Project in Biostatistics: Hyperparameter optimization

Anders Munch & Thomas Gerds

April 25, 2022

Outline of the setting

Let $\mathcal{P}$ be a collection of probability measures over $\mathbb{R}^{d+2}$, so that $O \sim P \in \mathcal{P}$, with $O = (Y, A, X)$, $Y \in \mathbb{R}$, $A \in \mathbb{R}$, and $X \in \mathbb{R}^d$.

We consider estimation of a parameter $\theta \colon \mathcal{P} \to \Theta$, where Θ is either a **high-** or a **low-dimensional** space:

1. Risk prediction model: For some suitable function space $\mathcal{F}$, the parameter of interest is $\nu \colon \mathcal{P} \to \mathcal{F}$,

$$\nu(P) = f, \text{ for } f(x, a) = \mathbb{E}_P[Y \mid X = x, A = a].$$

2. Low-dimensional (causal) target parameter, for instance the average treatment effect: The parameter of interest is $\Psi \colon \mathcal{P} \to \mathbb{R}$,

$$\begin{aligned}\Psi(P) &= \mathbb{E}_P\left[\mathbb{E}_P[Y \mid X, A = 1] - \mathbb{E}_P[Y \mid X, A = 0]\right] \\ &= \int \{\nu(x, 1) - \nu(x, 0)\} \, \mathrm{d}P_X(x), \quad \nu = \nu(P).\end{aligned}$$

Data

You will work with a random subset of the data described in [6]. The dataset contains data from 48,272 Danish women who gave birth twice. The main outcome is a binary variable called PPH which indicates if the woman had a postpartum haemorrhage (heavy bleeding) during the second delivery.

High- and low-dimensional inference problems

1. Predicting the risk of PPH at second delivery.
2. Estimating the causal effect of a planned cesarian section on PPH.

Hyperparameter and model selection/construction

Estimation of ν is a well-studied problem. Many (machine) learning approaches depend on one or more hyperparameters that has to be chosen in some way. We want to examine how to do this properly, and what the practical effects are of the different possible approaches.

Example

One of the simplest examples of hyperparameter selection is ridge regression, which estimates ν with

$$\hat{\nu}_\lambda(x) = x^\top \hat{\beta}, \quad \hat{\beta} := \operatorname*{argmin}_\beta \frac{1}{n} \sum_{i=1}^{n} L(y_i, x_i^\top \hat{\beta}) + \lambda \|\beta\|^2.$$

We can also consider a collection of models with different strengths to construct more robust estimators (e.g., the Super Learner [2, 4]).

Target parameter and causal inference

Target parameter

By explicitly defining a parameter of interest as an answer to a concrete scientific question we ensure that we can draw relevant and meaningful conclusions from our statistical analysis.

High-dimensional nuisance parameter

By allowing high-dimensional nuisance parameters we avoid having to make unrealistic model assumptions.

Causal inference

One example is the average treatment effect,

$$\Psi(P) = \int \{\nu(x,1) - \nu(x,0)\} \, \mathrm{d}P_X(x), \quad \nu = \nu(P) \in \mathcal{F}.$$

