Estimation bayésienne ciblée de modèles structurels marginaux

Herbert Susmann [1] & Antoine Chambaz [2]

[1] *University of Massachusetts Amherst and MAP5 (UMR CNRS 8145), Université de Paris, hsusmann@umass.edu*
[2] *MAP5 (UMR CNRS 8145), Université de Paris, antoine.chambaz@u-paris.fr*

Résumé. L'une des tâches principales de l'inférence causale est d'estimer l'effet d'un traitement sur une issue d'intérêt. On peut aussi chercher la relation entre la magnitude de l'effet et les covariables individuelles. Les modèles structurels marginaux permettent de définir une telle relation via l'introduction d'un modèle de travail. Les paramètres du modèle de travail peuvent être estimés de manière efficace par maximum de vraisemblance ciblée. Dans cette étude, nous présentons un nouvel estimateur bayésien des paramètres d'un modèle structurel marginal qui combine les avantages de l'inférence bayésienne avec les propriétés souhaitables de l'approche fréquentiste.
Mots-clés. analyse causale, aprentissage ciblé, inférence bayésienne, statistique semi-paramétrique

Abstract. One of the principal tasks in causal inference is to estimate the effect of a treatment on an outcome. It may also be of interest to learn about the relationship between treatment effect size and individual covariates. Marginal structural models provide a way to summarize the relationship between treatment effects and covariates via a working model. The parameters of the working model can be estimated efficiently via Targeted Maximum Likelihood Estimation (TMLE). In this work, we present a novel Bayesian estimator of the parameters of a marginal structural model that combines the benefits of Bayesian inference with the desirable properties of the frequentist TMLE.
Keywords. bayesian inference, causal analysis, semi-parametric statistics, targeted learning

1 Background

1.1 Introduction

A common goal in causal inference is to estimate the effect of a treatment on an outcome. In the case of a binary treatment, the popular Average Treatment Effect (ATE) parameter summarizes the average treatment effect over the population. In many cases it is also of interest to know how the treatment effect differs in various subgroups of the population. In particular, there may be a set of potential *treatment effect modifiers*, variables which are thought to change the size of the treatment effect. Ideally we would be able to simply estimate treatment effects within each stratum of effect modifier we are interested in. In practice this is often difficult, especially if there are strata with few observations. Marginal structural models (MSMs) provide one possible path forward. In this approach, we introduce a working model that summarizes the relationship between the potential effect modifiers and the treatment effects. We then seek to estimate a set of parameters for the working model that minimizes the loss between the treatment effect for each stratum of covariates and the modeled treatment effects given by the working model.

We seek to estimate the MSM parameters within a non-parametric model. The Targeted Learning framework, specifically Targeted Maximum Likelihood Estimation (TMLE), provides a general blueprint for constructing efficient estimators of low-dimensional parameters in non-parametric models (Van der Laan & Rose, 2011). Relevant to our work, a frequentist TMLE for estimating the parameters of a marginal structural model was presented by Rosenblum and van Der Laan (2010). Liu et al. (2021) apply a similar approach to estimate MSM parameters for potential treatment effect modifiers.

While most of the development of targeted learning methodology has taken place in a frequentist framework, there has been work on extending TMLE to a Bayesian context. Diaz, Hubbard, and van der Laan (2011) describes a Bayesian TMLE for estimating the average treatment effect, and Díaz, Savenkov, and Kamel (2020) a targeted Bayesian approach for estimating class proportions in an unlabeled dataset. The general approach we take in this paper follows that of both of these papers, which combines the benefits of Bayesian inference (subjective interpretation of probability, incorporation of priors) with the benefits of the frequentist approach (efficient, non-parametric inference.)

1.2 Inferential Problem

First, we introduce the structure of the observed data. Let W be a vector of covariates, $Y \in \{0,1\}$ a binary indicator of the outcome, and $A \in \{0,1\}$ a binary indicator of treatment. Let $V = (V_1, \ldots, V_k)$ be a subset of the covariates chosen by the analyst to be possible treatment effect modifiers. Let $X = (1, V_1, \ldots, V_k)$, with size $p = k + 1$. Let $O = (W, A, Y)$, which we assume is drawn from a distribution P_0, where P_0 is taken to be within a non-parametric model $\mathcal{M}$. The observed data set $\{O_1, O_2, \ldots, O_n\}$ is assumed to be composed of n i.i.d. draws from P_0.

Before we define the parameter of interest, we introduce notation that will simplify the exposition. Let the conditional mean and so called "blip" function be

$$\bar{Q}_P(a, w) = \mathbb{E}_P[Y \mid A = a, W = w] \quad \text{and} \quad \tilde{Q}_P(w) = \bar{Q}_P(1, w) - \bar{Q}_P(0, w).$$

Write $\bar{Q}_0 \equiv \bar{Q}_{P_0}$ and $\tilde{Q}_0 \equiv \tilde{Q}_{P_0}$. We summarize the relationship between treatment effect and treatment effect modifiers via a working model. Specifically, for all $P \in \mathcal{M}$, define the parameter $\boldsymbol{\beta}(P) \in \mathbb{R}^p$ as the solution to the optimization problem:

$$\boldsymbol{\beta}(P) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}_P \left[\left(\tilde{Q}_0(W) - X'\boldsymbol{\beta} \right)^2 \right]. \tag{1}$$

For convenience we denote $\boldsymbol{\beta}_0 \equiv \boldsymbol{\beta}(P_0)$. While the focus of this paper is on the statistical estimation of $\boldsymbol{\beta}_0$, we note that the parameter is causally identifiable under standard assumptions (i.e. exchangeability and positivity).

