Inference in Marginal Structural Models by Automatic Targeted Bayesian and Minimum Loss-Based Estimation

Herbert Susmann, Antoine Chambaz

Inference in Marginal Structural Models by Automatic Targeted Bayesian and Minimum Loss-Based Estimation

Herbert Susmann[1], Antoine Chambaz[2]

[1] Department of Biostatistics & Epidemiology, University of Massachusetts Amherst
[2] Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

January 23, 2023

Abstract

Two of the principle tasks of causal inference are to define and estimate the effect of a treatment on an outcome of interest. Formally, such treatment effects are defined as a possibly functional summary of the data generating distribution, and are referred to as *target parameters*. Estimation of the target parameter can be difficult, especially when it is high-dimensional. Marginal Structural Models (MSMs) provide a way to summarize such target parameters in terms of a lower dimensional working model. We introduce the semi-parametric efficiency bound for estimating MSM parameters in a general setting. We then present a frequentist estimator that achieves this bound based on Targeted Minimum Loss-Based Estimation. Our results are derived in a general context, and can be easily adapted to specific data structures and target parameters. We then describe a novel targeted Bayesian estimator and provide a Bernstein von-Mises type result analyzing its asymptotic behavior. We propose a universal algorithm that uses automatic differentiation to put the estimator into practice for arbitrary choice of working model. The frequentist and Bayesian estimators have been implemented in the Julia software package `TargetedMSM.jl`. Finally, we illustrate our proposed methods by investigating the effect of interventions on family planning behavior using data from a randomized field experiment conducted in Malawi.

1 Introduction

Causal inference is concerned with defining and estimating the effect of a treatment on an outcome of interest. For example, consider the *Conditional Average Treatment Effect* parameter. Let $O_{1:n}$ be n i.i.d. draws from a distribution P_0 of a generic variable $O = (X, A, Y)$ where X is a vector of covariates, A is a binary treatment indicator, and Y a binary outcome. The CATE is defined as the following functional summary of P,

$$\Psi_P^\sharp(X) := \mathbb{E}_P\left[Y|A=1,X\right] - \mathbb{E}_P\left[Y|A=0,X\right],$$

and is interpretable as the expected difference in outcome given treatment vs. non-treatment within strata of covariates X.

Estimating $\Psi_P^\sharp(X)$ may be challenging, especially when X is high-dimensional. One way to proceed would consist in assuming (recklessly) that $\Psi_P^\sharp(X)$ belongs to a known low-dimensional parametric model. Suppose there are $V \subset X$ potential *treatment effect modifiers*. A parametric model for the marginal distribution of $\Psi_P^\sharp(X)$ that assumes a linear functional form conditional on V could be

$$\mathbb{E}_P[\Psi_P^\sharp(X) \mid V] = \mathrm{MSM}_{\boldsymbol{\beta}}(V) := \boldsymbol{\beta}^\top (1, V)^\top. \tag{1}$$

The target of estimation would then be the parameter $\boldsymbol{\beta} \in \mathcal{B}$, which is of lower dimension than $\Psi_P^\sharp(X)$. This approach was originally introduced by Robins (1998); in this line of research, the assumed parametric model on the marginal distribution of the target parameter is referred to as a *Marginal Structural Model* (MSM).

In practice, the correct functional form of the parameter of interest will not be known. Any parametric model that is adopted will usually be misspecified. Non-parametric Marginal Structural Models build on

MSMs without assuming correct specification (Neugebauer and van der Laan, 2007; van der Laan and Rose, 2011). The idea is to summarize the parameter of interest in terms of a lower-dimensional working model with respect to a loss function. For example, in the spirit of (1), we could define a new parameter $B(P)$ as

$$B(P) := \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}_P \left[\left(\Psi^\sharp_P(X) - \text{MSM}_{\boldsymbol{\beta}}(V) \right)^2 \right].$$

That is, $\text{MSM}_{B(P)}(V)$ is the best linear approximation of $\Psi^\sharp_P(X)$ in terms of V with respect to squared-error loss function. Note that, in contrast to the first approach, the definition of $B(P)$ introduces no statistical assumptions on the form of $\Psi^\sharp_P$. We adopt this non-parametric approach to MSMs in the current work, providing a general definition and notation in Section 2.

Continuing in the same spirit, we seek to estimate $B(P_0)$ as non-parametrically as possible, without making any parametric assumptions on the distribution P_0. We only assume that P_0 is within a non-parametric model $\mathcal{M}$ of laws defined on the support of the observed data. The goal is to construct an asymptotically normal and efficient estimator of $B(P_0)$. The first step is to derive the *efficient influence function* (EIF) $D^*(P)$ of the target parameter at any $P \in \mathcal{M}$. Knowledge of the EIF is key because its variance defines the semi-parametric efficiency bound of estimating $B(P_0)$, and it serves as a building block for constructing efficient non-parametric estimators. We review relevant ideas from semi-parametric efficiency theory and present the form of the EIF for B in Section 3.

Next, we draw on the Targeted Learning framework, specifically Targeted Minimum Loss-Based Estimation (TMLE), to build an efficient estimator of $B(P_0)$ (van der Laan and Rose, 2011, 2018). Suppose we have an initial estimate of P_0, $P^\circ_n \in \mathcal{M}$. While the plug-in estimator $B(P^\circ_n)$ will typically be biased, the core insight of TMLE is that it is possible to mitigate this bias and achieve asymptotic normality by carefully fluctuating the initial estimate P°_n. In broad strokes, we define a parametric fluctuation $P^\circ_{n,\boldsymbol{\epsilon}}$ with parameter $\boldsymbol{\epsilon} \in \mathbb{R}^p$. An estimate $\boldsymbol{\epsilon}^*_n$ of $\boldsymbol{\epsilon}$ is found by minimizing an empirical risk associated with a carefully chosen loss function. An updated estimate of $B(P_0)$ is then derived by plugging in the fluctuated law, as $B(P^\circ_{n,\boldsymbol{\epsilon}^*_n})$. The key property of the fluctuation and loss function, from which the desirable properties of TMLE are derived, is that the fluctuated law empirically solves the efficient influence function D^*; that is,

$$\frac{1}{n}\sum_{i=1}^{n} D^*(P^\circ_{n,\boldsymbol{\epsilon}^*_n})(O_i) = o_p(n^{-1/2}).$$

It is for this reason that the estimator is called *targeted*, because it solves the efficient influence function of the parameter of interest. Under conditions on the estimator P°_n, it is possible to show that the updated estimate is asymptotically normal and efficient with asymptotic variance given by the variance of the efficient influence function.

Estimators based on TMLE have been developed for a number of marginal structural models in various contexts. These include estimators for vector-valued treatment assignments, time-varying treatments, and instrumental variable designs (Stein et al., 1956; Rosenblum and van der Laan, 2010; Petersen et al., 2014; Toth and van der Laan, 2016; Zheng et al., 2016, 2018). In practice, users may wish to test several alternative working model specifications and loss functions. This is difficult to do using existing approaches, as the estimators (and related software) are constructed for specific classes of working models and loss functions. Our results build off these works by proposing a TMLE based estimation strategy in a generalized data structure and for arbitrary MSM loss functions and working models. In Section 4 we provide a blueprint for building targeted estimators for MSMs with arbitrary working models and loss functions. Our quite general software implementation is made possible by an algorithm that adapts to the choice of loss function and working model using automatic differentiation (Baydin et al., 2018).

Statistical inference can be loosely categorized into frequentist and Bayesian approaches, corresponding to different interpretations of probability (Bayarri and Berger, 2004; Hájek, 2019). TMLE was developed within the frequentist paradigm, drawing on a long line of research in semi-parametric efficiency theory; see Bickel et al. (1993); van der Vaart and Wellner (1996); van der Vaart (1998) among many others. A Bayesian version of TMLE was first discussed in van der Laan (2008), based on the observation that, if a likelihood is adopted for the parametric fluctuation $P^\circ_{n,\boldsymbol{\epsilon}}$, then the parameter $\boldsymbol{\epsilon}$ can be estimated via Bayesian inference. Such an approach was later extended to estimation of average treatment effects and class proportions in an unlabeled dataset (Diaz et al., 2011; Díaz et al., 2020). We build on this research, and describe a novel Bayesian estimator in Section 5 that is the first Bayesian targeted estimator to be proposed for MSMs.

As an illustrative example, we consider the problem of estimating the association between a multi-faceted family planning intervention on contraceptive use in a cohort of women in Malawi using data collected in a randomized field experiment (Karra et al., 2022). We define a Marginal Structural Model to summarize how the treatment effect depends on the participant's number of children. The results are presented using both frequentist and Bayesian approaches in Section 8.

In summary, the rest of the article proceeds as follows. In Section 2 we provide a general definition and notation for Marginal Structural Models. In Section 3 we use results from semi-parametric efficiency theory to derive the efficiency bound for estimating MSMs in non-parametric models. In Section 4 we present an estimator based on Targeted Minimum Loss-Based Estimation that achieves the efficiency bound. In Section 5 we develop a novel targeted Bayesian estimator for MSMs. In Section 6 we discuss details of the software implementation. In Section 7 we share the results of a simulation study. In Section 8 apply the estimators to an example. Finally, we conclude with a discussion in Section 9.

2 Marginal Structural Models

2.1 Statistical Viewpoint

Recall that we have n i.i.d. draws $O_{1:n}$ from a distribution P_0 of a generic variable O. Let $\mathcal{O}$ be the support of P_0. We assume only that P_0 falls in a non-parametric model $\mathcal{M}$ of laws on $\mathcal{O}$. Further, suppose O can be decomposed into two variables, $O = (X, Z)$ (this is always possible, as Z can be set to the trivial set). Let $\mathcal{X}$ and $\mathcal{Z}$ denote the spaces to which X and Z belong. For any $P \in \mathcal{M}$, let $\Psi_P : \mathcal{X} \to \mathbb{R}$ be a functional summary of P, with Ψ_P a member of the function class $\mathcal{F}$. For later convenience, we introduce $\mathcal{T} := \cup_{P\in\mathcal{M}}\Psi_P(\mathcal{X})$ (that is, the union of the images of all Ψ_Ps). We focus on instances where $\mathcal{T}$ is an open set. We also assume that Ψ_P depends on P through a finite tuple of parameters $\bar{Q}_P = \left(\bar{Q}_P^{(0)}, \ldots, \bar{Q}_P^{(J)}\right)$.

> **Example.** Conditional Average Treatment Effect. Let $Z = (A, Y)$, where A is a binary treatment indicator and Y is a binary outcome. Let $g_P(a, x) := P(A = a|X = x)$, and suppose that $g_P(a, X) > 0$ holds P-almost-surely for all $P \in \mathcal{M}$ and $a \in \{0, 1\}$. Let $\bar{Q}_P^{(a)}(X) := \mathbb{E}_P[Y|A = a, X]$, which is defined almost surely on $\mathcal{X}$. Define $\Psi_P^\sharp(x) := \bar{Q}_P^{(1)}(x) - \bar{Q}_P^{(0)}(x)$, interpretable as the expected difference in the outcome for treatment vs. non-treatment within the strata of covariates $X = x$. Appendix 11.5 considers the case when $Y \in \mathbb{R}$.

For convenience, we will write $\psi_0 := \Psi_{P_0}$ to denote the functional under the true data generating distribution P_0. Estimating the functional ψ_0 across its entire domain may be challenging, especially if $\mathcal{X}$ is high dimensional. The idea with MSMs is to seek an approximation of the functional ψ_0 defined by a lower-dimensional working model. In other words, instead of estimating ψ_0, we estimate a parameter for the working model that leads to the best possible approximation of ψ_0 within the working model as measured by the risk induced by a well-chosen, problem-specific loss function.

