Machine Learning and Causal Inference: Wrapping Up

Ashley I Naimi

June 2022

Contents

1 Alternative Estimands

In this workshop, we focused mostly on the average treatment effect as an estimand, and (to an extent) the effect of treatment on the treated (and untreated).

However, there are a large number of estimands one can consider when seeking to quantify causal effects. We discuss a few here based on several resources (Kennedy, 2022):

1.1 Stochastic Intervention Effects

Stochastic intervention effects provide information on the outcomes that would be observed if we intervention changed the probabilty or the distribution of being exposed in the population. As an example, consider the generic effect:

$$\psi = E(E(Y \mid X^*, C)) = \int \int E(Y \mid X = x, C = c) dG(x \mid c) dP(x)$$

This effect answers the question about what would be observed if we were to set everyone's exposure status to some values drawn from an exposure distribution $G(x \mid c)$. Several results and techniques for implementing stochastic intervention effects have been published, e.g., Munoz and van der Laan (2012), Haneuse and Rotnitzky (2013), Naimi et al. (2014), Young et al. (2014), Kennedy (2019).

1.2 Instrumental Variable Effects

Instrumental variables can be used to target effects without the need for "no unmeasured confounding" assumptions. Under positivity, consistency, no interference, the exclusion restriction assumption, and monotonicity, a valid instrument can yield a local average treatment effect that is equivalent to the complier average causal effect in trial settings (Hernán and Robins, 2006):

$$\psi = E(Y^{x=1} - Y^{x=0} \mid X^{r=1} > X^{r=0}) = \frac{E(E(Y \mid C, R = 1) - E(Y|C, R = 0))}{E(E(X \mid C, R = 1) - E(X|C, R = 0))}$$

Instead of monotonicity, if we assume that the effect of X is constant

across levels of R (effect homogeneity assumption), then it follows that the LATE estimator above quantifies the effect of treatment on the treated (Ogburn et al., 2015). Additionally, slightly stronger effect homogeneity assumption and a different LATE estimator can provide an estimate of the ATE (Wang and Tchetgen Tchetgen, 2018).

2 Treatment Effect Bounds

What happens when the effect we want to estimate is not identifiable? Suppose, for example, exchangeability is violated because we could not randomize our exposure and were aware of the absence of key (unmeasured) confounders? Or perhaps there was some loss to follow-up that could not be accounted for with absolute certainty? More likely there is both unmeasured confounding and loss to follow-up. When this happens, we get a point estimate for the causal effect of interest, but it could either be smaller or larger in magnitude due to the influence of the unmeasured confounder and loss to follow-up.

In order to get a precise measure of **all the values the point estimate can possibly take** as a result of unmeasured confounding and loss to follow-up, we can estimate bounds for the point estimate of interest.[1] Confidence intervals are bounds on the point estimate of interest that capture the uncertainty that results from random variation (Wasserman, 2004). In contrast, identification bounds capture the uncertainty that results from potential violations in some of the identification conditions (Manski, 2003).

[1] Another way of phrasing this is: what range of point estimate values is compatible with the data?

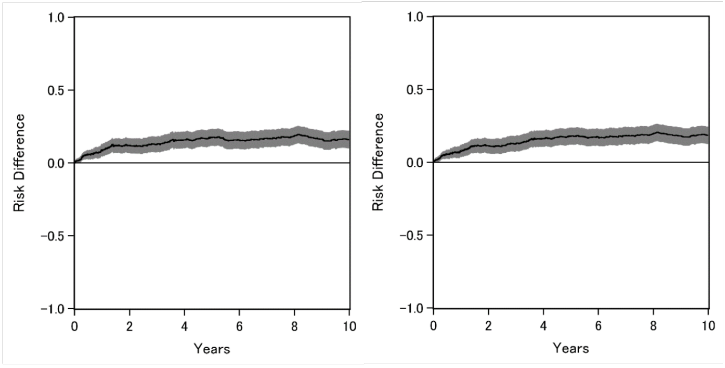
Consider a study by Cole et al. (2019) in which they sought to quantify the effect of injection drug use on time to AIDS or death in a cohort of 1164 adult HIV-positive, AIDS-free women. These women were followed for AIDS or death up to 10 years from 12/6/95 in the Women's Interagency HIV Study (Barkan et al., 1998). Overall, 127 of 1164 women (11%) were lost to follow up. Adjusted risk differences were obtained via inverse probability weighting. Adjustment was made for age, race and nadir CD4 cell count.

