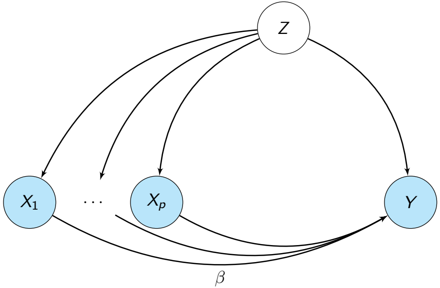
Adjusting for Multi-Cause Confounding

Jeff Adams
Ongoing work with Niels R. Hansen

Department of Mathematical Sciences
University of Copenhagen

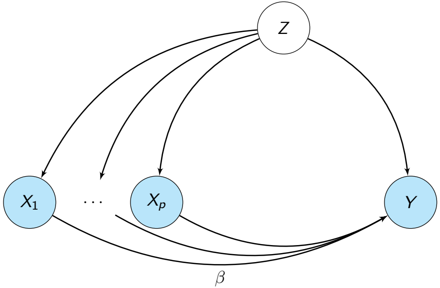
October 2022

The Confounding Model



- i.i.d. samples of p treatments X_i and a response Y.
- Goal: Estimate the treatment effect of X on Y.
- Problem: There is an unobserved confounder Z.
- Assumption: Treatments are non-adjacent.
- Assumption: Z is a "multi-cause" confounder.
- Assumption: Additive treatment effects β.

The Intuition



- If Z were measured, we could regress $Y = \hat{\beta}_j X_j + \hat{g}(Z)$ for any j of interest.
- If $f(x) = \arg\max_z p(Z = z | X = x)$ were known, then we could plug in $\hat{Z} = f(X)$.
- Since X are assumed conditionally independent given Z, maybe we can learn $\hat{f}$ from X.
- Together, the regression is $Y = \hat{\beta}_j X_j + \hat{g}(\hat{f}(X))$.

Related Work

Recent related papers include:

- Wang and Blei (2019): Advocate non-parametric estimation of Z, but give no finite-sample guarantees.
- Ogburn et al. (2020) and Grimmer et al. (2020): Critical response.
- Ćevid et al. (2020): A spectral transform and LASSO-based approach.

Step 1: Learn $\hat{f}$ by Tensor Decomposition

- Modeling choice: Assume Z is discrete in $\{1, ..., K\}$.
- Partition X into thirds: $X = (X_i, X_j, X_k)$.
- By conditional independence,

$$\mathbb{E}[X_i \otimes X_j] = \sum_{z=1}^{K} \omega^{(z)} \mu_i^{(z)} \otimes \mu_j^{(z)}$$

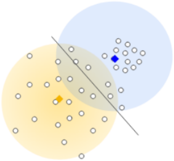
$$\mathbb{E}[X_i \otimes X_j \otimes X_k] = \sum_{z=1}^{K} \omega^{(z)} \mu_i^{(z)} \otimes \mu_j^{(z)} \otimes \mu_k^{(z)}$$

- Kruskal's Theorem tells us $\omega^{(z)}$ and $\mu^{(z)}$ are (generically) identifiable.
- We can learn $\mu^{(z)}$ and $\omega^{(z)}$ with provable sample complexities in p and K. (Anandkumar et al., 2014) (Guo et al., 2022)

Step 2: Latent Labeling

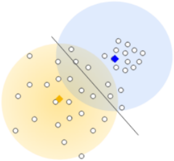
- Suppose estimates $\hat{\mu}^{(z)}$ of $\mu^{(z)}$ satisfy $\|\hat{\mu}^{(z)} - \mu^{(z)}\|_2 < \epsilon$.
- Simplest labeling algorithm: pick the nearest mean!

$$\hat{Z} = \arg\min_z \|X - \hat{\mu}^{(z)}\|_2$$



Step 2: Latent Labeling

- Suppose estimates $\hat{\mu}^{(z)}$ of $\mu^{(z)}$ satisfy $\|\hat{\mu}^{(z)} - \mu^{(z)}\|_2 < \epsilon$.
- Simplest labeling algorithm: pick the nearest mean!



$$\hat{Z} = \arg\min_z \|X - \hat{\mu}^{(z)}\|_2$$

  - Bound the mislabeling rate with standard concentration inequalities:

$$\begin{aligned}\mathbb{P}[\hat{Z} = 2 | Z = 1] &\leq \mathbb{P}\left[\left\|X - \hat{\mu}^{(2)}\right\|_2 < \left\|X - \hat{\mu}^{(1)}\right\|_2 \middle| Z = 1\right] \\ &= \mathbb{P}\left[(X - \hat{\mu}^{(1)})^T(\hat{\mu}^{(1)} - \hat{\mu}^{(2)}) < -\tfrac{1}{2}\left\|\hat{\mu}^{(1)} - \hat{\mu}^{(2)}\right\|_2^2 \middle| Z = 1\right] \\ &\leq \frac{\left(\left\|\mu^{(1)} - \mu^{(2)}\right\|_3 + 2\epsilon\right)^2}{\left(\left\|\mu^{(1)} - \mu^{(2)}\right\|_2 - \epsilon\right)^4}\left\|\sigma^{(1)}\right\|_3\end{aligned}$$

  - Decreasing in p for fixed $\hat{\mu}$; rate depends on true separation in means.

