Using Embeddings to Correct for Unobserved Confounding in Networks

Victor Veitch, Yixin Wang, David M. Blei

Columbia University

Setup

- Goal: infer (causal) average treatment effect of treatment T on outcome Y

$$\psi = \mathbb{E}[Y \mid \mathrm{do}(T=1)] - \mathbb{E}[Y \mid \mathrm{do}(T=0)]$$

- Problem: unobserved confounder Z might affect T and Y
- Salvation: we observe a proxy X for Z

Wrinkle

What if the proxy is nasty?

- the proxy has non-iid structure
  $\implies$ standard non-parametrics = :(
- no plausible generative model for $(X, \{Z_i\}, \{T_i\}, \{Y_i\})$:
  $\implies$ standard parametrics = :(
- easy to think of hacks, but do they work?

Example: Social Network

- Observe treatments and outcomes of people in a social network.
- Also observe the network.
- Possible latent confounders; e.g., age, sex, affinity for punk rock.
- These are associated with the social network itself.

If life was easy

If we observed $\{Z\}$,

- estimate $\hat{Q}_n(t,z) \approx Q(t,z) = \mathbb{E}[Y \mid t, z]$
- estimate $\hat{g}_n(z) \approx g(z) = \mathrm{P}(T=1 \mid z)$
- plug-in to (non-parametrically efficient, robust) estimator

If we had a well-specified generative model,

- estimate $\{\hat{z}_{n,i}\}$.
- plug-in as though z observed.
- works if $\hat{z}_{n,i}$ is really good estimate.

Embeddings

- Can represent network by embeddings.
- Not justified by any generative model.
- But way more useful for supervised learning.
- Can we forget estimating z, and use an embedding instead?

References

[1] L. Takac and M. Zabovsky. Data Analysis in Public Social Networks. 2012

[2] Kang et al. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. 2018

[3] V. Veitch, M. Austern, W. Zhou, D.M. Blei, and P. Orbanz. Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data. 2018.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

Estimation of Average Treatment Effect ψ

For each unit i we observe a treatment T_i and a response Y_i. Additionally, a (possibly complicated, non-iid) X as a proxy for confounders.

1. Split the units $\mathcal{I}$ into $\mathcal{I}_0, \mathcal{I} \setminus \mathcal{I}_0$
2. Jointly train embeddings λ and nuisance parameters γ. Use all data for embeddings, but only $\mathcal{I} \setminus \mathcal{I}_0$ for nuisance parameters. Example:

$$\hat{\lambda}_n, \hat{\gamma}_n = \underset{\lambda,\gamma}{\operatorname{argmin}}\, \hat{R}_n(\lambda,\gamma)$$

$$\hat{R}_n(\lambda,\gamma) = L(\lambda, X) + \frac{1}{|\mathcal{I} \setminus \mathcal{I}_0|}\Big[\sum_{i \in \mathcal{I}\setminus\mathcal{I}_0} (y_i - \tilde{Q}(t_i, \lambda_i; \gamma^Q)))^2 + \sum_{i \in \mathcal{I}\setminus\mathcal{I}_0} (t_i - \tilde{g}(\lambda_i;\gamma^g))^2\Big]$$

3. Report

$$\hat{\psi}_n^{\mathrm{A}}(I_0) := \frac{1}{|I_0|}\sum_{i \in I_0}\Bigg[\tilde{Q}(1, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q) - \tilde{Q}(0, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q) + \Big(\frac{I[t_i=1]}{\tilde{g}(\hat{\lambda}_{n,i};\hat{\gamma}_n^g)} - \frac{I[t_i=0]}{1-\tilde{g}(\hat{\lambda}_{n,i};\hat{\gamma}_n^g)}\Big)\big(y_i - \tilde{Q}(t_i, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q)\big)\Bigg].$$

Validity

Assume

- Many technical conditions
- Observing Z would render the effect identifiable
- The model learns something, eventually. For $t \in \{0,1\}$:

$$\|\mathbb{E}[Y \mid T=t, Z] - \tilde{Q}(t, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q)\|_2 \cdot \|\mathrm{P}(T=1 \mid Z) - \tilde{g}(\hat{\lambda}_{n,i}; \hat{\gamma}_n^g)\|_2 = o(\frac{1}{\sqrt{n}})$$

- Asymptotically, the embeddings are independent-ish. For all bounded continuous f with mean 0:

$$\mathbb{E}[f(\hat{\lambda}_i) f(\hat{\lambda}_j)] = o(\frac{1}{n})$$

Then,

$$\hat{\psi}_n^{\mathrm{A}} \text{ converges to the ATE } \psi \text{ at a } \sqrt{n}\text{-rate.}$$

Example: Pokec

- Semi-synthetic data based on Pokec, a large social network [1].
- Confounders are age, region, and Pokec join date.
- Treatments and outcomes simulated from a variety of models.
- Embedding model based on relational ERM [3].
- Punchline: embedding-based method is significantly more accurate than baselines.

Estimator	Linear	Trig.	High Var.	t-noise
Naive	1.56 ± 0.27	0.021 ± 0.003	79.8 ± 25.1	1.55 ± 0.28
Parametric	7.27 ± 1.67	8.75 ± 1.02	13.5 ± 3.4	8.21 ± 1.04
$\hat{\psi}_n^{\mathrm{A}}$	0.10 ± 0.02	0.011 ± 0.003	5.3 ± 1.8	0.10 ± 0.02

Table entries are mean square error of treatment effect estimate, over 25 simulations. Column heads are simulation settings. Naive does not correct for confounding. Parametric is mixed-membership stochastic block model and linear regression.