Formally, a loss function is a function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that, for all $f \in \mathcal{F}$,

$$\mathbb{E}_P[L(\Psi_P(X), \Psi_P(X))] \leq \mathbb{E}_P[L(\Psi_P(X), f(X))].$$

The working model is a collection $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$ of functions $m_{\boldsymbol{\beta}} : \mathcal{X} \to \mathbb{R}$ with $\mathcal{B}$ a parameter of dimension p. For convenience, we write $L_m(t, \boldsymbol{\beta})(X) := L(t(X), m_{\boldsymbol{\beta}}(X))$ for all $t \in \mathcal{T}$ and $\boldsymbol{\beta} \in \mathcal{B}$. The target parameter $B(P) \in \mathcal{B}$ is defined as any solution to the following optimization problem:

$$B(P) \in \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}_P[L_m(\Psi_P, \boldsymbol{\beta})(X)]. \tag{2}$$

Note that $B(P)$ depends on P only through the functional summary Ψ_P and the marginal distribution of X, which we denote Q.

From now on, we assume that there is a unique minimizer of (2) in the case where $P = P_0$. We call this minimizer $\boldsymbol{\beta}_0 := B(P_0)$. Moreover, we focus on loss functions L such that $\boldsymbol{\beta} \mapsto L_m(\psi_0, \boldsymbol{\beta})(X)$ is twice differentiable with a Hessian at $\boldsymbol{\beta}_0$ denoted by $\ddot{L}_m(\psi_0, \boldsymbol{\beta})(X)$. In a nutshell, it therefore holds that $\boldsymbol{\beta}_0$ solves the implicit equation

$$\mathbf{0} = \frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}_{P_0}[L_m(\psi_0, \boldsymbol{\beta})(X)]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

Assuming that the gradient and expected value can be interchanged, then $\boldsymbol{\beta}_0$ also solves the equation

$$\mathbf{0} = \mathbb{E}_{P_0}\left[\ddot{L}_m(\psi_0, \boldsymbol{\beta}_0)(X)\right].$$

This equality will play an important role in the semi-parametric analysis of the target functional B.

Example (cont'd). Let $V \subset X$ be a set of $(p-1)$ *treatment effect modifiers*. We adopt a linear working model with an intercept term

$$m_{\boldsymbol{\beta}} : X \mapsto \boldsymbol{\beta}^\top (1, V)^\top \quad \text{for all } \boldsymbol{\beta} \in \mathcal{B} := \mathbb{R}^p, \tag{3}$$

and a squared-error loss function given by $L(a, b) := (a-b)^2$, hence

$$L_m(\Psi_P^\sharp, \boldsymbol{\beta})(X) = (\Psi_P^\sharp(X) - m_{\boldsymbol{\beta}}(X))^2. \tag{4}$$

The target parameter is then given by

$$B^\sharp(P) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \mathbb{E}_P\left[\left(\Psi_P^\sharp(X) - \boldsymbol{\beta}^\top (1, V)^\top\right)^2\right].$$

Note that when $V = \varnothing$ the parameter $B^\sharp(P)$ reduces to the Average Treatment Effect (ATE):

$$B^\sharp(P) = \mathbb{E}_P\left[\Psi_P^\sharp(X)\right] = \mathbb{E}_P\left[\bar{Q}_P^{(1)}(X) - \bar{Q}_P^{(0)}(X)\right].$$

2.2 Causal Viewpoint

While the main focus of this work is statistical, we review briefly the causal interpretation of the target parameter $B(P)$. The parameter $B(P)$ can be viewed as a projection of the parameter Ψ_P onto a lower-dimensional space. As such, $B(P)$ is causally identifiable under the same conditions for which Ψ_P is identifiable. Formally, let $\mathbb{P}$ the causal law from which P arises, that is, $\mathbb{P}$ is the law of the complete data (including unobservable potential outcomes) and P is a joint marginal law thereof. Let $\Psi_{\mathbb{P}}^C$ be a causal functional summary of $\mathbb{P}$ and $B^C(\mathbb{P})$ be the causal analog to $B(P)$, defined as

$$B^C(\mathbb{P}) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \mathbb{E}_P\left[L_m(\Psi_{\mathbb{P}}^C, \boldsymbol{\beta})(X)\right]. \tag{5}$$

Suppose that there are a set of identification assumptions sufficient to show that $\Psi_P(X) = \Psi_{\mathbb{P}}^C(X)$. It is then straightforward to see that $B(P) = B^C(\mathbb{P})$ under the same conditions by direct substitution of $\Psi_P(X)$ for $\Psi_{\mathbb{P}}^C(X)$ in (5).

Example (cont'd). The parameter $\Psi_P^\sharp$ is identifiable under standard causal assumptions which are reviewed below for completeness. Let $Y(a)$ be the potential outcome under treatment assignment $a \in \{0, 1\}$. Define the causal parameters

$$\Psi_{\mathbb{P}}^{C,\sharp}(X) := \mathbb{E}_{\mathbb{P}}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X],$$
$$B^{C,\sharp}(\mathbb{P}) := \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \mathbb{E}_{\mathbb{P}}\left[\left(\Psi_{\mathbb{P}}^{C,\sharp}(X) - \boldsymbol{\beta}^\top (1, V)^\top\right)^2\right].$$

We seek conditions under which $B^{C,\sharp}(\mathbb{P}) = B^\sharp(P)$. Assume

1. Consistency: $Y = Y(A)$.
2. Positivity: $g_P(a, X) > 0$ for both $a \in \{0, 1\}$, P-almost surely.
3. No unmeasured confounders: under $\mathbb{P}$, $Y(a) \perp\!\!\!\perp A \mid X$ for both $a \in \{0, 1\}$.

Under these conditions,

$$\begin{aligned}\Psi_{\mathbb{P}}^{C,\sharp}(X) &= \mathbb{E}_{\mathbb{P}}[Y(1) \mid X] - \mathbb{E}_{\mathbb{P}}[Y(0) \mid X] \\ &= \mathbb{E}_{\mathbb{P}}[Y(1) \mid A=1, X] - \mathbb{E}_{\mathbb{P}}[Y(0) \mid A=0, X] \text{ (assumptions 2 and 3)} \\ &= \mathbb{E}_P[Y \mid A=1, X] - \mathbb{E}_P[Y \mid A=0, X] \text{ (assumption 1)} \\ &= \Psi_P^{\sharp}(X).\end{aligned}$$

It then follows directly that $B^{C,\sharp}(\mathbb{P}) = B^{\sharp}(P)$.

3 Semi-Parametric Efficiency Theory

Our goal is to find an asymptotically efficient estimator of $\boldsymbol{\beta}_0 := B(P_0) \in \mathbb{R}^p$, the value of the parameter of interest under the true data generating distribution P_0. We first review some semi-parametric efficiency theory through which we can derive the non-parametric efficiency bound for estimating $\boldsymbol{\beta}_0$ (van der Vaart, 2002).

In this section, we derive the semi-parametric efficiency bound for estimating $\boldsymbol{\beta}_0$ in a non-parametric model $\mathcal{M}$ (that is, $\mathcal{M}$ is the set of all laws P on $\mathcal{X} \times \mathcal{Z}$ such that $B(P)$ is well-defined). The semi-parametric efficiency bound for estimating $B(P)$ is defined via the concept of the hardest parametric submodel. For any $P \in \mathcal{M}$, for any integer $k \geq 1$ and for all

$$s \in \mathcal{S} := \left\{h \in (L_0^2(P))^k : h \neq 0, \|h\|_\infty < \infty, \mathbb{E}_P[h(O)h(O)^\top] \text{ invertible}\right\},$$

define a parametric submodel $\mathcal{P}_s := \{P_{s,\epsilon} : \boldsymbol{\epsilon} \in \mathbb{R}^k, \|\boldsymbol{\epsilon}\|_\infty < \infty\} \subset \mathcal{M}$, characterized by $dP_{\epsilon,s} = (1 + \boldsymbol{\epsilon}^\top s)dP$. The invertibility condition in the definition of $\mathcal{S}$ amounts to a condition that $\mathcal{P}_s$ is identifiable. Note that $P_{s,\epsilon} = P$ at $\epsilon = 0$, and the score of P equals s at $\epsilon = 0$. That is, $\mathcal{P}_s$ is a fluctuation of P in the direction s.

Assume the derivative $\frac{\partial}{\partial \boldsymbol{\epsilon}} B(P_{\epsilon,s})$ exists at $\boldsymbol{\epsilon} = \mathbf{0}$ for any $s \in \mathcal{S}$. The Cramér-Rao bound for estimating $B(P)$ within the submodel $\mathcal{P}_s$ (that is, the lowest asymptotic variance possible for an unbiased estimator of $B(P)$, in the sense of the Loewner order: for any A, B Hermitian matrices, $A \geq B$ if $A - B$ is positive semi-definite) is given by

$$\left[\frac{\partial}{\partial \boldsymbol{\epsilon}} B(P_{\epsilon,s})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}}\right]^\top \mathbb{E}_P[s(O)s(O)^\top]^{-1} \left[\frac{\partial}{\partial \boldsymbol{\epsilon}} B(P_{\epsilon,s})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}}\right].$$

Assume further that there exists a continuous, bounded, linear map $\dot{B}$ mapping closure($\mathcal{S}$) to $\mathbb{R}^{k \times p}$ such that, for all $s \in \mathcal{S}$,

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} B(P_{\epsilon,s})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} = \dot{B}(P_{\epsilon,s}).$$

We then say that B is *pathwise differentiable* at P with respect to $\{\mathcal{P}_s : s \in \mathcal{S}\}$. By the Riesz representation theorem, the derivative can be expressed as

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} B(P_{\epsilon,s})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} = \mathbb{E}_P[s(O)D^*(P)(O)^\top]$$

for a function $D^*(P) \in (L_0^2(P))^p$ which is called the *efficient influence function* (EIF) of the parameter B at P. For later convenience, let $\lambda^*(P) := D^*(P)D^*(P)^\top$. We can then rewrite the Cramér-Rao bound as

$$\mathbb{E}_P[s(O)D^*(P)(O)^\top]^\top \mathbb{E}_P[s(O)s(O)^\top]^{-1} \mathbb{E}_P[s(O)D^*(P)(O)^\top].$$

The *hardest parametric submodel*, i.e. the submodel for which estimating $B(P)$ is the hardest, is the submodel with the largest Cramér-Rao bound (for the Loewner order). Slightly tedious algebra and the Cauchy-Schwarz inequality imply that the largest Cramér-Rao bound is

$$\mathbb{E}_P[\lambda^*(P)(O)].$$

The semi-parametric efficiency bound for estimating $B(P)$ within a non-parametric model $\mathcal{M}$ is then defined as $\mathbb{E}_P[\lambda^*(P)(O)]$. Deriving the form of $D^*(P)$ is therefore crucial if we wish to construct efficient non-parametric estimators of $B(P)$.

In the rest of the section the parametric submodel $\mathcal{P}_s$ is a theoretical tool used to derive the form of D^*, so for simplicity we can choose $\boldsymbol{\epsilon}$ to be of dimension one. In the next theorem, we present detailed conditions under which $B(P)$ is pathwise differentiable, and characterize the form of $D^*(P)$.

