Chapter 7

Bounded Continuous Outcomes

Susan Gruber, Mark J. van der Laan

This chapter presents a TMLE of the additive treatment effect on a bounded continuous outcome. A TMLE is based on a choice of loss function and a corresponding parametric submodel through an initial estimator, chosen so that the loss-function-specific score of this parametric submodel at zero fluctuation equals or spans the efficient influence curve of the target parameter. Two such TMLEs are considered: one based on the squared error loss function with a linear regression model, and one based on a quasi-log-likelihood loss function with a logistic regression submodel. The problem with the first TMLE is highlighted: the linear regression model is not a submodel and thus does not respect global constraints implied by the statistical model. It is theoretically and practically demonstrated that the TMLE with the logistic regression submodel is more robust than a TMLE based on least squares linear regression. Some parts of this chapter assume familiarity with the core concepts, as presented in Chap. 5. The less theoretically trained reader should aim to navigate through these parts and focus on the practical implementation and importance of the presented TMLE procedure. This chapter is adapted from Gruber and van der Laan (2010b).

7.1 Introduction

TMLE of a target parameter of the data-generating distribution, known to be an element of a semiparametric model, involves selecting a loss function (e.g., log-likelihood) and constructing a parametric submodel through an initial density estimator with parameter $\epsilon$, so that the loss-function-specific “score” at $\epsilon = 0$ equals or spans the efficient influence curve (canonical gradient) at the initial estimator. This $\epsilon$ represents an amount of fluctuation of the initial density estimator. The latter “score” constraint can be satisfied by many loss functions and parametric submodels, since it represents only a local constraint of the submodels’ behavior at zero fluctuation.
However, it is very important that the fluctuations encoded by the parametric model stay within the semiparametric model for the observed data distribution (otherwise it is not a submodel!), even if the target parameter can be defined on fluctuations that fall outside the assumed observed data model.

In particular, in the context of sparse data, by which we mean situations where the generalized Cramér–Rao lower bound is high, a violation of this property can significantly affect the performance of the estimator. We demonstrate this in the context of estimation of a causal effect of a binary treatment on a continuous outcome that is bounded. It results in a TMLE that inherently respects known bounds and consequently is more robust in sparse data situations than a TMLE using a naive parametric fluctuation working model that is actually not a submodel of the assumed statistical model.

Sparsity is defined as low information in a data set for the purpose of learning the target parameter. Formally, the Fisher information $I$ is defined as sample size $n$ divided by the variance of the efficient influence curve: $I = n / \text{var}(D'(O))$, where $D'(O)$ is the efficient influence curve of the target parameter at the true data-generating distribution. The reciprocal of the variance of the efficient influence curve can be viewed as the information one observation contains for the purpose of learning the target parameter. Since the variance of the efficient influence curve divided by $n$ is the asymptotic variance of an asymptotically efficient estimator, one can also think of the information $I$ as the reciprocal of the variance of an efficient estimator of the target parameter. Thus, sparsity with respect to a particular target parameter corresponds with small sample size relative to the variance of the efficient influence curve for that target parameter.

The following section begins with background on the application of TMLE methodology in the context of sparsity and its power relative to other semiparametric efficient estimators since it is a substitution estimator respecting global constraints of the semiparametric model. Even though an estimator can be asymptotically efficient without utilizing global constraints, the global constraints are instrumental in the context of sparsity with respect to the target parameter, motivating the need for semiparametric efficient substitution estimators, and for a careful choice of fluctuation function for the targeting step that fully respects these global constraints. A rigorous demonstration of the proposed TMLE of the causal effect of a binary treatment on a bounded continuous outcome follows, and the TMLE using a linear fluctuation function (i.e., that does not represent a parametric submodel) is compared with the proposed TMLE using a logistic fluctuation function. In Sect. 7.3, we carry out simulation studies that compare the two TMLEs of the causal effect, with and without sparsity in the data. Results for other commonly applied estimators discussed in Chap. 6 (MLE according to a parametric statistical model, IPTW, and A-IPTW) are also presented.