Modeling relationships from themes in text and covariates with an outcome

A Bayesian supervised topic model with covariates

Kenneth Tyler Wilcox

Ross Jacobucci

Zhiyong Zhang

Department of Psychology, University of Notre Dame

2020/05/26

Text Data in Psychology

- Text is an increasingly popular data source
    - Social media (Schwartz et al., 2013)
    - Open-ended questions (Popping, 2015)
    - Medical health records (Obeid et al., 2019)
- Various overviews exist on existing text mining algorithms for psychological research (Finch et al., 2018; Iliev et al., 2015; Kjell et al., 2019; Rohrer et al., 2017)
- These algorithms are often designed for large data sets
- Current challenge is to adapt these algorithms to psychological research

Modeling Text as Data

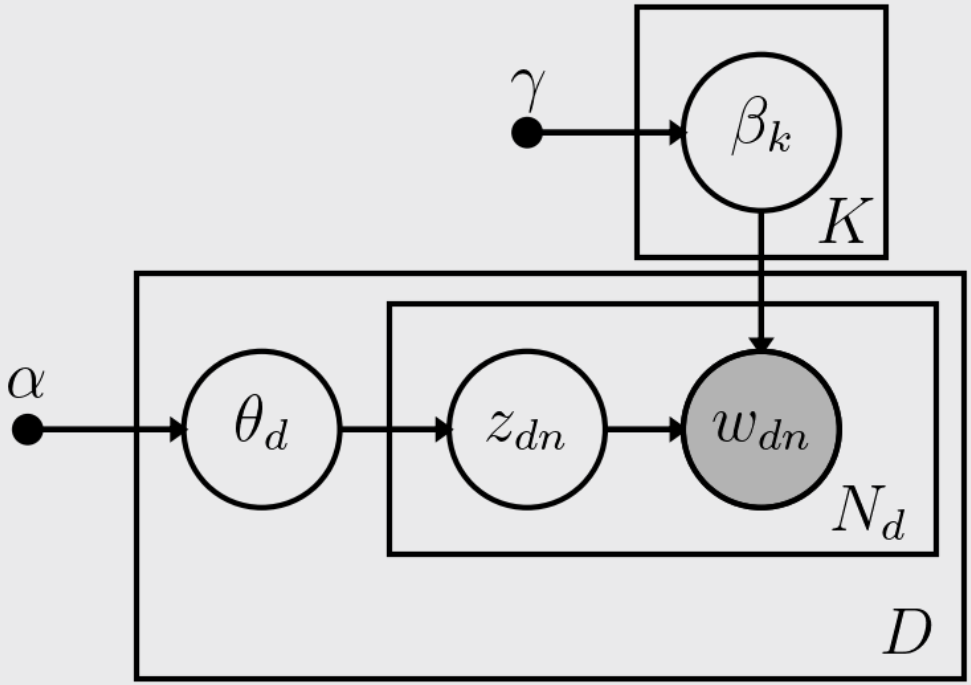
Top Down

- Dictionary methods
    - LIWC (Tausczik et al., 2010)
    - Sentiment analysis
- Dictionaries may not be valid for given data

Bottom Up

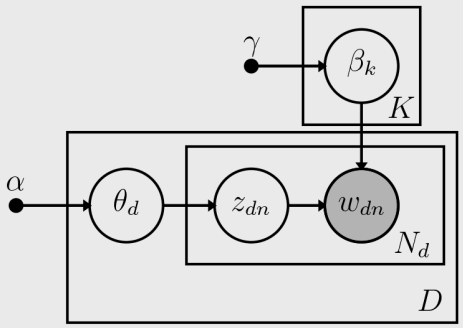
- Qualitative analysis
    - Gold standard
    - Time-consuming and expensive
    - Hard to reuse
- Quantitative models
    - Faster and cheaper
    - Reusable

Topic Modeling



Latent Dirichlet Allocation (LDA)

Seminal topic model (Blei et al., 2003)



$$L(\vec{\Theta}, \vec{B}) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{z_{dn}, w_{dn}} \theta_{d, z_{dn}}$$