Result 1 (Efficient Influence Function of B)**.** *Let us make the following assumption: for all $P \in \mathcal{M}$ and for all $s \in L_0^2(P)$ such that $s \neq 0$, $\|s\|_\infty < \infty$, if $\left\{P_\epsilon : |\epsilon| < \|s\|_\infty^{-1}\right\} \subset \mathcal{M}$ is characterized by $\frac{dP_\epsilon}{dP} = 1 + \epsilon s$ for every $\epsilon \in \mathbb{R}$ such that $|\epsilon| < \|s\|_\infty^{-1}$, then the mapping $\epsilon \mapsto \Psi_{P_\epsilon}(X)$ is almost surely differentiable at 0 and*

$$\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\bigg|_{\epsilon=0} = \mathbb{E}_P\left[\Delta^*(P)(O)s(O) \mid X\right]$$

for a function $\Delta^*(P) \in L_0^2(P)$. Under further assumptions stated in Theorem 3, the target functional $P \mapsto B(P)$ is pathwise differentiable at every $P \in \mathcal{M}$, with an efficient influence function $D^*(P)$ given by

$$D^*(P)(O) = M^{-1}\left[D_1^*(P)(O) + D_2^*(P)(X)\right],$$

where $D_1^*(P), D_2^*(P) \in L_0^2(P)$ are given by

$$\begin{aligned} D_1^*(P)(O) &= \nabla \dot{L}(\Psi_P(X), B(P))(X) \times \Delta^*(P)(O), \\ D_2^*(P)(X) &= \dot{L}(\Psi_P(X), B(P))(X), \end{aligned}$$

and the normalizing matrix M is given by

$$M = -\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B(P))(X)\right].$$

The full statement (Theorem 3) and its proof are given in Appendix 11.1.

Example (cont'd). We can derive the efficient influence function $D_\sharp^*(P)$ for $B^\sharp$ at P by applying Theorem 1. First, we prove a lemma giving the conditional efficient influence function $\Delta_\sharp^*(P)$ for $\Psi_P^\sharp$ at P.

Lemma 1 (Conditional Efficient Influence Function of $\Psi^\sharp$)**.** *The functional summary $\Psi^\sharp$ satisfies the first assumption of Result 1 with a conditional efficient influence function $\Delta_\sharp^*(P) \in L_0^2(P)$ at any $P \in \mathcal{M}$ given by*

$$\Delta_\sharp^*(P)(O) = \left\{\frac{\mathbb{I}(A=1)}{g_P(1,X)} - \frac{\mathbb{I}(A=0)}{g_P(0,X)}\right\}(Y - \bar{Q}_P^{(A)}(X)).$$

The proof is given in Appendix 11.2. Note that the form of $\Delta_\sharp^*$ is recognizable as a part of the efficient influence function for the Average Treatment Effect (van der Laan and Rose, 2011, Chapter 5). With this result in hand, we are ready to state the efficient influence function for the parameter $B^\sharp$.

Theorem 1 (Efficient Influence Function of $B^\sharp$)**.** *Suppose that Assumptions (2)-(6) of Theorem 3 are satisfied. Then the target functional $P \mapsto B^\sharp(P)$ is pathwise differentiable at every $P \in \mathcal{M}$, with an efficient influence function $D_\sharp^*(P)$ given by*

$$D_\sharp^*(P)(X, A, Y) = M^{-1}\left\{D_{1,\sharp}^*(P)(X, A, Y) + D_{2,\sharp}^*(P)(X)\right\},$$

where

$$\begin{aligned} D_{1,\sharp}^*(P)(X, A, Y) &= \left\{\frac{I(A=1)}{g_P(1,X)} - \frac{I(A=0)}{g_P(0,X)}\right\}(Y - \bar{Q}_P^{(A)}(X))(1, V)^\top, \\ D_{2,\sharp}^*(P)(X) &= (\Psi_P^\sharp(X) - B^\sharp(P)^\top(1, V)^\top)(1, V)^\top, \end{aligned}$$

and the normalizing matrix M is given by

$$M = -\mathbb{E}_P\left[(1, V)^\top(1, V)\right].$$

The proof is provided in Appendix 11.3.

Remark The form of the EIF resembles closely that of the Average Treatment Effect (ATE). In fact, when $V = \varnothing$ then $D^*_\sharp(P)$ reduces to the EIF of the ATE:

$$D^*_{1,\sharp}(P)(X, A, Y) = \left\{ \frac{I(A=1)}{g_P(1,X)} - \frac{I(A=0)}{g_P(0,X)} \right\} (Y - \bar{Q}_P^{(A)}(X)),$$
$$D^*_{2,\sharp}(P)(X) = \Psi^\sharp_P(X) - \mathbb{E}_P\left[\Psi^\sharp_P(X)\right].$$

Remark The variance of the EIF is given by

$$\mathbb{E}_P[\lambda^*(P)(O)] = M^{-1}\mathbb{E}_P\left[\left(\frac{\mathrm{Var}(Y \mid A, X)}{g_P(A,X)^2} + \left(\Psi^\sharp_P(X) - B^\sharp(P)^\top(1,V)^\top\right)^2\right)(1,V)(1,V)^\top\right] M^{-1}.$$

4 Targeted Minimum Loss-Based Estimation

In this section, we describe an efficient and asymptotically normal estimator for $\boldsymbol{\beta}_0$ based on Targeted Minimum Loss-Based Estimation (TMLE; van der Laan and Rose (2011, 2018)).

Notation The target parameter $B(P)$ and the efficient influence function $D^*(P)$ of B at P only depend on P through several nuisance parameters. As we will be only estimating these nuisance parameters, rather than all of P, it is helpful to introduce notation that emphasizes the parts of P necessary for estimating the target parameter and efficient influence function.

- Nuisance parameters for $B(P)$: note that $B(P)$ depends on P only through (a) the marginal distribution of X under P, which we call Q_P, and (b) the functional Ψ_P, which itself only depends on $\bar{Q}_P$ and possibly Q_P. For convenience, we will write $B(P)$ and $B(\bar{Q}_P, Q_P)$ interchangeably.
- Nuisance parameters for $D^*(P)$: recall that the EIF $D^*(P)$ is the sum of $D^*_1(P)$ and $D^*_2(P)$. The first component $D^*_1(P)$ may depend on P through Ψ_P, Q_P, and additional nuisance parameters we call η_P. The function $D^*_2(P)$ only depends on P through Ψ_P and Q_P. As such, we will write $D^*_1(\bar{Q}_P, Q_P, \eta_P)$, $D^*_2(\bar{Q}_P, Q_P)$, and $D^*(\bar{Q}_P, Q_P, \eta_P)$ interchangeably with $D^*_1(P)$, $D^*_2(P)$, and $D^*(P)$.

For convenience, let $\mathcal{Q} = \{Q_P : P \in \mathcal{M}\}$ denote the parameter space of Q_P, denote $Q_0 := Q_{P_0}$, and define $\eta_0 := \eta_{P_0}$.

Suppose we have initial estimates $\bar{Q}^0_n \in \mathcal{Q}$, Q^0_n, and η_n of $\bar{Q}_0$, Q_0, and η_0. From now on, we will always choose Q^0_n to be the empirical distribution of $X_1, \ldots, X_n$, but we do not yet choose a particular estimating strategy for $\bar{Q}^0_n$ and η_n. Given these initial estimates, we could estimate $\boldsymbol{\beta}_0$ via the plug-in estimator $B(\bar{Q}^0_n, Q^0_n)$ (which is indeed well-defined). However, although $\bar{Q}^0_n$ and Q^0_n may be good estimators of $\bar{Q}_0$ and Q_0, there is no guarantee that plugging them into the target parameter yields a good estimator of $\boldsymbol{\beta}_0$. The idea of TMLE is to reduce the bias of the plug-in estimator by iteratively updating the initial estimates $\bar{Q}^0_n$ and Q^0_n so as to target the parameter of interest $\boldsymbol{\beta}_0$. Here, the iterative updates may incorporate the additional nuisance parameters η_n. Given mild conditions on $\bar{Q}^0_n$ and η_n the resulting targeted estimator is asympotically normal and efficient.

Now we describe in more detail the updating procedure at the heart of TMLE. The iterative updates are based on fluctuations of the current estimators of $\bar{Q}_0$ and Q_0. First, choose loss functions $\mathcal{L}_j : \mathbb{R} \times \mathcal{O} \to \mathbb{R}$ corresponding to each component $\bar{Q}^{(j)}_P$ of $\bar{Q}_P = \left(\bar{Q}^{(0)}_P, \ldots, \bar{Q}^{(J)}_P\right)$ such that the following conditions are satisfied for all $j = 0, \ldots, J$:

(L1) $\bar{Q}^{(j)}_0 = \arg\min \left\{ \mathbb{E}_{P_0}[\mathcal{L}_j(\bar{Q}_j(X), O)] : \bar{Q}_j \in \{\bar{Q}^{(j)}_P : P \in \mathcal{M}\} \right\}$

(L2) For each $P \in \mathcal{M}$, it holds P-almost surely that $t' \mapsto \mathcal{L}_j(t', O)$ is differentiable at every $t \in \mathbb{R}$ with derivative $\dot{\mathcal{L}}_j(t, O) \in \mathbb{R}$.

Next, introduce a parametric fluctuation model satisfying the following conditions:

(M1) For any $P \in \mathcal{M}$ with corresponding $\bar{Q}_P$, Q_P, and η_P, we can define a fluctuation model

$$\left\{(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}, \ldots, \bar{Q}^{(J)}_{P,\boldsymbol{\epsilon}}, Q_{P,\boldsymbol{\epsilon}}) : \boldsymbol{\epsilon} \in \mathbb{R}^p\right\}$$

such that:

(a) If $\boldsymbol{\epsilon} = \mathbf{0}$, then $\bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}} = \bar{Q}^{(j)}_P$ for all $j = 0, \ldots, J$ and $Q_{P,\boldsymbol{\epsilon}} = Q_P$.

(b) It holds P-almost surely that the mappings $\boldsymbol{\epsilon} \mapsto \mathcal{L}_j(\bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}}(X), O)$ $(j = 0, \ldots, J)$ and $\boldsymbol{\epsilon} \mapsto -\log Q_{P,\boldsymbol{\epsilon}}(X)$ are differentiable at $\boldsymbol{\epsilon} = \mathbf{0}$, and $D^*(P)(O)$ belongs to

$$\text{Span}\left(\frac{\partial}{\partial \boldsymbol{\epsilon}}\mathcal{L}_0(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}(X), O)\Big|_{\boldsymbol{\epsilon}=\mathbf{0}}, \ldots, \frac{\partial}{\partial \boldsymbol{\epsilon}}\mathcal{L}_J(\bar{Q}^{(J)}_{P,\boldsymbol{\epsilon}}(X), O)\Big|_{\boldsymbol{\epsilon}=\mathbf{0}}, \frac{\partial}{\partial \boldsymbol{\epsilon}}\left(-\log Q_{P,\boldsymbol{\epsilon}}\right)(X)\Big|_{\boldsymbol{\epsilon}=\mathbf{0}}\right).$$

Given loss functions and a fluctuation model satisfying the above conditions, conduct the following iterative procedure:

1. Start with the initial estimates $\bar{Q}^{(j),0}_n$ $(j = 0, \ldots, J)$, Q^0_n, η_n.
2. For $k \geq 1$, recursively let $\bar{Q}^{(j),k}_n = \bar{Q}^{(j),k-1}_{n,\boldsymbol{\epsilon}^k_n}$ $(j = 0, \ldots, J)$ and $Q^k_n = Q^{k-1}_{n,\boldsymbol{\epsilon}^k_n}$, where

$$\boldsymbol{\epsilon}^k_n = \arg\min_{\boldsymbol{\epsilon} \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left[\sum_{j=0}^J \left[\mathcal{L}_j(\bar{Q}^{(j),k-1}_{n,\boldsymbol{\epsilon}}(X_i), O_i)\right] - \log Q^{k-1}_{n,\boldsymbol{\epsilon}}(X_i)\right],$$

and stop when $\|\boldsymbol{\epsilon}^k_n\| \approx 0$ (rigorously, when $\boldsymbol{\epsilon}^k_n = o_P(n^{-1/2})$).

