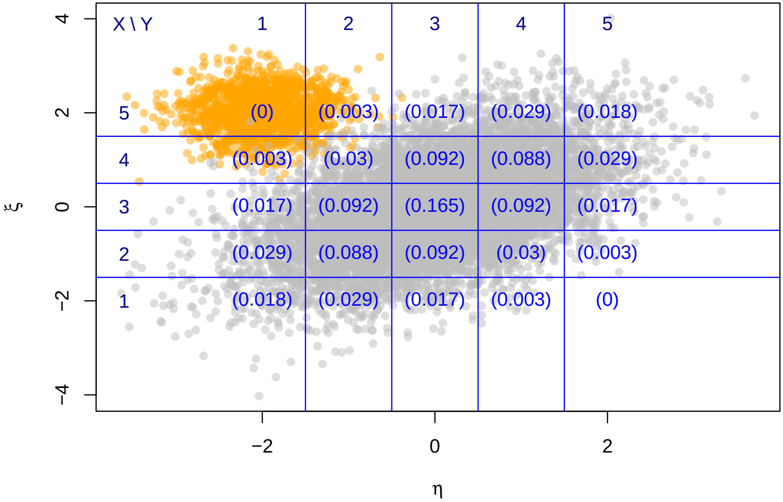
Simulation design (1/2) back

For misspecification fraction ε,

- Sample fraction $1 - \varepsilon$ from bivariate standard normal with $\rho_* = 0.5$
- Sample fraction ε from $\mathrm{N}_2\left((2, -2)^\top, \boldsymbol{I}\right)$
- Use discretization process to obtain (Z_1, Z_2)

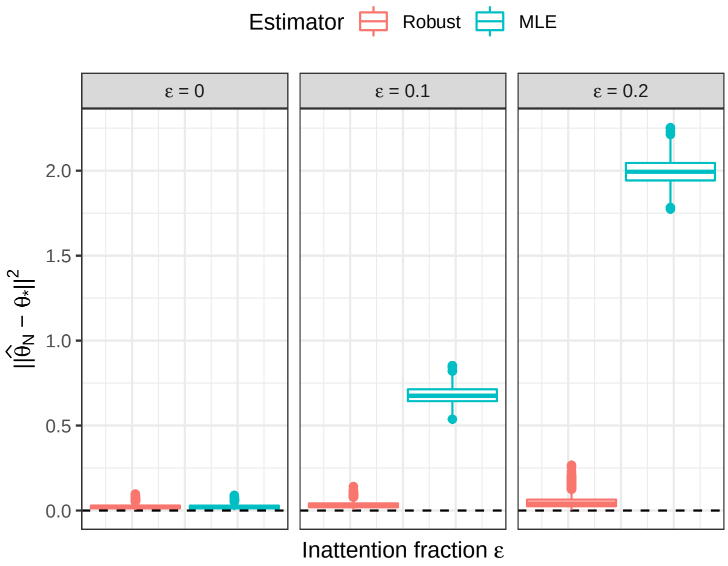
$\Longrightarrow$ Cells $(Z_1, Z_2) = (5, 1), (5, 2), (4, 1)$ are inflated, but some overlap with model distribution

Simulation design (2/2)

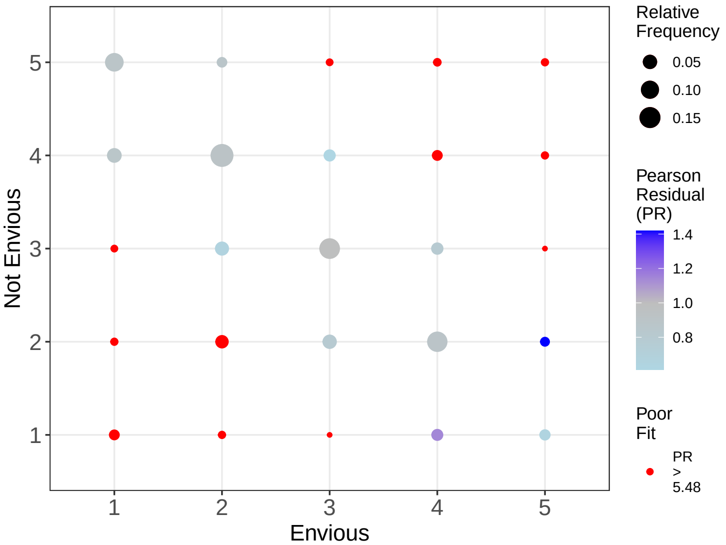


Simulation: Correlation of bivariate 5-point scale

back

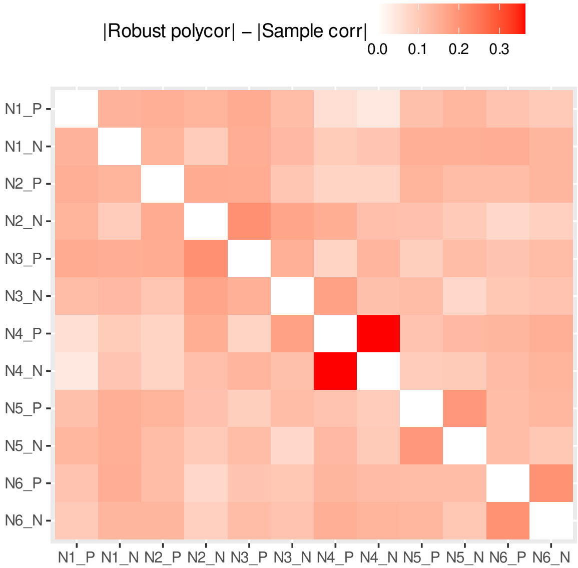


Analysis of "envious" vs. "not envious" in Arias et al. (2020)



back

Estimated correlation matrix for the *Neuroticism* scale



- Robustly estimated correlations are all stronger (ave. 0.130)
- MLE polyhoric correlations are similar to sample correlations (expected; cf. Rhemtulla et al., 2012)

back

Loadings of the neuroticism factor in data of Arias et al. (2020)

Item	Sample corr	Robust polycor
N1_P	0.70	0.80
N1_N	0.56	0.66
N2_P	0.76	0.86
N2_N	0.68	0.78
N3_P	0.77	0.88
N3_N	0.66	0.74
N4_P	0.35	0.46
N4_N	0.46	0.54
N5_P	0.69	0.77
N5_N	0.67	0.73
N6_P	0.57	0.66
N6_N	0.64	0.71
Proportion variance	0.40	0.53
Cronbach's α	0.89	0.93