Topics:

$$\vec{\beta}_k = \Pr\left[w_{dn} = m | z_{dn} = k\right]$$
$$\vec{\beta}_k \sim \mathrm{Dir}(\vec{\gamma})$$

Topic proportions:

$$\vec{\theta}_d = \Pr\left[z_{dn} = k\right]$$
$$\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$$

Topic assignments:

$$\left(z_{dn} | \vec{\theta}_d\right) \sim \mathrm{Cat}(\vec{\theta}_d)$$

Words:

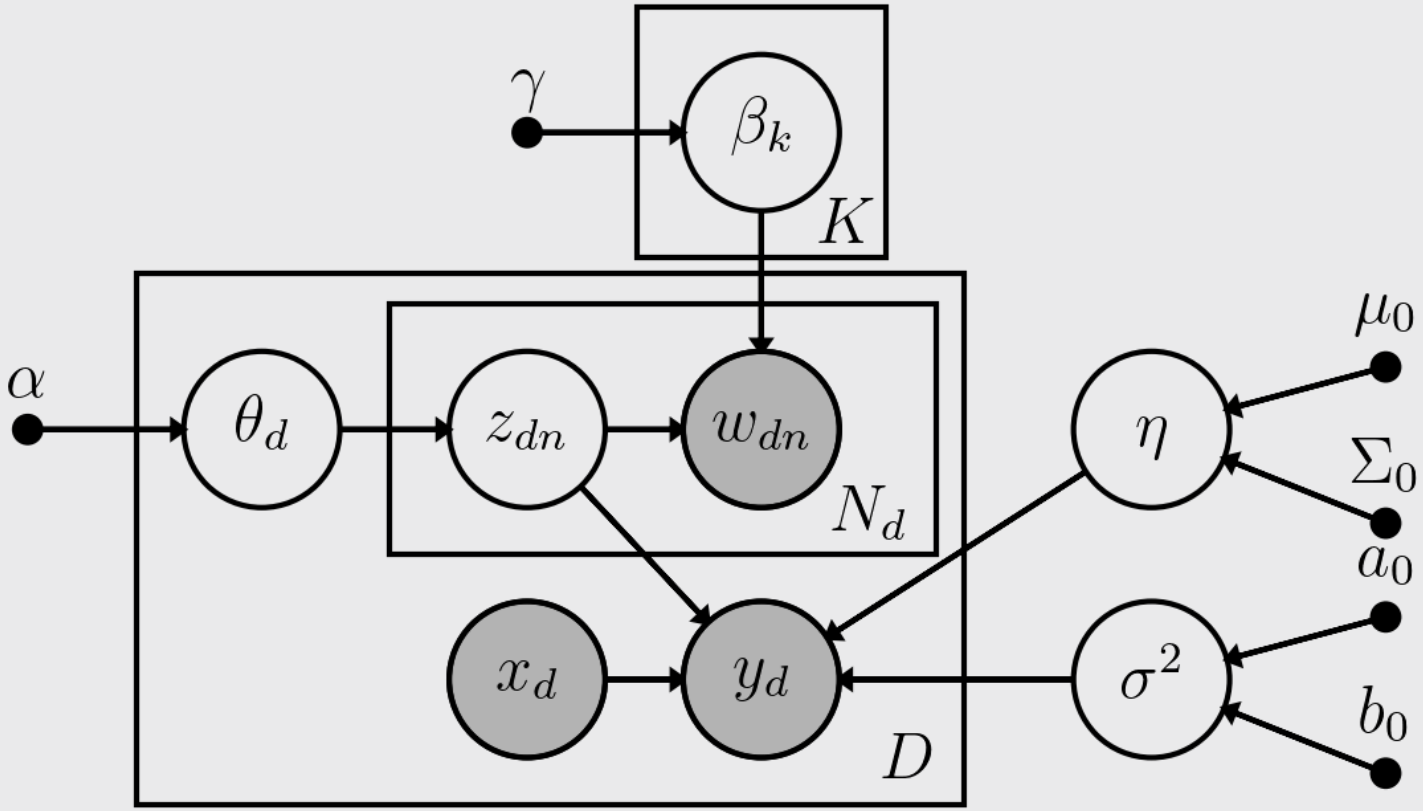
$$\left(w_{dn} | z_{dn} = k, \vec{\beta}_k\right) \sim \mathrm{Cat}(\vec{\beta}_k)$$