3. Set $\bar{Q}^*_n = \left(\bar{Q}^{(0),k}_n, \ldots, \bar{Q}^{(J),k}_n\right)$ and $Q^*_n = Q^k_n$, where k is the final iteration of the above step.

The targeted estimator is the plug-in estimator given by $\hat{\boldsymbol{\beta}}^*_n = B(\bar{Q}^*_n, Q^*_n)$. The desirable statistical properties enjoyed by this estimator derive from the fact that the final update approximately solves the estimating equation $P_n\left[D^*(\bar{Q}^*_n, Q^*_n, \eta_n)\right] = o_P(n^{-1/2})$.[1] The following theorem states conditions under which $\boldsymbol{\beta}^*_n$ is asymptotically normal and efficient, which serves as the basis for conducting valid statistical inference.

Theorem 2 (Asymptotic normality and efficiency of $\boldsymbol{\beta}^*_n$)**.** *Let us assume that*

1. *Rate of convergence of second-order remainder:* $\hat{\boldsymbol{\beta}}^*_n - \boldsymbol{\beta}_0 + P_0 D^*(P^*_n) = o_P(n^{-1/2})$.
2. *Donsker conditions:* $D^*(P^*_n)$ *is in a* P_0*-Donsker class with probability tending to one, and the random squared norm* $P_0[(D^*(P^*_n) - D^*(P_0))^2] = o_P(n^{-1/2})$.

Then the estimator $\hat{\boldsymbol{\beta}}^*_n$ *is asymptotically linear with the form*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^*_n - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n D^*(P_0)(O_i) + o_P(1),$$

which implies that $\hat{\boldsymbol{\beta}}^*_n$ *is asymptotically normal and efficient:*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^*_n - \boldsymbol{\beta}_0\right) \rightsquigarrow N\left(0, P_0[\lambda^*(P_0)]\right).$$

The proof is given in Appendix 11.4. Note that the Donsker conditions can be relaxed through the use of cross-validation (Zheng and van der Laan, 2010).

So far we have not given an explicit form for the loss functions $\mathcal{L}_0, \ldots, \mathcal{L}_J$ nor the fluctuation model. The explicit forms of the $\mathcal{L}_0, \ldots, \mathcal{L}_J$ will depend on the structure of O and the form of the functional summary Ψ_P. Likewise, we cannot give a complete specification of the fluctuation model as it too depends on the choices for $\mathcal{L}_0, \ldots, \mathcal{L}_J$ and the form of Ψ_P. In addition, there may be multiple fluctuation models that satisfy the required conditions. We give a blueprint for one particular fluctuation model satisfying

[1] From now on we will write $P[f]$ for the integral $\int f dP$.

the required conditions that can be easily adapted to specific problems. This blueprint also yields a useful constraint on the form of the $\mathcal{L}_0, \ldots, \mathcal{L}_J$.

For any $P \in \mathcal{M}$ with corresponding Ψ_P, $\bar{Q}_P$, Q_P, and η_P, characterize the fluctuation model by setting

$$\bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}}(O) = \phi^{-1}\left(\phi(\bar{Q}^{(j)}_P(O)) + H_j(O)\boldsymbol{\epsilon}^\top \nabla \dot{L}_m(\Psi_P, B(P))(X)\right) \quad \text{for each } j = 0, \ldots, J,$$
$$\text{and} \quad Q_{P,\boldsymbol{\epsilon}}(X) = C(\boldsymbol{\epsilon}) \exp\left(\boldsymbol{\epsilon}^\top \dot{L}_m(\Psi_P, B(P))(X)\right) Q_P(X),$$

where the so-called "clever-covariates", which need to be chosen, are functions $H_j : \mathcal{O} \to \mathbb{R}$ $(j = 0, \ldots, J)$ and the constant $C(\boldsymbol{\epsilon})$ is chosen such that $Q_{P,\boldsymbol{\epsilon}}$ is well-defined. We also assume that the transformation link has an inverse ϕ^{-1}, and that this inverse has derivative $\dot{\phi}^{-1}$. It is easy (sic) to see that $\bar{Q}^{(j)}_{P,\mathbf{0}} = \bar{Q}^{(j)}_P$ $(j = 0, \ldots, J)$ and $Q_{P,\mathbf{0}} = Q_P$, so (M1a) is satisfied. Next we derive a constraint under which condition (M1b) is satisfied. First, see that

$$\begin{aligned}\frac{\partial}{\partial \epsilon} \log Q_{P,\boldsymbol{\epsilon}}(X)\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} &= \dot{L}_m(\Psi_P, B(P))(X) \\ &= D^*_2(\Psi_P, Q_P)(X).\end{aligned}$$

In addition, by condition (L1) we can compute

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\epsilon}} \mathcal{L}_j(\bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}}(O), O)\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} &= \dot{\mathcal{L}}_j(\bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}}(O), O)\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} \frac{\partial}{\partial \boldsymbol{\epsilon}} \bar{Q}^{(j)}_{P,\boldsymbol{\epsilon}}(O)\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} \\ &= \dot{\mathcal{L}}_j(\bar{Q}^{(j)}_P(O), O) \times \dot{g}^{-1}(g(\bar{Q}^{(j)}_P(O)) \times H_j(O) \times \nabla \dot{L}_m(\Psi_P, B(P))(X).\end{aligned}$$

Recall that condition (M1b) requires that $D^*(P)(O)$ be included in the linear span of the two gradients from the two above displays. Therefore, the only way for (M2b) to be satisfied is if

$$\sum_{j=0}^{J} \dot{\mathcal{L}}_j(\bar{Q}^{(j)}_P(O), O)\dot{\phi}^{-1}(\phi(\bar{Q}^{(j)}_P(O)))H_j(O) \propto \Delta^*(P)(O).$$

In practice, Δ^* often includes a term resembling a residual. The loss functions $\mathcal{L}_j$ $(j = 0, \ldots, J)$ are then chosen such that their summed derivatives equals the residual term, and the remaining part of Δ^* is integrated into the H_j $(j = 0, \ldots, J)$.

Example (cont'd). Now that we are working in the context of a particular example we can fill in the details of the targeted estimator. First, note that $\Psi^\sharp_P$ depends on P only through the parameters $\bar{Q}_P = (\bar{Q}^{(0)}_P, \bar{Q}^{(1)}_P)$, and $D^*_{1,\sharp}(P)$ depends on P through the additional nuisance parameters $\eta_P(X) = (g_P(0, X), g_P(1, X))$. For any $P \in \mathcal{M}$ with corresponding Ψ_P, $\bar{Q}_P$, Q_P and η_P, characterize the fluctuation model as follows, using the transformation $\phi = \text{logit}$:

$$\begin{aligned}\text{logit}\left(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}(X)\right) &= \text{logit}\left(\bar{Q}^{(0)}_P(X)\right) + H_0(O)\boldsymbol{\epsilon}^\top(1, V)^\top, \\ \text{logit}\left(\bar{Q}^{(1)}_{P,\boldsymbol{\epsilon}}(X)\right) &= \text{logit}\left(\bar{Q}^{(1)}_P(X)\right) + H_1(O)\boldsymbol{\epsilon}^\top(1, V)^\top, \\ Q_{P,\boldsymbol{\epsilon}}(X) &= C(\boldsymbol{\epsilon}) \exp\left(\boldsymbol{\epsilon}^\top(\psi^\sharp_P(X) - B(P)^\top(1, V)^\top)(1, V)^\top\right) Q_X(X).\end{aligned}$$

The clever covariates are given by

$$\begin{aligned}H_0(O) &= -\frac{I(A = 0)}{g_P(0, X)}, \\ H_1(O) &= \frac{I(A = 1)}{g_P(1, X)}.\end{aligned}$$

Note that the fluctuations are quite similar to those commonly adopted to define a TMLE for the ATE (van der Laan and Rose, 2011, Chapter 5). The loss functions are chosen to be characterized

by

$$\mathcal{L}_0(t, O) = -I(A = 0)\left[Y \log(t) + (1 - Y)\log(1 - t)\right],$$
$$\mathcal{L}_1(t, O) = -I(A = 1)\left[Y \log(t) + (1 - Y)\log(1 - t)\right]$$

(both $t \in (0, 1)$). To ensure this is a valid setup, we need to check conditions (L1), (L2), and (M1). (L1) is satisfied for this choice of log-likelihood loss function (van der Laan and Rose, 2011). The derivatives of the loss functions are

$$\dot{\mathcal{L}}_0(t, O) = -I(A = 0)\left[\frac{Y - t}{(1 - t)t}\right],$$
$$\dot{\mathcal{L}}_1(t, O) = -I(A = 1)\left[\frac{Y - t}{(1 - t)t}\right]$$

(both $t \in (0, 1)$). Therefore, (L2) is satisfied. To check (M1), see that, for $\phi = \text{logit}$, $\dot{\phi}^{-1}(\phi(t)) = (1 - t)t$. Therefore

$$\dot{\mathcal{L}}_0(\bar{Q}_P^{(0)}(X), O)\dot{\phi}^{-1}(\phi(\bar{Q}_P^{(0)}(X)))H_0(O) + \dot{\mathcal{L}}_1(\bar{Q}_P^{(1)}(X), O)\dot{\phi}^{-1}(\phi(\bar{Q}_P^{(1)}(X)))H_1(O)$$
$$= -\left\{\frac{I(A = 1)}{g_P(1, X)} - \frac{I(A = 0)}{g_P(0, X)}\right\}(Y - \bar{Q}_P^{(A)}(X)) = -\Delta_P^*(X),$$

which shows that (M1) is satisfied.

5 Bayesian Inference

The key observation behind Bayesian Targeted Maximum Likelihood Estimation is that when the TMLE loss functions can be interpreted as log-likelihoods, then the optimization step in the iterative procedure of TMLE simply corresponds to maximum likelihood estimation of $\boldsymbol{\epsilon}$. Furthermore, due to the likelihood interpretation, we can also use Bayesian inference to estimate $\boldsymbol{\epsilon}$ via a straightforward application of Bayes' theorem. In this section, we formalize this application of Bayesian inference, and present an oracle Bernstein von-Mises type result that suggests the resulting Bayesian targeted estimator converges asymptotically to the truth with optimal variance given by the variance of the efficient influence function. By oracle result, we mean that the Berstein von-Mises theorem concerns an idealized version of the Bayesian targeted estimator that we define below.

