Project in biostatistics: Targeted learning and cross-fitting

Anders Munch & Thomas Alexander Gerds

Department of Biostatistics, University of Copenhagen

April 13, 2024

Background and motivation

The central philosophy of targeted learning is to clearly separate the definition of the parameter of interest, *the target parameter*, and the models/algorithms used for estimation. Let $\mathcal{P}$ be a collection of probability measures and $\Psi \colon \mathcal{P} \to R$ a pathwise differentiable parameter of interest, called *the target parameter*. In this project we consider estimation problems, where the parameter Ψ can be written as

$$\Psi(P) = \int \phi(x, \nu(P)) P(\mathrm{d}x) = P[\phi(\cdot, \nu(P))]. \tag{1}$$

In this setting $\nu : \mathcal{P} \to \mathcal{F}$ is a function-valued nuisance parameter, such that $v(P)$ is a regression function. Note also that we use the notation $P[f] = \int f \mathrm{d}P$. By using tools from semiparametric efficiency theory it is possible to choose the function ϕ such that valid statistical inference for Ψ can be obtained, even when ν is estimated with data-adaptive algorithms and machine learning (van der Laan and Rose, 2011; Chernozhukov et al., 2018).

Let P_n denote the empirical measure of an iid. sample $\{O_i\}_{i=1}^n$ from some $P \in \mathcal{P}$ and suppose we have been given an estimator $\hat{\nu}_n$ of the nuisance parameter ν. Then, a natural estimator of the target parameter Ψ is

$$\hat{\Psi}_n = P_n[\phi(\cdot, \hat{\nu}_n)].$$

This estimator will be consistent and asymptotically normal under a set of regularity assumptions about the function ϕ and the set $\mathcal{P}$ if the estimator $\hat{\nu}_n$ converges sufficiently fast to $\nu(P)$ for all $P \in \mathcal{P}$. One assumption is that $\hat{\nu}_n$ takes values in a so-called *Donsker class* of functions. To avoid this assumption it has been suggested to instead employ *cross-fitting*, where P_n and $\hat{\nu}_n$ are constructed using separate parts of the data (Chernozhukov et al., 2018).

The project

In this project we aim to learn about the theoretical and practical properties of cross-fitting in the context of targeted learning. You will work with the data described in (Wikkelsø et al., 2014). The dataset available for this project contains data from 48,272 Danish women who all gave birth twice. The main outcome is a binary variable indicating if the woman had a postpartum haemorrhage (heavy bleeding) during the second delivery. The target parameter is the causal effect of a planned cesarian section on the risk of postpartum haemorrhage at the second delivery. This parameter is a special case of equation 1 where $\nu(P)$ is the conditional probability of a planned cesarian section at the second delivery given characteristics of the first delivery. Under the usual assumptions needed for causal inference, the causal effect coincides with the one of a hypothetical study which randomizes women to either planned cesarian section or intended vaginal birth.

You will derive a cross-fitting algorithm and construct a targeted minimum loss based estimator (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). You will study the role of cross-fitting for the asymptotic inference theoretically and the effect of cross-fitting in finite samples. In this project we work with elements of causal inference and asymptotic semiparametric theory. Good introductions to the theory of targeted learning are Kennedy (2016, 2022) and Hines et al. (2022). Recent papers on cross-fitting are Chen et al. (2022) and Zivich and Breskin (2021).

References

Chen, Q., V. Syrgkanis, and M. Austern (2022). Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*.

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

Hernán, M. and J. Robins (2020). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Hines, O., O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*.

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

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

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

van der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *The international journal of biostatistics 2*(1).

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

Zivich, P. N. and A. Breskin (2021). Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*.