Fusing Topic Models and Regression

$$Y = \boxed{\phantom{\vec{\Theta}}}\eta + X\beta + \epsilon$$

- Two-stage approach (Packard et al., 2020; Rohrer et al., 2017)
  - Use estimated $\vec{\Theta}$ to predict Y
  - Could include other manifest predictors $\vec{X}$
- One-stage approach
  - Supervised topic model (SLDA; Blei et al., 2010)
  - Does not include $\vec{X}$
- We propose the SLDAX model
  - One-stage approach
  - Allow topics and manifest predictors of Y

SLDAX


γ
β_k
K
α
θ_d
z_{dn}
w_{dn}
N_d
μ_0
η
Σ_0
a_0
x_d
y_d
σ^2
b_0
D

Gibbs Sampler for $Y|\cdot \sim \mathrm{N}(\cdot)$

- $f\left(z_{d,n}=k|\cdot\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(y_d-\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right)^2\right\}\times \left(n_{dk}^{(-n)}+\alpha\right)\left(\frac{n_{kv}^{(-n)}+\gamma}{n_k^{(-n)}+V\gamma}\right)$
- $f\left(\sigma^2|\cdot\right)=\mathrm{IG}\left(\frac{a_0+D}{2},\frac{1}{2}\left(b_0+\sum_d\left[y_d-\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right]^2\right)\right)$
- $f\left(\vec{\eta}|\cdot\right)=\mathrm{N}\left(\vec{\eta}_1,\vec{\Sigma}_1\right)$
  - $\vec{\Sigma}_1=\left(\vec{\Sigma}_0^{-1}+\sigma^{-2}\left(\vec{\bar{Z}},\vec{X}\right)'\left(\vec{\bar{Z}},\vec{X}\right)\right)^{-1}$
  - $\vec{\eta}_1=\vec{\Sigma}_1\left(\vec{\Sigma}_0^{-1}\vec{\mu}_0+\sigma^{-2}\left(\vec{\bar{Z}},\vec{X}\right)'\vec{y}\right)$

MH-in-Gibbs for $Y|\cdot \sim \text{Ber}(\cdot)$

- $f\left(z_{dn}=k|\cdot\right) \propto \frac{\exp\left\{y_d\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right\}}{1+\exp\left\{\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right\}}\left(n_{dk}^{(-n)}+\alpha\right)\left(\frac{n_{kv}^{(-n)}+\gamma}{n_{k}^{(-n)}+V\gamma}\right)$
- $f\left(\vec{\eta}|\cdot\right) \propto \prod_d \left[\frac{\exp\left\{y_d\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right\}}{1+\exp\left\{\left(\vec{\bar{z}}_d,\vec{x}_d\right)'\vec{\eta}\right\}}\right] \times \exp\left\{-\frac{1}{2}(\vec{\eta}-\vec{\mu}_0)'\vec{\Sigma}_0^{-1}(\vec{\eta}-\vec{\mu}_0)\right\}$
  - Use Metropolis-Hastings algorithm to sample
  - Independent proposal distributions
    - $\eta_j \sim \text{N}\left(\mu_j,\tau_j\right)$
    - Tune τ_j during burn-in

Software

- **psychtm** R package in *early* development
- Features
  - LDA, SLDA, SLDAX MCMC algorithms implemented in **C++**
    - Normal and dichotomous outcomes supported
  - Estimation and visualization of $\vec{\Theta}$ and $\vec{B}$
  - Model selection by WAIC (Watanabe, 2010)
- Available from Github

```
devtools::install_github("ktw5691/psychtm")
```

Simulation Study

Goal

- Compare SLDAX with two-stage approach (LDA + OLS regression)
  - SLDAX from our R package **`psychtm`**
  - LDA model from R package **`topicmodels`**
- Conditions
  - # topics K: 2 and 5
  - # documents D: 200, 800, and 1500
  - Mean # words $\bar{N}_d$: 15, 80, and 150
  - Vocabulary V: 500 and 1000