Figure 1 shows the results from the analysis (obtained via personal communication with Stephen R. Cole; only a subset of these were presented in the manuscript). The top left panel shows the unadjusted risk difference over follow-up. The top right panel shows the corresponding risk difference after

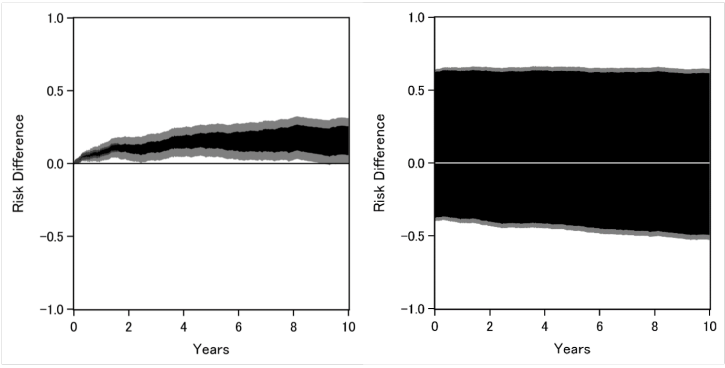
adjusting for loss to follow-up and measured confounders. The bottom left panel shows the identification bounds that result from loss to follow-up. And the bottom right panel shows the identification bounds that result from both loss to follow-up and unmeasured confounding. Specifically, the black area shows all possible risk differences that could arise given the data.

Figure 1: Bounds figure

Figure 4: Difference in risk of AIDS or death by injection drugs use, as a function of time on study, Women's Interagency HIV Study, 1995 to 2006. Courtesy of Stephen R. Cole.



Black line or area is the point or set estimate, and grey area is the 95% confidence interval. Left panel is the crude results, right panel is IP-weighted results.



Black line or area is the point or set estimate, and grey area is the 95% confidence interval. Left panel is bounding for selection bias alone, and right panel is bounding for confounding and selection bias.

The bottom right panel in Figure 1 tells us something critically important that we often fail to consider when conducting an empirical study. Without assumptions, data alone rarely provide much information about a causal effect of interest. Rather, when we interpret that a point estimate from a statistical model as a causal effect estimate, we are invoking a whole set of assumptions (knowingly or unknowingly) that allow us to get a single number out of our data, rather than a range of possible values. One of these sets of assumptions

we discussed here (counterfactual consistency, no interference, positivity, exchangeability, correct model specification). Nonparametric bounds such as those depicted in the study by Cole et al. (2019) help us understand exactly how much support our data provide for an effect of interest, and how much of our results rely on unverifiable assumptions.

3 Time Dependent Exposures and Confounders

Often times, we are interested in the effect of an exposure or intervention measured repeatedly over the course of a (possibly lengthy) follow-up period on an outcome of interest. In this case, to identify the effect of a time-dependent exposure, we need information on all relevant time-dependent confounders. These latter variables can sometimes behave as both confounders and intermediates of the overall effect of interest:

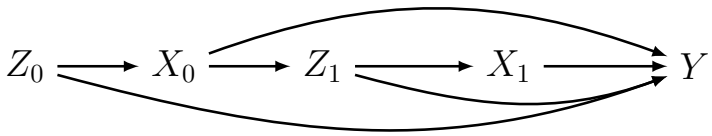


Figure 2: Basic causal structure representing a time dependent exposure (X) and a time dependent confounder (Z) that simultaneously confounds and mediates the relationship between X and Y.

These causal structures are often analyzed using inverse probability weighting (**?**) or g computation (Robins, 1986, **?**) for time-dependent exposures. However, to date, there is only one software program (Lendle et al., 2017) that implements targeted minimum loss-based estimation for time-dependent exposure data.