Suppose a set of loss functions $\mathcal{L}_j$, $j = 0, \ldots, J$ has been chosen such that conditions (L1) and (L2) are satisfied. In addition, suppose we have already chosen for any $P \in \mathcal{M}$ with corresponding $\bar{Q}_P$, Q_P, and η_P a well-defined fluctuation submodel

$$\left\{(\bar{Q}_{P,\boldsymbol{\epsilon}}^{(0)}, \ldots, \bar{Q}_{P,\boldsymbol{\epsilon}}^{(J)}, Q_{P,\boldsymbol{\epsilon}}) : \boldsymbol{\epsilon} \in \mathbb{R}^p\right\}$$

satisfying condition (M1). In this section, in order to reframe Targeted Maximum Likelihood Estimation in a Bayesian framework, we focus on the case where it is possible, for any $P \in \mathcal{M}$, to define an additional nuisance parameter σ_P and a submodel $\mathcal{F} = \{F_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in \mathbb{R}^p\} \subset \mathcal{M}$ such that the log-likelihood of O under each $F_{\boldsymbol{\epsilon}}$ writes as

$$\log f(O \mid \boldsymbol{\epsilon}) = h_1(\sigma_P(O)) - h_2(\sigma_P(O))\sum_{j=1}^{J}\mathcal{L}_j(\bar{Q}_{P,\boldsymbol{\epsilon}}^{(j)}(O), O) + \log Q_{P,\boldsymbol{\epsilon}}(X) + \text{constant}, \tag{6}$$

where h_1 and h_2 are transformations of the nuisance parameter σ_P. It is often possible to construct such $F_{\boldsymbol{\epsilon}}$ satisfying the above condition by choosing a conditional probability distribution for Y given X with a conditional log-likelihood resembling the TMLE loss functions.

Example (cont'd). The TMLE loss functions $\mathcal{L}_0$ and $\mathcal{L}_1$ in this example can be interpreted as the negative conditional log-likelihoods of Y under Bernoulli distributions. This suggests constructing $F_{\boldsymbol{\epsilon}}$ such that Y given A and X follows a Bernoulli distribution. Fix arbitrarily $P \in \mathcal{M}$ and let $F_{\boldsymbol{\epsilon}}$

be characterized by

$$\begin{aligned} X &\sim Q_{P,\boldsymbol{\epsilon}}, \\ A \mid X &\sim \text{Bernoulli}(1/2), \\ Y \mid X, A &\sim \text{Bernoulli}\left(\bar{Q}^{(A)}_{P,\boldsymbol{\epsilon}}(X)\right), \end{aligned}$$

where

$$\begin{aligned} Q_{P,\boldsymbol{\epsilon}}(X) &= C(\boldsymbol{\epsilon}) \exp\left(M^{-1}\boldsymbol{\epsilon}^\top(\Psi^\sharp_P(X) - B(P)^\top(1,V)^\top)(1,V)^\top\right) Q_X(X), \\ \text{logit}(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}(X)) &= \text{logit}(\bar{Q}^{(0)}_P(X)) + H_0(O)\boldsymbol{\epsilon}^\top M^{-1}(1,V)^\top, \\ \text{logit}(\bar{Q}^{(1)}_{P,\boldsymbol{\epsilon}}(X)) &= \text{logit}(\bar{Q}^{(1)}_P(X)) + H_1(O)\boldsymbol{\epsilon}^\top M^{-1}(1,V)^\top, \end{aligned}$$

with $M = -\mathbb{E}_P[(1,V)^\top(1,V)]$. The only difference between these fluctuations and those used for the frequentist TMLE are in the inclusion of the normalizing matrix M, which is included so that the conditions of the forthcoming Bernstein von-Mises theorem are satisfied. The conditional distribution of A is included so that the model is fully specified, although as in the frequentist TMLE this conditional distribution is not fluctuated. No additional nuisance parameters σ_P are needed in this example; as such, we set $h_1(\cdot) = 0$ and $h_2(\cdot) = 1$ (see Appendix 11.5 for an example where σ_P is non-empty). The conditional log-likelihood of O under $F_{\boldsymbol{\epsilon}}$ is then given by

$$\begin{aligned} \log f\left(O|\boldsymbol{\epsilon}\right) &= Y \log\left(\bar{Q}^{(A)}_{P,\boldsymbol{\epsilon}}(X)\right) + (1-Y)\log\left(1 - \bar{Q}^{(A)}_{P,\boldsymbol{\epsilon}}(X)\right) + \log Q_{P,\boldsymbol{\epsilon}}(X) + \text{constant} \\ &= \left(\mathcal{L}_0\left(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}(X), O\right) + \mathcal{L}_0\left(\bar{Q}^{(0)}_{P,\boldsymbol{\epsilon}}(X), O\right)\right) + \log Q_{P,\boldsymbol{\epsilon}}(X) + \text{constant}, \end{aligned}$$

which satisfies (6).

Next, we discuss Bayesian estimation of the parameter $\boldsymbol{\epsilon}$ within the model $\mathcal{F} = \{F_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in \mathbb{R}^p\}$. Let π_ϵ be a prior distribution for $\boldsymbol{\epsilon}$. Fix arbitrarily $P \in \mathcal{M}$, with corresponding $\bar{Q}_P$, Q_P, η_P, and σ_P. Application of Bayes' theorem yields a posterior distribution of $\boldsymbol{\epsilon}$ given by

$$\Pi_\epsilon(\boldsymbol{\epsilon}|O_{1:n}, \bar{Q}_P, Q_P, \eta_P, \sigma_P) \propto \pi_\epsilon(\boldsymbol{\epsilon}) \prod_{i=1}^n f(O_i \mid \boldsymbol{\epsilon}, \bar{Q}_P, Q_P, \eta_P, \sigma_P).$$

We can then find a posterior distribution for $\boldsymbol{\beta}$, the object of interest, by mapping the posterior of $\boldsymbol{\epsilon}$ onto a posterior for $\boldsymbol{\beta}$. For convenience, we write $\vartheta : \boldsymbol{\epsilon} \mapsto B(F_{\boldsymbol{\epsilon}})$. The posterior distribution Π_β of $\boldsymbol{\beta}$ is then the image of Π_ϵ under ϑ.

We are unlikely to have any prior information for $\boldsymbol{\epsilon}$ directly. On the contrary, we might have prior information for $\boldsymbol{\beta}$. As such, we set a prior on $\boldsymbol{\beta}$, which we then map back to a prior on $\boldsymbol{\epsilon}$. If ϑ has an inverse and if ϑ is differentiable with derivative $\dot{\vartheta} \in \mathbb{R}^{p\times p}$, then a prior distribution π_β for $\boldsymbol{\beta}$ is mapped to a prior distribution π_ϵ on $\boldsymbol{\epsilon}$ by the formula

$$\pi_\epsilon(\boldsymbol{\epsilon}) = \pi_{\boldsymbol{\beta}}\left(\vartheta(\boldsymbol{\epsilon})\right)\left|\det\left(\dot{\vartheta}(\boldsymbol{\epsilon})\right)\right|. \tag{7}$$

Next, we present a Bernstein von-Mises type theorem for an oracular fluctuation model that is built using the **true** nuisance parameters under P_0. The proof shows that in this setting, the Bayesian TMLE converges to the truth with variance given by the variance of the efficient influence function. Denote by $\mathcal{F}^0 = \{F^0_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in \mathbb{R}^p\}$ the submodel through P_0 satisfying (6). The following theorem only holds if the form of $F^0_{\boldsymbol{\epsilon}}$ satisfies certain conditions, one of which being that $\mathcal{F}^0$ be *locally asymptotically normal*. We provide a definition for this property in Appendix 11.6, where we also state and prove Theorem 4, the full version of the result below.

Let $N(\mu, \Sigma)$ denote the multivariate normal distribution with mean vector μ and covariance matrix Σ. Let

$$\Pi^0_{\sqrt{n}(\beta-\beta_0)}\left(\cdot|O_{1:n}\right) := \Pi_{\sqrt{n}(\beta-\beta_0)}\left(\cdot\middle|O_{1:n}, \bar{Q}_0, Q_0, \eta_0, \sigma_0\right)$$

be the posterior for $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ corresponding to the submodel $\mathcal{F}^0$.

Result 2 (Oracular Bernstein von-Mises)**.** *Under assumptions stated in Theorem 4,*

$$\|\Pi^0_{\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}(\cdot \mid O_{1:n}) - N(\Delta^0_n, P_0[\lambda^*(P_0)])\|_1 = o_P(1)$$

where

$$\Delta^0_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} P_0[\lambda^*(P_0)]^{-1} D^*(P_0)(O_i).$$

The full statement (Theorem 4) and its proof are given in Appendix 11.6.

> **Example (cont'd).** We also show in Appendix 11.6 that the above result applies to the running example.

So far we have defined a family of posterior distributions of $\boldsymbol{\beta}$ indexed by $\bar{Q}_P$, Q_P, η_P, and σ_P. The oracular Bernstein von-Mises proof shows that, for a fluctuation $\mathcal{F}^0$ built using the true nuisance parameters $\bar{Q}_0, Q_0, \eta_0$, and σ_0, the posterior distribution of $\boldsymbol{\epsilon}$ converges to the truth with variance given by the variance of the efficient influence function. Since in practice we have to estimate the values of these nuisance parameters, we turn to a closely related posterior from the full family of posterior distributions, the one indexed by, and conditional on, the estimates $\bar{Q}^*_n, Q^*_n$, η_n, and σ_n — recall that $\bar{Q}^*_n, Q^*_n$ are the final fluctuated estimates of $\bar{Q}_0$ and Q_0 from the frequentist TMLE, η_n is an estimate of the nuisance parameter η, and the newly introduced σ_n is an estimate of σ_{P_0}. We call this posterior a *targeted posterior*, and we hope that it inherits favorable properties from the frequentist TMLE.

Formally, define

$$\Pi^*_{\sqrt{n}(\beta-\beta_0)}(\cdot|O_{1:n}) := \Pi_{\sqrt{n}(\beta-\beta_0)}(\cdot|O_{1:n}, \bar{Q}^*_n, Q^*_n, \eta_n, \sigma_n)$$

to be the targeted targeted posterior of $\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$. Under an assumption that the estimates $\bar{Q}^*_n, Q^*_n, \eta_n$, and σ_n are consistent, it is reasonable to think that this posterior will converge to $\Pi^0_{\sqrt{n}(\beta-\beta_0)}(\cdot \mid O_{1:n})$. By appealing to Result 2 (that is, Theorem 4), it would therefore follow that the posterior converges to a distribution centered on the MLE and with optimal asymptotic variance given by the variance of the efficient influence function. While we leave a formal proof of this convergence to future work, we conduct a simulation study in a later section to investigate the properties of Π^*_β in a finite-sample context.

6 Computation

Putting the proposed estimators into practice requires a software implementation. The frequentist and Bayesian estimators have been implemented in the software package `TargetedMSM.jl` in the Julia programming language (Bezanson et al., 2017).

6.1 Universal Algorithm

The efficient influence function $D^*(P)$ depends on four components: the function $\Delta^*(P)$ and the derivatives $\dot{L}$, $\ddot{L}$, and $\nabla\dot{L}$. Each of the derivatives depends on the form of the working model and loss function chosen by the analyst. Writing these derivatives in software by hand is tedious, and makes it more difficult for users to implement new working models and loss functions.

In a unified approach, we propose a universal algorithm that encompasses all loss functions and working models that are well-chosen by the analyst. The algorithm uses automatic differentiation to compute the required derivatives, alleviating the need for the analyst to supply them (Baydin et al., 2018; Margossian, 2019). When using our universal algorithm, the analyst only needs to provide software implementations of the MSM working model and loss function. Our Julia package is available at `https://github.com/herbps10/TargetedMSM.jl`. Its implementation uses forward mode automatic differentiation through the Julia package `ForwardMode.jl` (Revels et al., 2016).