Simulation Study

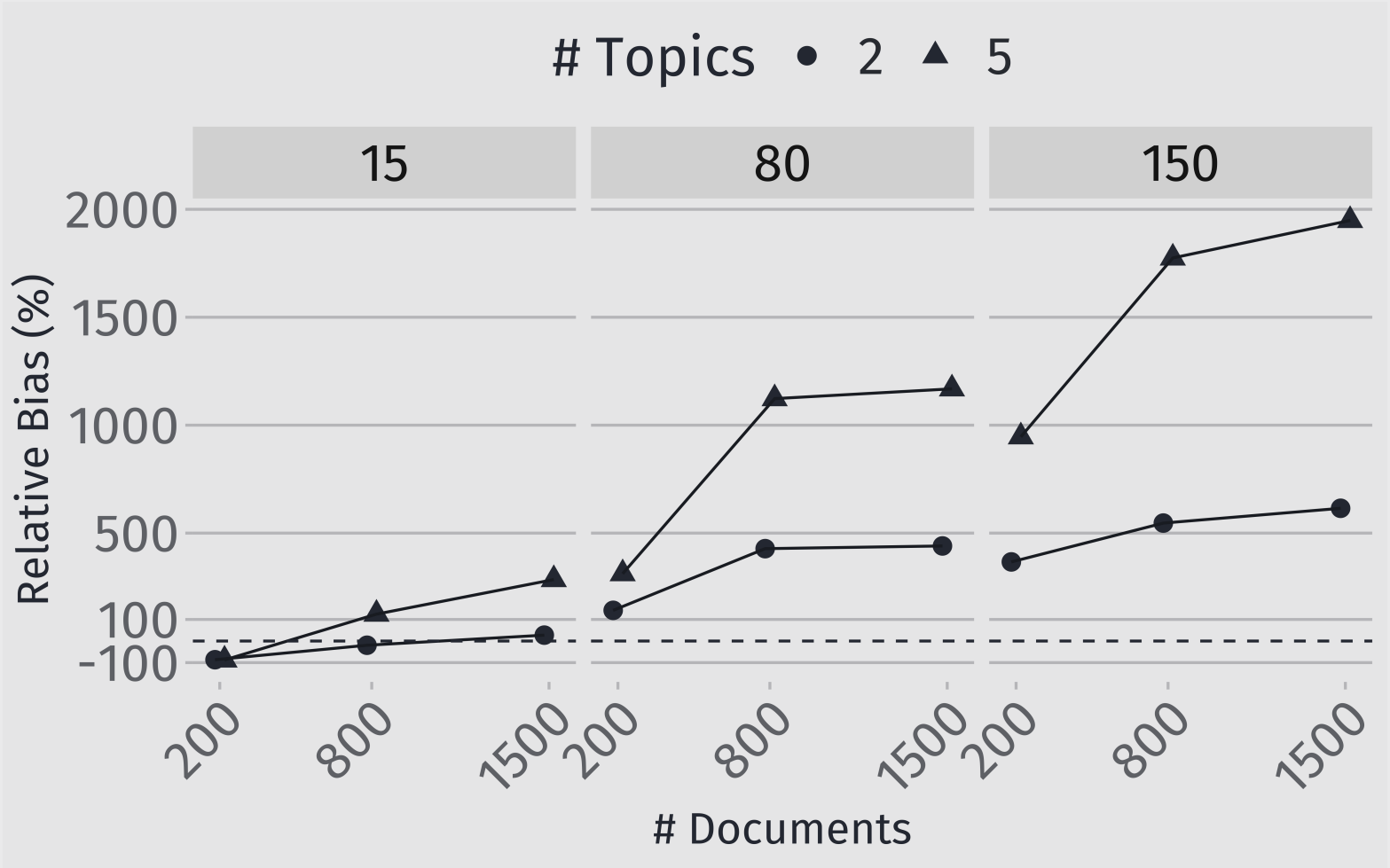
Data Generation

- SLDAX model
  - $X \sim \mathrm{N}(0, 1)$ w/ R^2 = .15
  - $Y \sim \mathrm{N}(\cdot)$
  - K topics w/ joint R^2 = .35