2 Targeted Maximum Likelihood Estimation

In this section we present two estimators for $\boldsymbol{\beta}_0$. The first is a frequentist TMLE (Liu et al., 2021; Rosenblum & van Der Laan, 2010). The second, which is the main contribution of our work, is a Bayesian TMLE. In both approaches, our goal is to efficiently estimate $\boldsymbol{\beta}_0$ within the non-parametric model $\mathcal{M}$. Semi-parametric efficiency theory teaches us that the efficiency bound is given by $\mathrm{var}_{P_0}(D_0(O))$, where D_0 is called the *efficient influence function* (EIF) of the target parameter β at P_0 and is given by

$$D_0(O) = M^{-1} \left[\frac{2A - 1}{g_0(A, W)} \left(Y - \bar{Q}_0(A, W) \right) + \tilde{Q}_0(W) - \boldsymbol{\beta}^\top X \right] X$$

where $g_0(a, w) \equiv E_{P_0}[A = a \mid W = w]$ and the normalizing matrix M is given by $M = \mathbb{E}_{P_0}[X^\top X]$ (Liu et al., 2021; Rosenblum & van Der Laan, 2010).

2.1 Frequentist Inference

In this section we sketch a brief outline of the frequentist TMLE. First, we need a set of initial estimators of all the nuisance parameters involved in the target parameter and the efficient influence function. For the marginal distribution $Q_{W,n} \equiv P(W = w)$ we use the empirical distribution of the covariates, and for $\bar{Q}_n$ and g_n estimates could be obtained through parametric estimators like logistic regression, or through flexible machine learning methods (or an ensemble of methods). Next, we set a parametric sub-model indexed by a finite dimensional parameter $\boldsymbol{\epsilon}$ which fluctuates the initial estimates, designed such that it equals the initial estimators $\bar{Q}_n$, g_n, and $Q_{W,n}$ at $\boldsymbol{\epsilon} = \mathbf{0}$ and such that its score at $\boldsymbol{\epsilon} = \mathbf{0}$ spans the EIF of the target parameter. We then estimate the parameter $\boldsymbol{\epsilon}$ using maximum-likelihood estimation. The initial estimates are then updated according to the parametric sub-model to yield a new set of estimates. This process is repeated until convergence (until the MLE of $\boldsymbol{\epsilon}$ becomes sufficiently close to zero). The final set of estimates are denoted $\bar{Q}_n^*$, g_n^*, and $Q_{W,n}^*$. We then estimate the value of the target parameter $\boldsymbol{\beta}_n^*$ by solving the optimization problem posed in (1) using $\bar{Q}_n^*$ and $Q_{W,n}^*$ as estimates of the nuisance parameters (as the optimization problem depends only on them). Then $\boldsymbol{\beta}_n^*$ is asymptotically normal and efficient (Van der Laan & Rose, 2011).

2.2 Bayesian Inference

The Bayesian version of the frequentist TMLE is derived from the fact that the parametric sub-model used in the frequentist TMLE defines a likelihood. As such, Bayesian inference can be performed as usual to find a posterior distribution for the parameter $\boldsymbol{\epsilon}$. Bayesian inference requires specifying a prior distribution on the parameter $\boldsymbol{\epsilon}$. Rather than putting a prior on $\boldsymbol{\epsilon}$ directly, we prefer to map a prior on the parameter $\boldsymbol{\beta}$ back to $\boldsymbol{\epsilon}$. Let $\boldsymbol{\beta}(\boldsymbol{\epsilon}) : \mathbb{R}^p \mapsto \mathbb{R}^p$ be a function that maps $\boldsymbol{\epsilon}$ to a parameter value $\boldsymbol{\beta}$. We assume that this function is invertible. Then a prior distribution $\pi_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is mapped to a prior distribution $\pi_{\boldsymbol{\epsilon}}$ on $\boldsymbol{\epsilon}$ by the formula

$$\pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \pi_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}(\boldsymbol{\epsilon})\right) \left|\det\left(J(\boldsymbol{\epsilon})\right)\right|$$

where J is the Jacobian of the transformation $\boldsymbol{\beta}(\boldsymbol{\epsilon})$. The posterior distribution of $\boldsymbol{\epsilon}$ is then given by

$$\pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|O_1, \ldots, O_n) \propto \pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \prod_{i=1}^{n} p_{\epsilon}(O_i \mid \boldsymbol{\epsilon}).$$

Sampling techniques such as Markov-Chain Monte Carlo (MCMC) can be used to draw a set of samples $\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(\ell)}$ from the posterior distribution of $\boldsymbol{\epsilon}$, which can then be used to generate a set of samples of $\boldsymbol{\beta}$ through the mapping $\boldsymbol{\beta}(\boldsymbol{\epsilon})$. Our main finding is a Bernstein-von Mises type result that the posterior distribution of $\boldsymbol{\beta}$ converges to a multivariate normal distribution centered on the truth and with variance given by $\mathrm{var}_{P_0}(D_0(O))$. We will also present results from a simulation study illustrating the performance of the Bayesian TMLE in finite sample settings.

References

Diaz, I., Hubbard, A. E., & van der Laan, M. J. (2011). Targeted Bayesian learning. In *Targeted learning* (pp. 475–493). Springer.

Díaz, I., Savenkov, O., & Kamel, H. (2020). Nonparametric targeted Bayesian estimation of class proportions in unlabeled data. *Biostatistics*.

Liu, Y., Schnitzer, M., Wang, G., Kennedy, E., Viiklepp, P., Vargas, M. H., ... Benedetti, A. (2021). Modeling treatment effect modification in multidrug-resistant tuberculosis in an individual patient data meta-analysis. *arXiv preprint arXiv:2101.03997*.

Rosenblum, M., & van Der Laan, M. J. (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The international journal of biostatistics*, *6*(2).

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