6.2 Markov-Chain Monte Carlo

In practice, the targeted posterior distribution will not have a closed form. However, 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

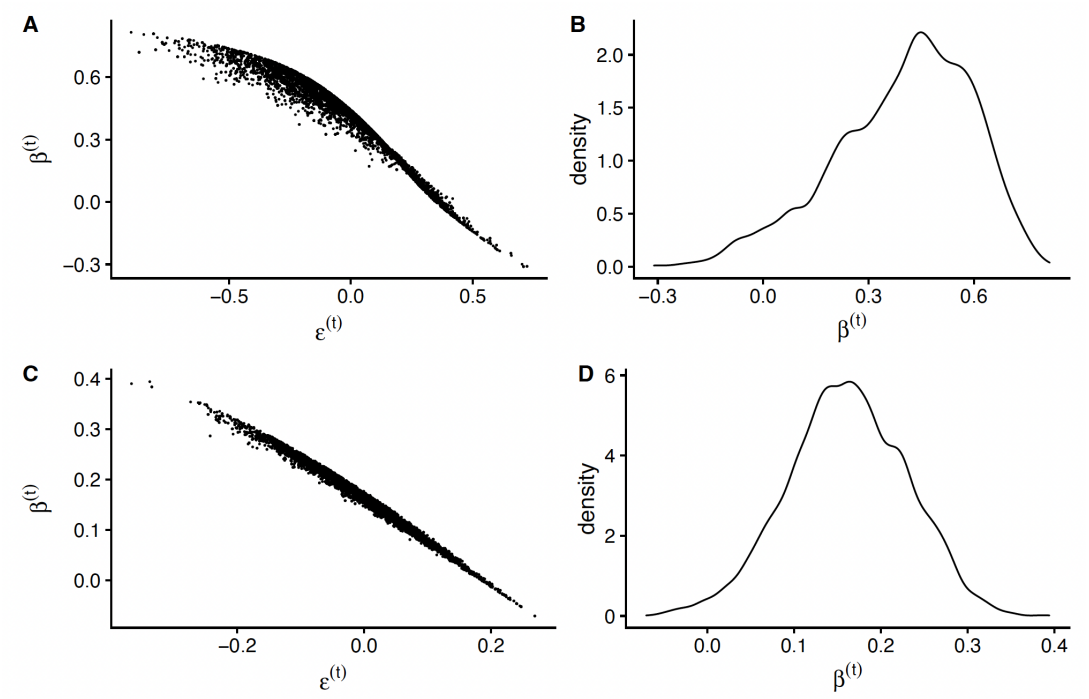

Figure 1: In finite samples, the targeted posterior can fail to accurately characterize the uncertainty in estimating $\boldsymbol{\beta}$ because the map $\boldsymbol{\epsilon} \mapsto B(F_{\boldsymbol{\epsilon}})$ cannot reach the tails of the posterior distribution. In (A), a diagnostic plot shows joint posterior samples $\boldsymbol{\epsilon}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$, in which it can be seen that $\boldsymbol{\beta}$ is bounded above. The marginal posterior distribution of $\boldsymbol{\beta}$, shown in (B), is therefore skewed. Plots (C) and (D) show an example where this pathological behavior is not observed.

the posterior distribution of $\boldsymbol{\epsilon}$, which can then be used to generate a set of samples from the posterior distribution of $\boldsymbol{\beta}_0$ through the mapping B. In this work, we implement a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; Robert, 2015) with a Gaussian proposal distribution to generate samples of $\boldsymbol{\epsilon}$ which are then mapped to samples of $\boldsymbol{\beta}$ from the posterior distribution. A detailed description of the algorithm is given in Appendix 11.7.

6.3 Diagnostics

In finite samples, it is possible that $\{F_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in \mathbb{R}^p\}$ will exhibit pathological behavior in which $B(F_{\boldsymbol{\epsilon}})$ is bounded in such a way that there is no value of $\boldsymbol{\epsilon}$ such that $B(F_{\boldsymbol{\epsilon}})$ reaches the tails of the posterior distribution. However, a simple visual diagnostic can be used to identify this behavior in practice. First, we assume that we have at hand a sample of ℓ draws $\boldsymbol{\epsilon}^{(1)}, \dots \boldsymbol{\epsilon}^{(\ell)}$ and $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(\ell)}$ from the posterior distributions of $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$, perhaps generated by the Metropolis-Hastings algorithm discussed in Section 6.2. The diagnostic simply consists in plotting each sample with $\boldsymbol{\beta}^{(t)}$ on the y-axis and $\boldsymbol{\epsilon}^{(t)}$ on the x-axis. We expect to see one-to-one relationship between $\boldsymbol{\epsilon}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$. In a pathological setting, the $\boldsymbol{\beta}^{(t)}$ reach a threshold. If this is observed, then the draws $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(\ell)}$ should not be trusted as an accurate characterization of the uncertainty of estimating $\boldsymbol{\beta}_0$, as they collectively fail to capture the tails of the posterior distribution. An example of the proposed diagnostic is shown in Figure 1.

7 Simulation Study

We present a simulation study that compares the finite sample performance of the frequentist and Bayesian targeted estimators for estimating Conditional Treatment Effects, as in the running example. All the code necessary to reproduce the simulation study is available at `https://github.com/herbps10/targeted_msms_paper`. The data generating distribution for the simulations is defined by the following

conditional distributions:

$$\begin{aligned}(X_1, X_2, X_3, X_4) &\sim N(\mathbf{0}, \boldsymbol{I}_4),\\ A|X_1, X_2, X_3, X_4 &\sim \text{Bernoulli}(\text{logit}^{-1}(0.5X_1 - 0.5X_2 + 0.2X_3 - 0.1X4)),\\ Y|A, X_1, X_2, X_3, X_4 &\sim \text{Bernoulli}(\text{logit}^{-1}(X_2 + X_3 + 3A + 1.5AX_4)).\end{aligned}$$

A total of 500 datasets were drawn from the data generating distribution for each sample size $n \in \{50, 100, 250, 500, 750, 1000\}$. We define a marginal structural model with linear working model (3) and squared-error loss function (4) which yield the target parameter

$$B(P) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\, \mathbb{E}_P\left[\left(\Psi_P^\sharp(X) - (1, X_4)\boldsymbol{\beta}\right)^2\right],$$

where $\Psi_P^\sharp$ is defined as in the example and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$.

In order to investigate the behavior of the estimators under inconsistent estimation of the nuisance parameters, the parameters g_P and $\bar{Q}_P$ were estimated in each sample using generalized linear models in four configurations: g_P and $\bar{Q}_P$ estimated using parametric models with

(a) both models for g_P and $\bar{Q}_P$ specified correctly,

(b) the model for g_P specified correctly, that for $\bar{Q}_P$ misspecified,

(c) the model for g_P misspecified, that for $\bar{Q}_P$ specified correctly,

(d) both models for g_P and $\bar{Q}_P$ misspecified.

The correctly specified regressions included all covariates, while the misspecified regressions included only X_1 and X_4 as covariates. For the Bayesian estimator, the following priors were used for β_1 and β_2:

$$(\beta_1, \beta_2) \sim N(\mathbf{0}, \boldsymbol{I}_2).$$

The estimators of $\boldsymbol{\beta}_0$ are evaluated by the empirical coverage of the 95% credible (or confidence) intervals and in terms of absolute bias of the point estimators. For the Bayesian estimator, the posterior median is taken as a point estimator.

Results Figure 2 shows the empirical coverage of the 95% credible (confidence) intervals, and Figure 3 shows the absolute bias. The results are also summarized as Table 1 in the appendix. Both the estimators achieve optimal empirical coverage of the 95% credible (confidence) interval and achieve low absolute bias when both nuisance parameters are correctly specified. Of note, for the smallest sample size the Bayesian estimator achieves slightly better coverage than the frequentist estimator, suggesting the Bayesian approach may be more conservative in some finite sample settings. When either of the estimators are estimated with misspecified regressions the absolute bias still approaches zero as the sample size increases, demonstrating the double-robust properties of the estimators. Interestingly, the 95% empirical coverage is not greatly affected by the misspecification, a result that is not guaranteed theoretically. The good performance in this case may be caused by the specific setup of the simulation study. Finally, when both nuisance parameters are estimated inconsistently the absolute bias is larger than in the other scenarios and the empirical coverage is highly degraded for β_1. The fact that the empirical coverage of β_2 remains near-optimal is likely a result of the simulation setup.

8 Application

As an example, we define and estimate the causal effect of a broad-based family planning intervention on postpartum contraceptive use using data from a randomized field experiment conducted in Lilongwe, Malawi (Karra et al., 2022). All the code necessary to reproduce the data analysis is available at `https://github.com/herbps10/targeted_msms_paper`.

Postpartum women were randomized into intervention and control groups in which intervention consisted of receiving a family planning information package and counseling sessions, free transportation to a family planning clinic, free family planning services at the clinic, and phone consultations with a doctor and reimbursement for any treatment necessary due to side effects. Contraceptive use was measured as an outcome after a 2-year interval.

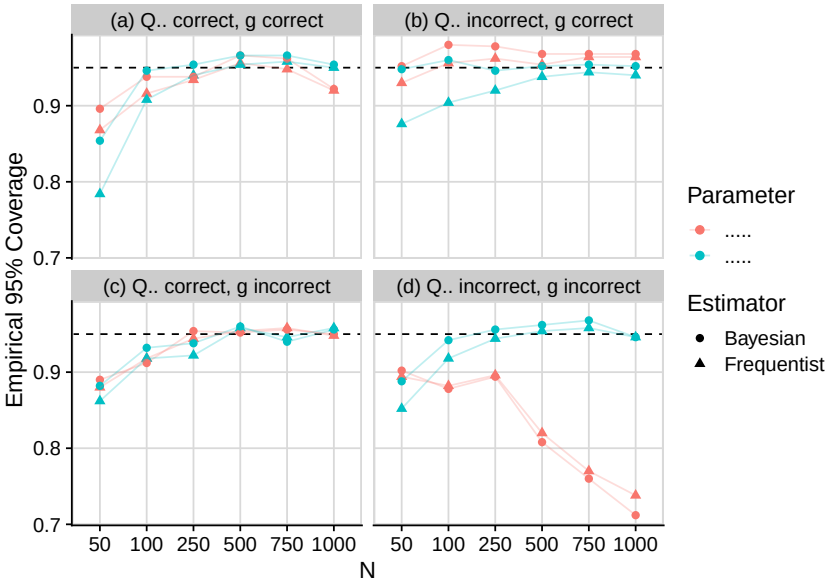

Figure 2: Frequentist empirical coverage of 95% credible and confidence intervals in the simulation study for scenarios (a), (b), (c), and (d).

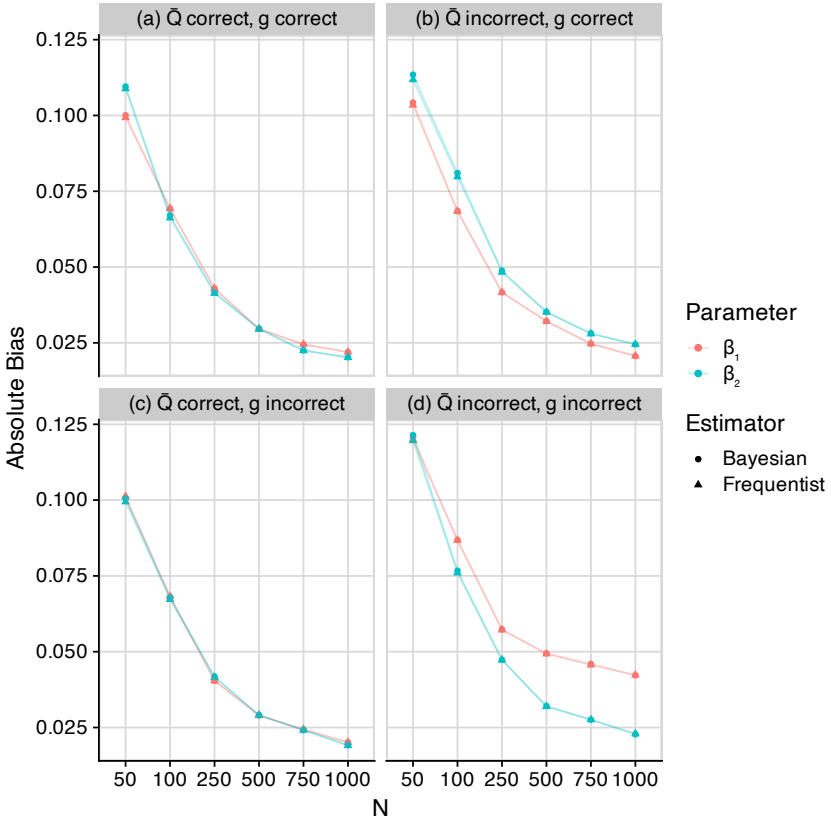

Figure 3: Absolute bias of estimator in the simulation study for scenarios (a), (b), (c), and (d).