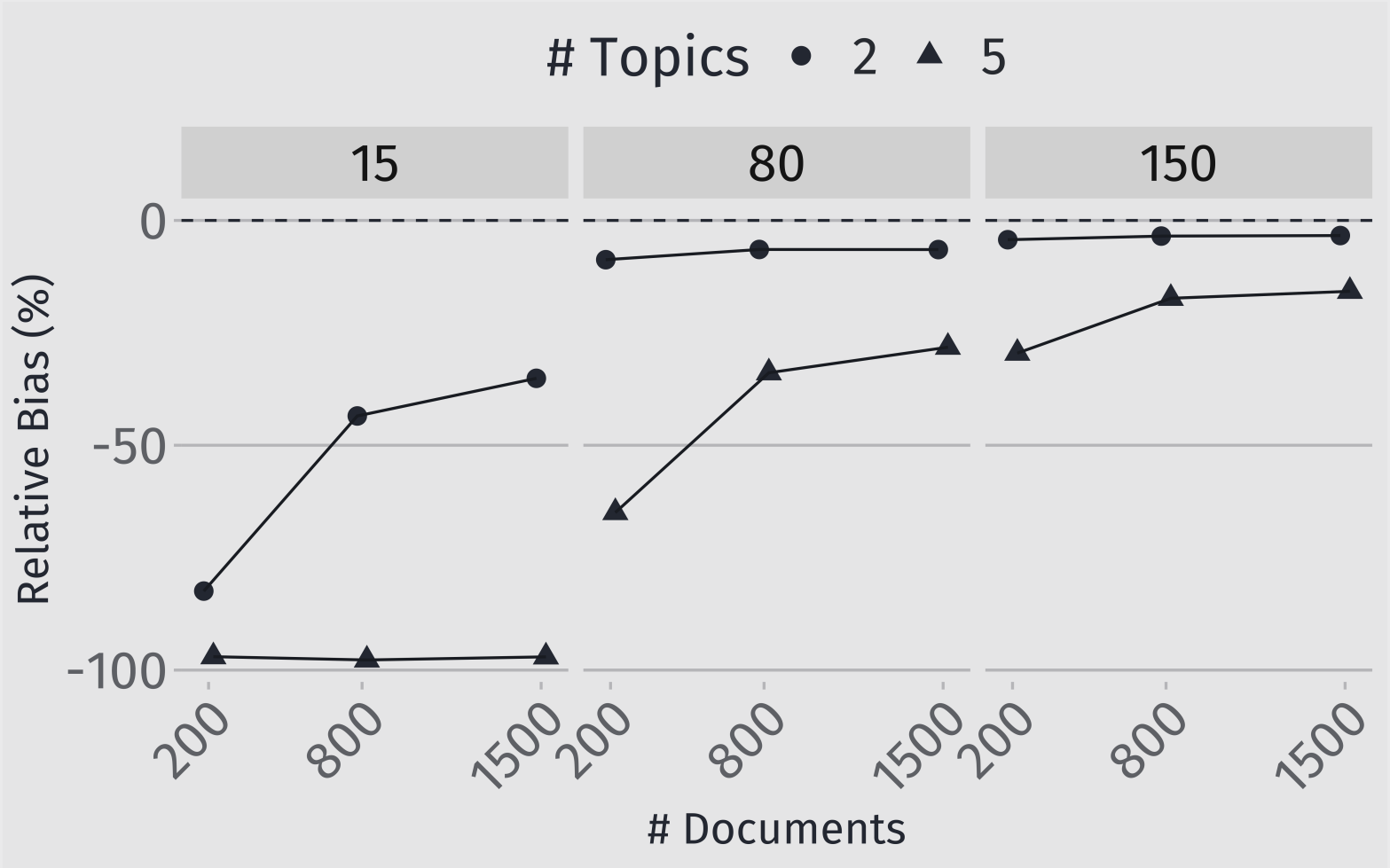
Estimation

- SLDAX with flat priors
- Two-stage
  1. LDA: estimated w/ variational EM (same hyper-parameters)
  2. OLS regression