Step 3: OLS with Measurement Error

- Suppose $\mathbb{P}[\hat{Z} = z' | Z = z] \leq \zeta$ for any distinct z, z'.
- Consider the OLS estimator $\hat{\beta}_j^{\mathrm{OLS}} = \frac{\hat{\mathbb{E}}\widehat{\mathrm{Cov}}[X_j, Y | \hat{Z}]}{\hat{\mathbb{E}}\widehat{\mathrm{Var}}[X_j | \hat{Z}]}$.

$$\left| \hat{\beta}_j^{\mathrm{OLS}} - \hat{\beta}_j \right| \leq \underbrace{\left| \hat{\beta}_j^{\mathrm{OLS}} - \hat{\beta}_j^{\mathrm{oracle}} \right|}_{\leq O(K\sqrt{\zeta}) \to 0} + \left| \hat{\beta}_j^{\mathrm{oracle}} - \beta_j \right|$$

with probability $1 - \exp\left\{ -\frac{N_2}{8bK} \right\}$ if $K\sqrt{\zeta} \to 0$ and $\omega_z \geq \frac{1}{bK}$.

Step 3: OLS with Measurement Error

- Suppose $\mathbb{P}[\hat{Z} = z' | Z = z] \leq \zeta$ for any distinct z, z'.
- Consider the OLS estimator $\hat{\beta}_j^{\mathrm{OLS}} = \frac{\hat{\mathbb{E}}\widehat{\mathsf{Cov}}[X_j, Y | \hat{Z}]}{\hat{\mathbb{E}}\widehat{\mathsf{Var}}[X_j | \hat{Z}]}$.

$$\left|\hat{\beta}_j^{\mathrm{OLS}} - \hat{\beta}_j\right| \leq \underbrace{\left|\hat{\beta}_j^{\mathrm{OLS}} - \hat{\beta}_j^{\mathrm{oracle}}\right|}_{\leq O(K\sqrt{\zeta}) \to 0} + \left|\hat{\beta}_j^{\mathrm{oracle}} - \beta_j\right|$$

  with probability $1 - \exp\left\{-\frac{N_2}{8bK}\right\}$ if $K\sqrt{\zeta} \to 0$ and $\omega_z \geq \frac{1}{bK}$.
- Since $\hat{\beta}_j^{\mathrm{oracle}} - \beta_j$ is unbiased and asymptotically (with respect to N_2) normal:
  - We have consistency under the above conditions.
  - We have asymptotic normality with oracle variance if further $K\sqrt{N_2\zeta} \to 0$.

Example: Easy Bounds

Suppose for all z there exist a, b such that $\|\mu^{(z)} - \mu^{(z')}\|_2^2 \geq ap$ and $\omega_z \geq \frac{1}{bK}$.

- Given $O(k^3/\delta)$* or $O(p^2/\delta)$ samples, we can learn $\mu^{(z)}$ to $O(\sqrt{p})$ with probability $1-\delta$. (Anandkumar et al., 2014; Guo et al., 2022)
- This gives us mislabeling probabilities ζ of $O(1/p)$.
- Given $O\big(K \log \frac{1}{\delta}\big)$ additional samples, the bias for any $\hat{\beta}_j^{\mathrm{OLS}}$ is $O(K/\sqrt{p})$ with probability $1-2\delta$.

Discussion

- We have a flexible 3 step pipeline.
  - Better tensor decomposition methods for step 1?
  - Sub-Gaussian bounds for step 2?
  - Nonlinear mechanism in step 3?
  - Extend to continuous Z?

Discussion

- We have a flexible 3 step pipeline.
  - Better tensor decomposition methods for step 1?
  - Sub-Gaussian bounds for step 2?
  - Nonlinear mechanism in step 3?
  - Extend to continuous Z?
- We have a trajectory in N, p, and K.
  - Wang and Blei (2019) and Grimmer et al. (2020) require either $p = \infty$ or $N = \infty$.
  - How: Tolerating mislabeled $\hat{Z}$ and carrying error forward.
  - Plausibility: We allow K to increase with p.

Discussion

- We have a flexible 3 step pipeline.
  - Better tensor decomposition methods for step 1?
  - Sub-Gaussian bounds for step 2?
  - Nonlinear mechanism in step 3?
  - Extend to continuous Z?
- We have a trajectory in N, p, and K.
  - Wang and Blei (2019) and Grimmer et al. (2020) require either $p = \infty$ or $N = \infty$.
  - How: Tolerating mislabeled $\hat{Z}$ and carrying error forward.
  - Plausibility: We allow K to increase with p.
- Compare to semiparametric regression $Y = \hat{\beta}_j X_j + g(\hat{f}(X))$.
  - Conditional independence restricts the function class for f in a principled way.
  - This drastically reduces the variance of $\hat{\beta}$.

Bibliography

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral deconfounding via perturbed sparse linear models. *The Journal of Machine Learning Research*, 21(1):9442–9482, 2020.

J. Grimmer, D. Knox, and B. M. Stewart. Na\" ive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv preprint arXiv:2007.12702*, 2020.

B. Guo, J. Nie, and Z. Yang. Learning diagonal gaussian mixture models and incomplete tensor decompositions. *Vietnam Journal of Mathematics*, 50:421–446, 2022.

E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen. Counterexamples to" the blessings of multiple causes" by wang and blei. *arXiv preprint arXiv:2001.06555*, 2020.

Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.