TARGETED LEARNING FOR CAUSAL EFFECT ESTIMATION

Project instructors:

Helene Rytgaard (hely@sund.ku.dk) [1] and Anders Munch (a.munch@sund.ku.dk)[1]

[1]*Section of Biostatistics, University of Copenhagen*

In this project we consider causal effect estimation in semiparametric settings where the target parameter is low-dimensional but estimation has to deal with potentially high-dimensional nuisance parameters often taking the form of a regression function. In these settings, efficient influence function based estimation provides a popular basis for combining machine learning techniques with valid statistical inference. We will consider and investigate these methods for estimation of the average treatment effect (ATE) based on observational (i.e., non-randomized) data.

1 The problem to be studied

In randomized trials, where trial participants are enrolled and randomized to a treatment or a placebo arm, the randomization ensures that the differences in outcome occur as a result of treatment differences only. Randomized trials are often considered the golden standard for causal inference, but may not always be feasible or ethical. In this project, the question we want to address is whether having a planned cesarian section (intended cesarian section) among women who gave birth twice changes the risk of postpartum haemorrhage during the second delivery. For this question, performing a randomized trial would not be ethical. With tools from causal inference and semiparametric efficiency theory, we can formulate and optimally estimate causal effects from observational data.

As part of the project, you will:

- Understand key concepts from semiparametric efficiency theory and the bias-correction abilities of influence function based estimation.
- Translate a real-world data application into a mathematical and statistical formulation of the estimation problem that needs to be solved.
- Assess model misspecification and estimation performance via simulations in `R`.
- Conduct an analysis of the Danish Birth Registry data in `R` for average treatment effect estimation.

2 Targeted learning

The targeted learning framework has been developed as a general template for combining machine learning and causal inference. In this project we consider a particular data structure as follows. Suppose we observed an iid sample $O_1, \ldots, O_n$, $n \in \mathbb{N}$ of a random variable $O \in \mathcal{O}$ distributed according to an unknown distribution function P_0 belonging to a statistical model $\mathcal{M}$. Each observation consists of $O = (X, A, Y)$ where $X \in \mathbb{R}^d$ are covariates, $A \in \{0, 1\}$ is a binary exposure and $Y \in \{0, 1\}$ is a binary outcome variable.

The target parameter can be written as functional $\Psi : \mathcal{M} \to \mathbb{R}$ of distributions $P \in \mathcal{M}$, for our purposes defined as

$$\Psi(P) = \mathbb{E}_P\big[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}[Y \mid A = 0, X]\big], \tag{1}$$

where $\mathbb{E}_P[\cdot]$ denotes the expectation operator under the distribution $P \in \mathcal{M}$. The parameter defined by (1) is commonly referred to as the *average treatment effect* (ATE); under structural assumptions, the ATE can be interpreted causally (Hernan and Robins, 2020). A straightforward estimator of the ATE would be

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n(1, X_i) - \hat{f}_n(0, X_i)\right\},$$

where $\hat{f}_n$ denotes some estimator of the conditional expectation $\mathbb{E}[Y \mid A, X]$. Note that the task of estimating $\mathbb{E}[Y \mid A, X]$ is a standard classification problem, and hence we have a large collection of well-studied estimators of at our disposal, e.g., logistic regression, random forests or neural networks. Perhaps surprisingly, nice performance for estimation of $\mathbb{E}[Y \mid A, X]$ does not necessarily translate into nice performance for estimation of the target parameter $\Psi(P)$. Basing the estimation procedure on the so-called efficient influence function, on the other hand, shifts the performance optimization to the target parameter specifically. This requires estimating both the exposure distribution $P(A = a \mid X)$ as well as the outcome regression above (potentially using machine learning), but then provides the basis for asymptotic linearity and efficiency of the resulting estimator and thus inference based on the limiting normal distribution. Thus, estimation based on influence functions allow us to use flexible estimators for P_0, to ensure optimal asymptotic behavior of the estimator without strict parametric assumptions.