Two-Stage Estimation Bias for $\eta_{\bar{z}}$



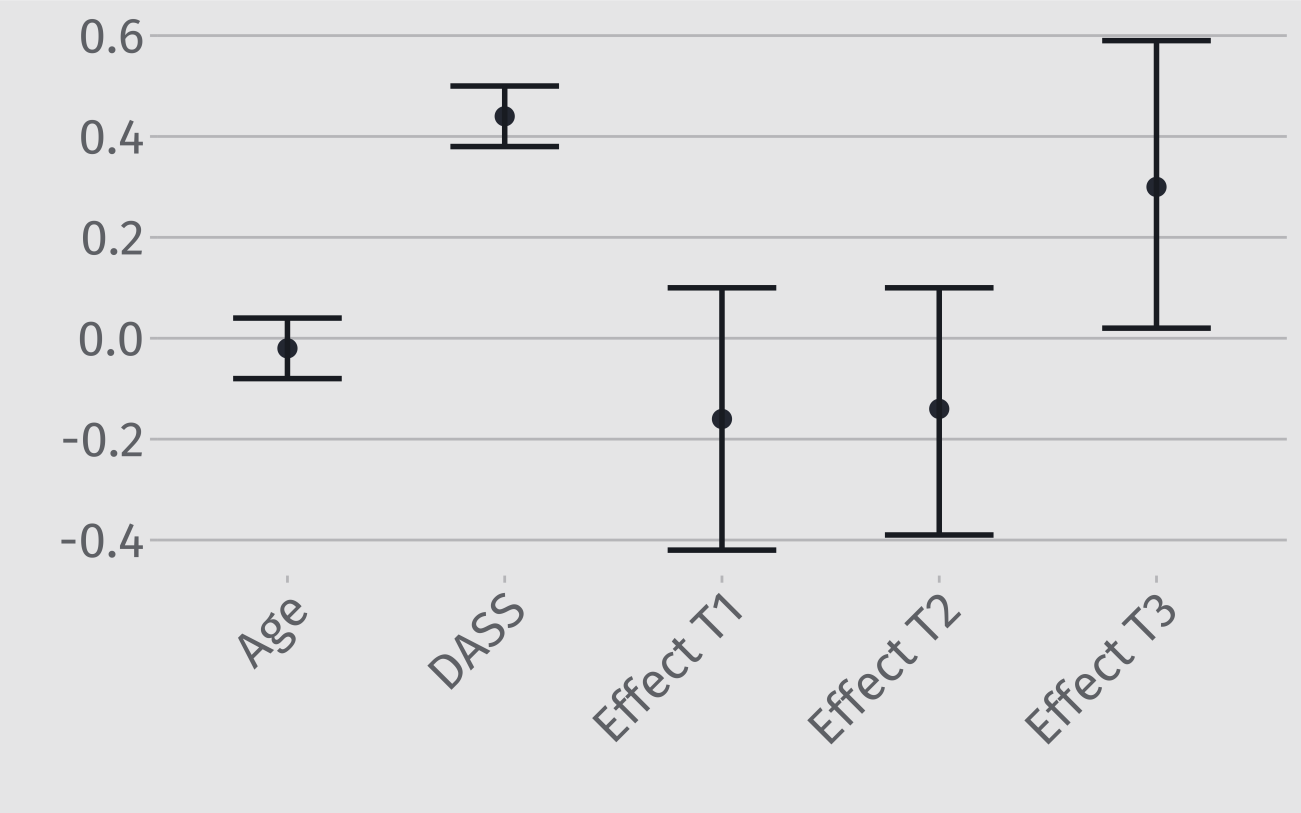
SLDAX Estimation Bias for $\eta_{\bar{z}}$