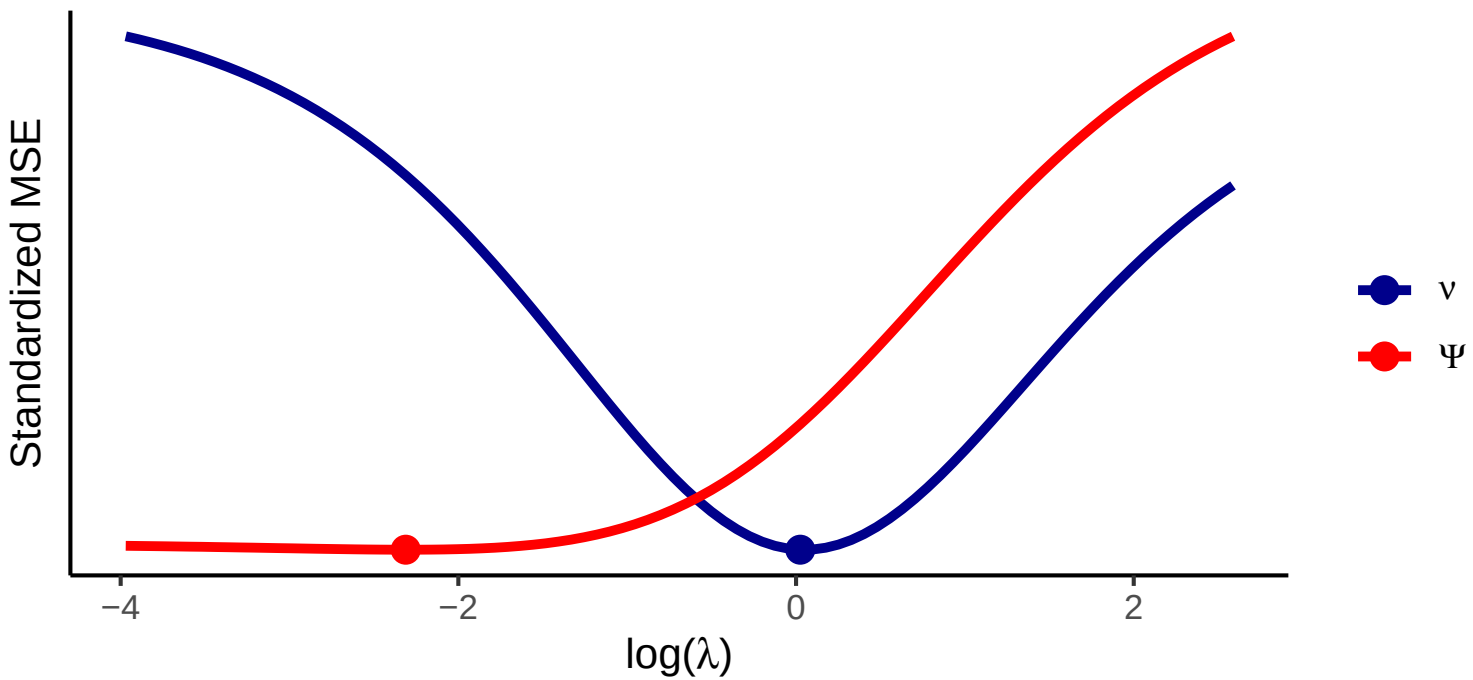
- How do we formalize a causal question?
- Which assumptions are needed for the above parameter to have a causal interpretation?
- How likely are these assumptions to hold for the data at hand?

Targeted inference and hyperparameter selection

Finding a good estimator $\hat{\nu}$ of ν does *not* necessarily provide a good plug-in estimator of Ψ! $\rightarrow$ targeted learning and debiased ML [5, 4, 1, 3].

Let $\hat{\nu}_\lambda$ be the ridge regression estimator of ν with penalty parameter λ.
Let $\hat{\Psi}_\lambda$ denote the plug-in estimator of Ψ based on $\hat{\nu}_\lambda$, i.e.,

$$\hat{\Psi}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\nu}_\lambda(X_i, 1) - \hat{\nu}_\lambda(X_i, 0)\}.$$



References

[1] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[2] Katherine Hoffman. Khstats. `https://www.khstats.com/blog/sl/superlearning/`. Accessed: 2022-04-19.

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

[4] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[5] Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[6] Anne J Wikkelsø, Sofie Hjortøe, Thomas A Gerds, Ann M Møller, and Jens Langhoff-Roos. Prediction of postpartum blood transfusion–risk factors and recurrence. *The Journal of Maternal-Fetal & Neonatal Medicine*, 27(16):1661–1667, 2014.