But there are important challenges in quantifying average treatment effects nonparametrically in longitudinal data, particularly when the follow-up period is long. One important challenge is that the `ltmle` function stratifies the time-dependent estimator by time. Conceptually, one can think of the process as sub-setting the data and fitting a TMLE function within each stratum. However, when the follow-up time is long (e.g., 60 weeks of follow-up), then creating strata by time leads to important data sparsity problems similar to those encountered with a nonparametric maximum likelihood estimator.

Currently, there are several groups doing research on optimal approaches to

modeling time nonparametrically in datasets with lengthy follow-up periods.

4 Mediation Analysis

Mediation questions focus on the extent to which an exposure effect is attributable to some intermediate variable. Mediation analysis has become extremely popular in recent years, and there is now a vast literature on the topic. In particular, careful consideration needs to be given to defining, identifying, and estimating mediation effects. They are seven in total, and represent two foundational articles, two conceptual articles, and three methods articles.

Before 1990, most researchers employed the "Baron and Kenny" approach to mediation and it is (unfortunately) still fairly popular today. Since then, however, extensive research has shown the pitfalls of using the BK approach, which results largely from the fact that the methods used does not allow for consideration of potential confounding of the exposure-outcome, mediator-outcome, and exposure-mediator confounders. Robins and Greenland (Robins and Greenland, 1992) and Pearl (Pearl, 2001) were among the first formal treatments of mediation analysis in the field of causal inference. In many ways, these set the stage for much of the research that followed.

The reviews by Hafeman and Schwartz (Hafeman and Schwartz, 2009) and by VanderWeele and Vansteelandt [2009b] are important reviews/introductions to mediation analysis, albeit at different technical levels. Hafeman and Schwartz (2009) stands out because it highlights the distinction between natural, pure, and total direct and indirect effects. This distinction is often not considered, but can be important when there is exposure-mediator interaction.

Their are also several research papers on how to estimate direct and indirect effects using different methods. Vansteelandt's method (Vansteelandt, 2009) is easy to use, and relies on modeling the outcome to estimate direct effects. VanderWeele's method is also straightforward (VanderWeele, 2009), but relies on modeling the exposure and mediator, instead of the outcome. Finally, the article by Naimi et al reviews both of these, plus two additional approaches (g estimation, TMLE), and illustrates the importance of double-robustness (Naimi et al., 2016). While focused on health disparities, the methods in this latter paper can be applied to any substantive context where mediation is of interest.

5 Machine Learning Reading List

Besides the articles, websites, and books cited throughout this seminar, particularly the ones I pointed out during the course of the lectures, here is a list of reading materials on machine learning that I think are excellent (Asterisks indicated "must reads"!).

5.1 Articles:

Seminal work by Brieman on the distinction between a more classical statistical modeling approach and more recent "algorithmic" modeling approach.

1) *Brieman (2001) Statistical Modeling: The Two Cultures.

Excellent introduction to concepts and issues in using machine learning for epidemiologists.

2) Bi et al (2019) What is Machine Learning?: A Primer for Epidemiologists

Attempt to demonstrate the fundamentals behind the super learner.

3) Naimi & Balzer (2018) Stacked Generalization: An Introduction to Super Learning

Detailed resource on using the super learner (Polley's version) in real data settings.

4) Kennedy (2017) Guide to Super Learner. URL: https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

Important example of some fundamental constraints on using data with algorithms to predict outcomes fairly.

5) Chouldechova (2016) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. https://arxiv.org/abs/1610.07524

Excellent introduction to machine learning (emphasis on econometrics but very useful for epidemiologists).

6) Mullainathan, S. and J. Spiess, Machine learning: an applied econometric approach. Journal of Economic Perspectives, 2017. 31(2): p. 87-106

Important example of how ML algorithms can yield very misleading predictions when deeper aspects of the data-modeling complex are not taken into account.

7) Caruana, R., et al. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. ACM.

8) Kennedy EH. Semiparametric doubly robust targeted double machine learning: a review. arxiv:2203:06469

9) Kennedy EH. Semiparametric theory and empirical processes in causal inference. In Statistical Causal Inferences and Their Applications in Public Health Research, edited by He H, Wu P, Chen D. New York: Springer. 2016; 141-167.