Illustrative Example

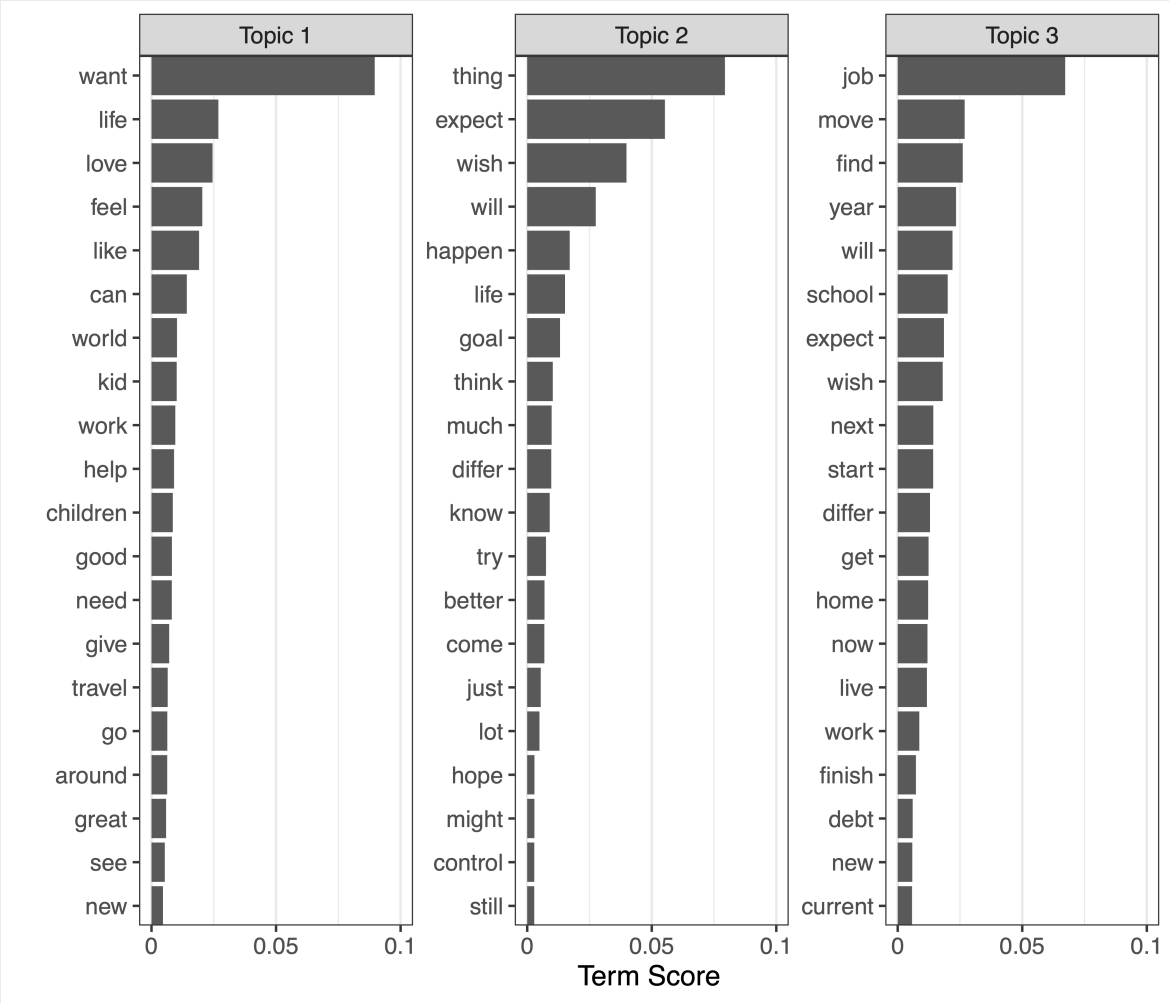
- 882 adults recruited on MTurk
- Y: Beck Hopelessness Scale (Beck et al., 1989)
- "What are your expectations for the future?"
  - *M* = 50 words, *SD* = 24, *Range* = 5 – 186
  - After stopword removal and stemming:
    - Median length was 18 words (*M* = 21, *SD* = 10, *Range* = 2 – 76)
    - Vocabulary of 3096 stems (98% of original vocabulary)
- Manifest predictors
  - Depression Anxiety Stress Scales (Lovibond et al., 1995)
  - Age (*M* = 33, *SD* = 10, *Range* = 18 – 79)