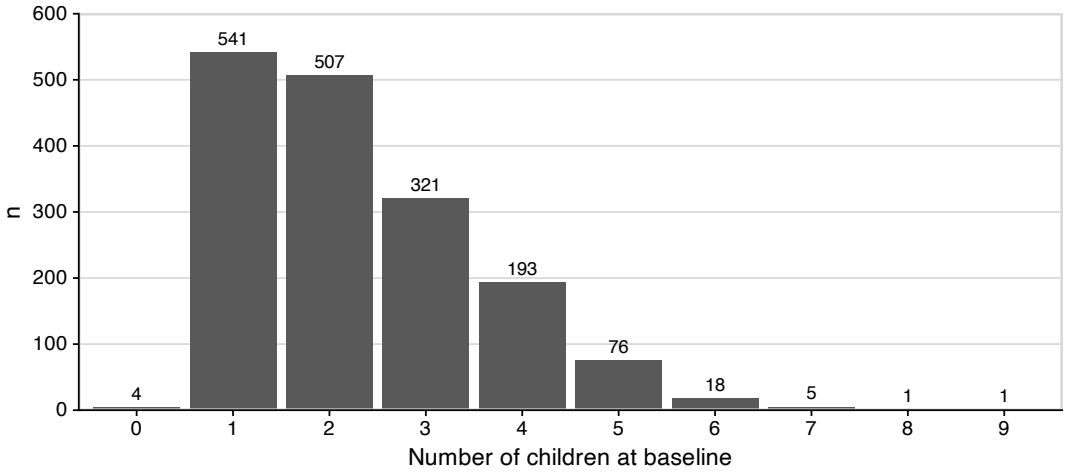

Figure 4: Distribution of the number of children at baseline among the cohort of women in the example.

We use the publicly available replication dataset from the original paper to conduct our analysis (Karra and Canning, 2022). We refer to the paper and published study protocol for detailed information regarding the study design and data processing (Karra et al., 2020, 2022). The variables for each participant used in our analysis are:

- X: number of children who are alive, educational attainment (primary or less vs. secondary or higher), age (three bins), age of sexual debut, ever used family planning, religion, work status, tribal group, neighborhood, and current contraceptive use, all measured at baseline;
- A: binary indicator of inclusion in intervention group;
- Y: binary indicator of contraceptive use at endline.

Following the original analysis, 5 observations were excluded due to missing data, leaving an analysis dataset of 1667 observations. The intervention group included 781 participants (46.9%). At endline, 1240 participants were recorded as using contraceptives (74.3%). The number of children alive at baseline (from now on referred to as "number of children at baseline"), which we consider as a potential treatment effect modifier, ranged from 0 to 9 (mean: 2.3, standard deviation: 1.3). Figure 4 shows the full distribution of children at baseline.

Target Parameter Let $\Psi_P^\sharp$ be the Conditional Average Treatment Effect as defined in the running example, interpretable as the expected difference in contraceptive use at endline conditional on inclusion vs. non-inclusion in the intervention group. We consider the total number of children who are alive at baseline, denoted X_{children}, to be a potential treatment effect modifier, and set $V = X_{\text{children}}$. A linear working model and squared-error loss function are adopted, as in the example. The target parameter is therefore

$$B^\sharp(P) = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\arg\min}\, \mathbb{E}_P \left[\left(\Psi_P^\sharp(X) - \boldsymbol{\beta}^\top (1, X_{\text{children}})^\top \right)^2 \right].$$

The MSM parameters β_1 and β_2 can be interpreted as the intercept and slope of a line that best summarizes (in terms of a squared-error loss) the relationship between the CATE and number of children at baseline.

Causal Identification The statistical parameter $\boldsymbol{\beta}_0$ can be given a causal interpretation if the identification assumptions for $\Psi^\sharp$ hold. Positivity ($g_P(1, X) > 0$ for all X) in this setting can be interpreted as the assumption that the probability of randomization into the intervention group is positive for all strata of covariates. The no unmeasured confounders assumption requires that the potential outcomes $Y(0)$ and $Y(1)$ are independent of intervention status conditional on the covariates. Both assumptions reasonably hold given the randomized study design. It is possible that the consistency assumption, however, may not hold due to spillover effects between the intervention and control groups.

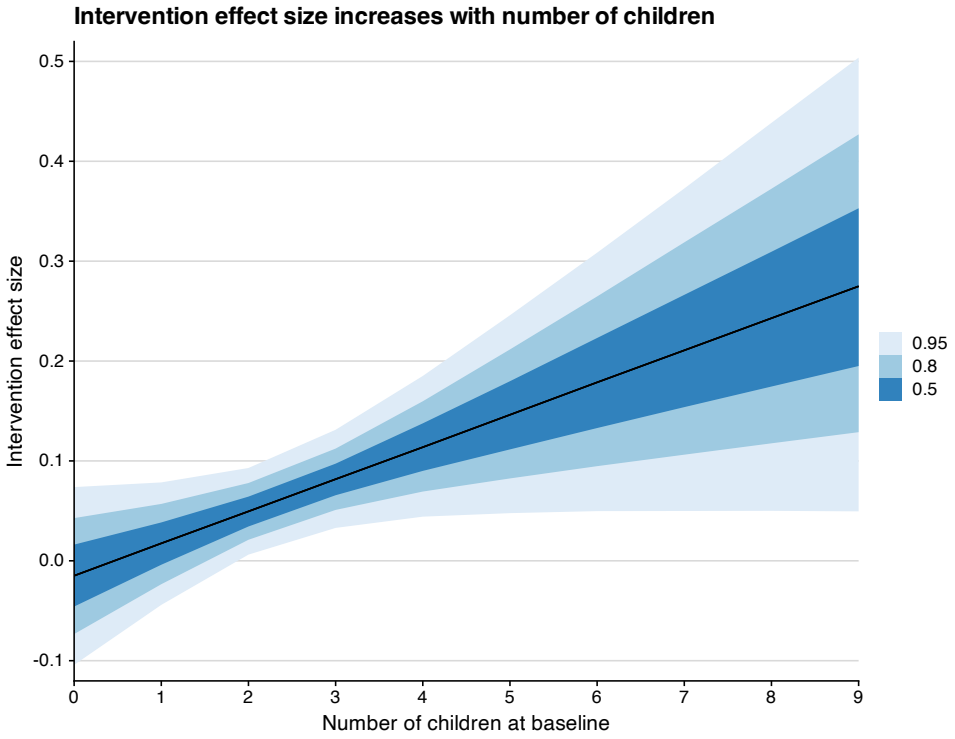

Figure 5: Posterior distribution of the parameter $\boldsymbol{\beta}$ of the working model $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$, which can be interpreted as a linear approximation under a squared error loss of the conditional average treatment effect of the intervention on contraceptive use at endline. The black line is the posterior median and the shaded regions are the 50%, 80%, and 95% credible intervals. The results suggest the treatment effect grows as the number of children increases.

Estimation The nuisance parameters g_P and $\bar{Q}_P$ were estimated using an ensemble of learners using the SuperLearner algorithm (van der Laan et al., 2007; Polley et al., 2021). The base learners included generalized linear models, ℓ_1-penalized regression (Friedman et al., 2010; Simon et al., 2011), random forests (Wright and Ziegler, 2017), gradient boosting trees (Chen et al., 2022), and the Highly Adaptive Lasso (Benkeser and van der Laan, 2016; Coyle et al., 2022; Hejazi et al., 2020). For the Bayesian estimator, the following weakly informative priors were applied, reflecting a lack of strong prior knowledge of the true values of the parameters:

$$(\beta_1, \beta_2) \sim N(\mathbf{0}, \boldsymbol{I}_2).$$

The Metropolis-Hastings algorithm was run for $50,000$ iterations to yield a set of joint draws from the posterior distribution.

Results The frequentist estimates for β_1 and β_2 were -0.016 (95% CI: $[-0.1, 0.07]$) and 0.033 (95% CI: $[0.0019, 0.063]$), respectively. The two-sided hypothesis test of H_0 : "$\beta_2 = 0$" vs. H_1 : "$\beta_2 \neq 0$" yields a p-value of $p = 0.038$. Thus there is sufficient evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis that $\beta_2 = 0$. The Bayesian posterior medians for β_1 and β_2 were -0.015 (95% CI: $[-0.1, 0.074]$) and 0.032 (95% CI: $[-0.001, 0.066]$), respectively. The credible intervals can be interpreted as the regions in which the true values of β_1 and β_2 reside with 95% probability. The posterior probability that $\beta_1 > 0$ is 97.1%. The posterior distribution of the working model $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$ is shown in Figure 5. Both the frequentist and Bayesian methods reach substantially the same conclusion, although care should be taken to interpret the results following the correct interpretation of probability for each approach.

Kernel density plots of the posterior distribution of the MSM parameters are shown in Figure 6. The figure also includes normal densities centered on the targeted point estimate $\boldsymbol{\beta}_n^*$ and with variance given by the variance of $D^*(P_n^*)$, the estimate of the efficient influence function. As expected under the Bernstein von-Mises theorem (Theorem 4), the posterior distribution and normal densities are similar.

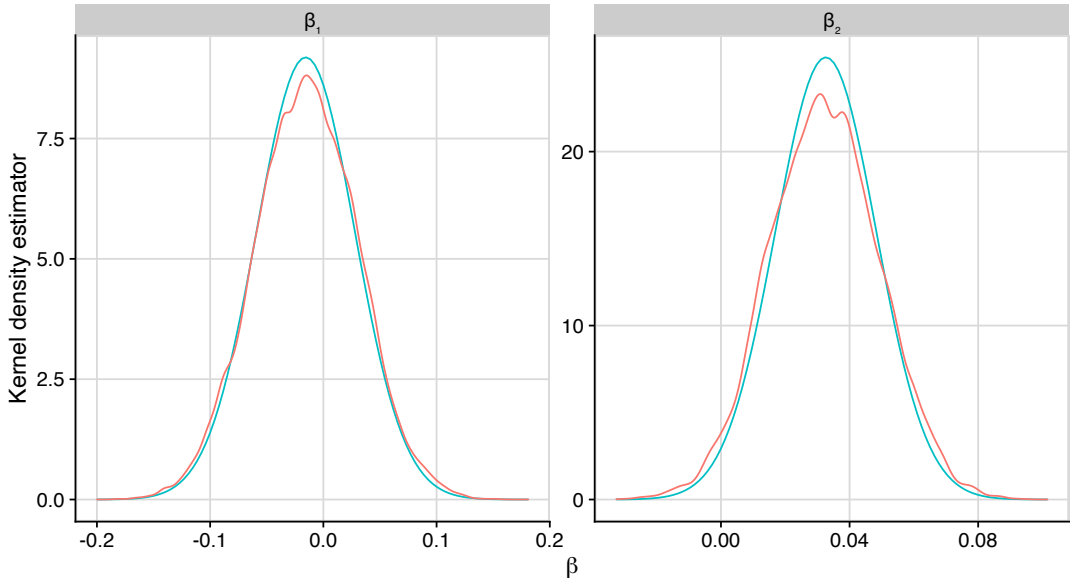

Figure 6: Kernel density plot of posterior distribution (red line) of β_1 and β_2 from the example application. A normal density is also pictured (blue line) centered on the frequentist point estimates $\hat{\boldsymbol{\beta}}_n^*$ with variance given by the targeted estimate of the variance of the efficient influence function $D^*(P_0)$.