5.2 Books:

Technical Skills

1) *Burkov (2019) The Hundred Page Machine Learning Book

2) Burkov (2021) Machine Learning Engineering

3) Kuhn and Johnson (2016) Applied Predictive Modeling

4) *Raschka et al (2022) Machine Learning with PyTorch and Scikit-Learn

5.3 Conceptual/Theoretical Understanding & Social Issues

1) *Mitchell (2019) Artificial Intelligence: A Guide for Thinking Humans

2) Broussard (2019) Artificial Unintelligence: How Computers Misunderstand the World

5.4 Advanced Texts

1) Wasserman (2006) All of Nonparametric Statistics
2) Shalev-Schwartz and Ben-David (2014) Understanding Machine Learning: From Theory to Algorithms
3) Efron and Hastie (2017) Computer Age Statistical Inference: Algorithms, Evidence, and Data Science
4) Hastie, Tibshirani, Friedman (2009) Elements of Statistical Learning
5) James, Witten, Hastie, Tibshirani (2017) Introduction to Statistical Learning

References

S E Barkan, S L Melnick, S Preston-Martin, K Weber, L A Kalish, P Miotti, M Young, R Greenblatt, H Sacks, and J Feldman. The women's interagency hiv study. wihs collaborative study group. *Epidemiology*, 9(2):117–125, Mar 1998.

Stephen R. Cole, Michael G Hudgens, Jessie K Edwards, M Alan Brookhart, David B Richardson, Daniel Westreich, and Adaora A Adimora. Nonparametric bounds for the risk function. *American Journal of Epidemiology*, 188(4): 632–636, 2019.

Danella M Hafeman and Sharon Schwartz. Opening the black box: a motivation for the assessment of mediation. *Int J Epidemiol*, 38(3):838–845, Jun 2009.

S Haneuse and A Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Stat Med*, 32(30):5260–5277, 2013.

M. A. Hernán and J. M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiol*, 17(4):360–72, 2006. Hernan, Miguel A Robins, James M United States Epidemiology (Cambridge, Mass.) Epidemiology. 2006 Jul;17(4):360-72.

Edward H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv:2203.06469 [stat]*, 2022.

Samuel D. Lendle, Joshua Schwab, Maya L. Petersen, and Mark J. van der Laan. ltmle: An r package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81(1):1 – 21, 2017.

Charles F Manski. *Partial identification of probability distributions*. Springer Science & Business Media, New York, NY, 2003.

Ivan Diaz Munoz and Mark van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

Ashley I. Naimi, Erica EM. Moodie, Nathalie Auger, and Jay S Kaufman. Stochastic mediation contrasts in epidemiologic research: Interpregnancy interval and the educational disparity in preterm birth. *Am J Epidemiol*, 180 (4):436–45, 2014.

Ashley I. Naimi, Mireille E. Schnitzer, Erica E. M. Moodie, and Lisa M. Bodnar. Mediation analysis for health disparities research. *American Journal of Epidemiology*, 184(4):315–324, 2016. DOI: 10.1093/aje/kwv329. URL http://aje.oxfordjournals.org/content/184/4/315.abstract.

Elizabeth L. Ogburn, Andrea Rotnitzky, and James M. Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

J. Pearl. Direct and Indirect Effects. In John Breese and Daphne Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–20. Morgan Kaufmann, San Francisco, CA, 2001.

J. M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiol*, 3(2):143–155, 1992.

T. J. VanderWeele. Marginal structural models for direct and indirect effects (erratum in *Epidemiology* 2009; 20(4):629). *Epidemiol*, 20(1):18–26, 2009.

Stijn Vansteelandt. Estimating direct effects in cohort and case–control studies [erratum in: *Epidemiol* 2010:21(2)]. *Epidemiol*, 20(6):851–860, 2009.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80 (3):531–550, 2018.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, New York, 2004.

Jessica G Young, Miguel Á Hernán, and James M Robins. Identification, estimation and approximation of risk under interventions that depend on the

natural value of treatment using observational data. *Epidemiologic Methods*, 3(1):1–19, 2014.