Regression Estimates



$$\text{Effect} = \hat{\eta}_k - K^{-1} \sum_{j \neq k}^{K} \hat{\eta}_j$$

Topic Estimates



Conclusions

- Hopelessness in responses associated with BHS
  - Convergent validity for topics
  - Text topics associated with BHS above and beyond DASS
- Topic effects may be attenuated based on simulation results
  - Large D, small $\bar{N}_d$
- Could predict on new data or update model using new data

Discussion

Key Findings

- We derived MCMC algorithms to estimate SLDAX models
- SLDAX models implemented in open-source **R** package
- The popular two-stage approach yields biased regression estimates
- SLDAX yields accurate estimates with shrinkage in small-data scenarios

Future Work

- SLDAX framework can be generalized
  - Integration with SEM
  - Longitudinal / EMA data
- Impact of text data quality on performance

Thanks!

kwilcox3@nd.edu

ktylerwilcox.netlify.app

@ktw5691

Slides:

ktylerwilcox.netlify.app/talk/2020-isdsa-sldax/

References

Beck AT, Brown G, Steer RA (1989). "Prediction of Eventual Suicide in Psychiatric Inpatients by Clinical Ratings of Hopelessness." *Journal of Consulting and Clinical Psychology*, *57*(2), 309-310. https://doi.org/10.1037/0022-006X.57.2.309.

Blei DM, McAuliffe JD (2010). "Supervised Topic Models." *arXiv*.

Blei DM, Ng AY, Jordan MI (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, *3*, 993-1022.

Finch WH, Finch MEH, McIntosh CE, Braun C (2018). "The Use of Topic Modeling with Latent Dirichlet Analysis with Open-Ended Survey Items." *Translational Issues in Psychological Science*, 4(4), 403-424. https://doi.org/10.1037/tps0000173.

Iliev R, Dehghani M, Sagi E (2015). "Automated Text Analysis in Psychology: Methods, Applications, and Future Developments." *Language and Cognition*, *7*(2), 265-290. https://doi.org/10.1017/langcog.2014.30.

Kjell ONE, Kjell K, Garcia D, Sikström S (2019). "Semantic Measures: Using Natural Language Processing to Measure, Differentiate, and Describe Psychological Constructs." *Psychological Methods*, *24*(1), 92-115. https://doi.org/10.1037/met0000191.

Lovibond PF, Lovibond SH (1995). "The Structure of Negative Emotional States: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories." *Behaviour Research and Therapy*, *33*(3), 335-343. https://doi.org/10.1016/0005-7967(94)00075-U.

Obeid JS, Weeda ER, Matuskowitz AJ, Gagnon K, Crawford T, Carr CM, Frey LJ (2019). "Automated Detection of Altered Mental Status in Emergency Department Clinical Notes: A Deep Learning Approach." *BMC Medical Informatics and Decision Making*, *19*(1), 164. https://doi.org/10.1186/s12911-019-0894-9.

Packard G, Berger J (2020). "Thinking of You: How Second-Person Pronouns Shape Cultural Success." *Psychological Science*. https://doi.org/10.1177/0956797620902380.

Popping R (2015). "Analyzing Open-Ended Questions by Means of Text Analysis Procedures." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *128*(1), 23-39. https://doi.org/10.1177/0759106315597389.

Roberts GO, Gelman A, Gilks WR (1997). "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *Annals of Applied Probability*, *7*(1), 110-120. https://doi.org/10.1214/aoap/1034625254.

Roberts ME, Stewart BM, Airoldi EM (2016). "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association*, *111*(515), 988-1003. https://doi.org/10.1080/01621459.2016.1141684.

Rohrer JM, Brümmer M, Schmukle SC, Goebel J, Wagner GG (2017). ""What Else Are You Worried about?" Integrating Textual Responses into Quantitative Social Science Research." *PLoS ONE*, *12*(7), e0182156. https://doi.org/10.1371/journal.pone.0182156.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman MEP, Ungar LH (2013). "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PloS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

Stephens M (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(4), 795-809.

Tausczik YR, Pennebaker JW (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology*, *29*(1), 24-54. https://doi.org/10.1177/0261927X09351676.

Watanabe S (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research*, *11*, 3571-3594.