In this project we focus on the specific problem of estimating the ATE, but the field of targeted learning can more broadly be seen as one approach to combine the advantages of machine learning (flexible and data-adaptive models with few assumptions) with the goal of more traditional inference based statistics, which seeks to establish valid confidence intervals for interpretable parameters. This is an important general problem to address if we want to utilize the advances of machine learning to answer scientifically interesting questions.

3 Data

In the project we will work with a random subset of the data described in Wikkelsø et al. (2014). The dataset contains data from 48272 Danish women who gave birth twice. The main outcome is a binary variable called `PPHbin` which indicates if the woman had a postpartum haemorrhage (to haemorrhage means to bleed very heavily) during the second delivery. The data are from the registries and have an observational character. In particular, note that the decision to plan a cesarian section was not randomized. With causal tools we want to analyze the causal effect of a planned cesarian section on the risk of postpartum haemorrhage, i.e., the effect that one would have observed in a hypothetical study which randomizes women to either intended cesarian section or intended vaginal birth.

4 What you need to know before the course

It is not required that you know semiparametric efficiency theory nor causal inference beforehand, but you must be able to work comfortably with:

* Derivatives of functions, directional derivative, chain rule.
* Probability distributions, conditional distributions, empirical distribution
* Central limit theorem

- Regression models, including logistic regression
- Statistical inference, parameter estimation, confidence intervals, p-values
- `R` programming

5 Project work

Figure 1 illustrates the roadmap of the data analysis.

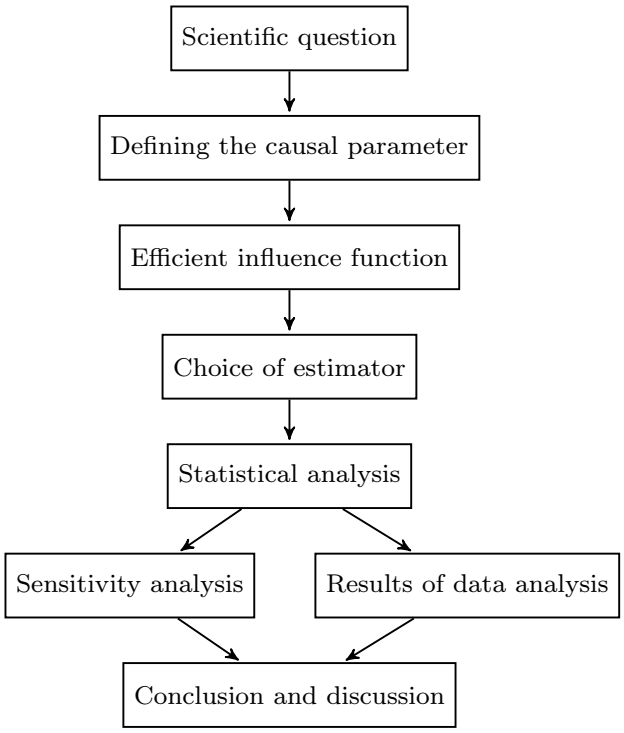

Figure 1: Roadmap for the data analysis of the project.

6 Some literature

There are many papers and some books about the mathematical background of the methods we consider. A very nice review is provided by Kennedy (2016). For causal notation, we follow the framework with counterfactual variables Hernan and Robins (2020). We further recommend the introduction (pp. 1–11) by Chernozhukov et al. (2018).

Other more in-depth references on semiparametric efficiency theory for the more interested are Bickel et al. (1993); van der Vaart (2000); van der Laan and Robins (2003); Tsiatis (2007); van der Laan and Rose (2011, Appendix A).

References

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. Efficient and adaptive inference in semiparametric models, 1993.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

M. A. Hernan and J. M. Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, Fl, 2020.

E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A. J. Wikkelsø, S. Hjortøe, T. A. Gerds, A. M. Møller, and J. Langhoff-Roos. Prediction of postpartum blood transfusion–risk factors and recurrence. *The Journal of Maternal-Fetal & Neonatal Medicine*, 27 (16):1661–1667, 2014.