9 Discussion

In this work, we proposed a novel targeted Bayesian estimator for the parameter of a Marginal Structural Model. The general notation for MSMs defines a parameter $B(P)$ which is the parameter of a user-specified working model $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$ that best approximates a functional summary Ψ_P with respect to a loss function L. A semi-parametric analysis of $P \mapsto B(P)$ derived its (already known) efficient influence function $D^*(P)$ at any P in the model. The variance $\text{Var}_P(D^*(P)(O))$ is the asymptotic efficiency bound for estimating $B(P)$ within a non-parametric model. We then presented a general framework for efficient estimation using Targeted Maximum Loss-Based Estimation, and introduced a novel targeted Bayesian estimator.

Bayesian TMLE blends frequentist and Bayesian approaches to statistical inference in a way that enjoys benefits of both approaches. From the frequentist TMLE it inherits a non-parametric estimation strategy that avoids unrealistic assumptions on the data generating process. Flexible estimation methods, including algorithms from machine learning, can be used (or combined, via ensemble methods) to estimate the required nuisance parameters, and are then naturally incorporated into the targeted Bayesian analysis. From the Bayesian paradigm, the targeted method inherits the ability to incorporate prior information into the analysis and the subjective interpretation of probability. An additional benefit is that contrary to traditional Bayesian non-parametric inference, which generally requires placing priors on complex probability spaces or abstract hyperparameters, in the targeted approach priors can be placed directly on the parameter of interest. For MSMs, priors are placed directly on $B(P)$.

We stress that the Bayesian TMLE should not be interpreted as "100% Bayesian" because Bayesian inference is not used to model the entire data-generating process (that is, the full joint distribution of X and Z is not modeled). Indeed, estimates of the required nuisance parameters are estimated separately and not necessarily within a Bayesian paradigm. In other words, the method is not "generative", as is typically the case for Bayesian models (Gelman et al., 2020). Bayesian TMLE is thus better understood as a hybrid approach that combines aspects of frequentist and Bayesian paradigms.

Our work leaves open multiple directions for future research. More investigation is warranted to understand the finite sample performance of both the frequentist and Bayesian estimators for more complex data generating distributions and working models of high dimension. In addition, further theoretical work is necessary to establish that the targeted posterior converges to a distribution centered on the frequentist MLE with optimal variance given by the variance of the EIF. Part of the challenge with this result is to show it holds under reasonable assumptions on the convergence of the nuisance parameters, such as that $\|\bar{Q}_n^* - \bar{Q}_0\|_2 = o_p(n^{-1/4})$.

Bayesian inference allows for the introduction of prior information in a principled manner. While in

the analyses presented here we used independent and weakly informative priors, an interesting direction for future work would be to investigate more complex priors. Hierarchical priors are commonly used in applied Bayesian analyses to share information between units, such that parameters are partially pooled towards a common mean (Gelman et al., 2011). This has the benefit of stabilizing estimates, particularly in small samples. Hierarchical priors could be applied to MSM parameters that exhibit hierarchical structure, which we hypothesize may lead to improved finite sample performance. As long as such hierarchical priors satisfy the (weak) assumptions necessary for the Bernstein von-Mises theorem, most importantly that the true parameter is included in the support of the prior, then such a prior choice will not affect the desirable asymptotic properties of the targeted posterior distribution.

10 Acknowledgements

We would like to thank Ismaël Castillo for helpful discussions, and Leontine Alkema and Laura Balzer for useful feedback. This article was written while the first author was a visiting researcher at MAP5 (UMR CNRS 8145, Université Paris Cité). This material is based upon research supported by the Chateaubriand Fellowship of the Office for Science & Technology of the Embassy of France in the United States.

References

M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

D. Benkeser and M. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696. IEEE, 2016.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi: 10.1137/141000671. URL `https://epubs.siam.org/doi/10.1137/141000671`.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6(none):1059 – 1099, 2012. doi: 10.1214/12-EJS703. URL `https://doi.org/10.1214/12-EJS703`.

T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan. *xgboost: Extreme Gradient Boosting*, 2022. URL `https://CRAN.R-project.org/package=xgboost`. R package version 1.6.0.1.

J. R. Coyle, N. S. Hejazi, R. V. Phillips, L. W. van der Laan, and M. J. van der Laan. *hal9001: The scalable highly adaptive lasso*, 2022. URL `https://github.com/tlverse/hal9001`. R package version 0.4.3.

I. Diaz, A. E. Hubbard, and M. J. van der Laan. Targeted Bayesian learning. In *Targeted Learning*, pages 475–493. Springer, 2011.

I. Díaz, O. Savenkov, and H. Kamel. Nonparametric targeted Bayesian estimation of class proportions in unlabeled data. *Biostatistics*, 23(1):274–293, 06 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa022. URL `https://doi.org/10.1093/biostatistics/kxaa022`.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/v33/i01/`.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis (3rd ed.)*. Chapman and Hall/CRC, 2011. URL `https://doi.org/10.1201/b16018`.

A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334940.

N. S. Hejazi, J. R. Coyle, and M. J. van der Laan. hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*, 2020. doi: 10.21105/joss.02526. URL https://doi.org/10.21105/joss.02526.

A. Hájek. Interpretations of Probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

M. Karra and D. Canning. Replication Data for: The causal effect of a family planning intervention on women's contraceptive use and birth spacing, 2022. URL https://doi.org/10.7910/DVN/KNNYZN.

M. Karra, D. Canning, et al. The effect of improved access to family planning on postpartum women: protocol for a randomized controlled trial. *JMIR Research Protocols*, 9(8):e16697, 2020.

M. Karra, D. Maggio, M. Guo, B. Ngwira, and D. Canning. The causal effect of a family planning intervention on women’s contraceptive use and birth spacing. *Proceedings of the National Academy of Sciences*, 119(22):e2200279119, 2022. doi: 10.1073/pnas.2200279119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2200279119.

C. C. Margossian. A review of automatic differentiation and its efficient implementation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1305, 2019.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

R. Neugebauer and M. van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M. van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185, 2014. doi: doi:10.1515/jci-2013-0007. URL https://doi.org/10.1515/jci-2013-0007.

E. Polley, E. LeDell, C. Kennedy, and M. van der Laan. *SuperLearner: Super Learner Prediction*, 2021. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-28.

J. Revels, M. Lubin, and T. Papamarkou. Forward-mode automatic differentiation in Julia. *arXiv:1607.07892 [cs.MS]*, 2016. URL https://arxiv.org/abs/1607.07892.

C. P. Robert. The Metropolis-Hastings algorithm, 2015. URL https://arxiv.org/abs/1504.01896.

J. M. Robins. Marginal structural models. *1997 proceedings of the American Statistical Association, section on Bayesian statistical science*, pages 1–10, 1998.

M. Rosenblum and M. J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2), 2010.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. doi: 10.18637/jss.v039.i05. URL https://www.jstatsoft.org/v39/i05/.

C. Stein et al. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 187–195, 1956.

B. Toth and M. J. van der Laan. TMLE for marginal structural models based on an instrument. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 350, 2016. URL https://biostats.bepress.com/ucbbiostat/paper350.

M. J. van der Laan. The construction and analysis of adaptive group sequential designs. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 232, 2008. URL https://biostats.bepress.com/ucbbiostat/paper232.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer New York Dordrecht Heidelberg London, 2011.

M. J. van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer International Publishing, 2018.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. doi: doi:10.2202/1544-6115.1309. URL `https://doi.org/10.2202/1544-6115.1309`.

A. van der Vaart. Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer, Berlin, 2002.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics, 1996.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

W. Zheng and M. J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 273, 2010. URL `https://biostats.bepress.com/ucbbiostat/paper273`.

W. Zheng, M. Petersen, and M. J. van der Laan. Doubly robust and efficient estimation of marginal structural models for the hazard function. *The International Journal of Biostatistics*, 12(1):233–252, 2016.

W. Zheng, Z. Luo, and M. J. van der Laan. Marginal structural models with counterfactual effect modifiers. *The International Journal of Biostatistics*, 14(1), 2018.

11 Appendix

11.1 Full Statement and Proof of Theorem 3

Here is the full version of Result 1. Recall that, for any $P \in \mathcal{M}$ and $s \in L_0^2(P)$ such that $s \neq 0$, $\|s\|_\infty < \infty$, $\left\{P_\epsilon : |\epsilon| < \|s\|_\infty^{-1}\right\} \subset \mathcal{M}$ fluctuates P in the direction of s. For clarity, we now introduce the mapping $B_P^\circ : \epsilon \mapsto B(P_\epsilon)$.

Theorem 3 (Efficient Influence Function of B)**.** *Let us assume that:*

1. *For every $P \in \mathcal{M}$, for every $t \in \mathcal{T}$, the following two conditions are met P-almost surely:*

   (a) $\boldsymbol{\beta}' \mapsto L_m(t, \boldsymbol{\beta}')(X)$ is differentiable at every $\boldsymbol{\beta} \in \mathcal{B}$ with derivative $\dot{L}_m(t, \boldsymbol{\beta})(X) \in \mathbb{R}^p$,

   (b) $\boldsymbol{\beta}' \mapsto \dot{L}_m(t, \boldsymbol{\beta}')(X)$ is differentiable at every $\boldsymbol{\beta} \in \mathcal{B}$ with derivative $\ddot{L}_m(t, \boldsymbol{\beta})(X) \in \mathbb{R}^{p \times p}$.

   Moreover, for every $P \in \mathcal{M}$, for every $\boldsymbol{\beta} \in \mathcal{B}$, it holds P-almost surely that

   (c) $t' \mapsto \dot{L}_m(t', \boldsymbol{\beta})(X)$ is differentiable at every $t \in \mathcal{T}$ with derivative $\nabla \dot{L}_m(t, \boldsymbol{\beta})(X) \in \mathbb{R}^p$.

2. *The function $(\epsilon, \boldsymbol{\beta}) \mapsto U(\epsilon, \boldsymbol{\beta}) := \mathbb{E}_{P_\epsilon}\left[\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)\right]$ is differentiable in the neighborhood of $(0, B_P^\circ(0))$.*

3. *For every $P \in \mathcal{M}$, $\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B(P))(X)\right]$ is invertible.*

4. *For every $P \in \mathcal{M}$, for all $s \in L_0^2(P)$ such that $s \neq 0$, $\|s\|_\infty < \infty$, it holds that $U(\epsilon, B_P^\circ(\epsilon)) = 0$ in an ϵ-neighborhood of 0.*

5. *(Dominated Convergence) For all $P \in \mathcal{M}$, there exists a random variable $G_1 \in \mathbb{R}_+^{p \times p}$ with $\mathbb{E}_P[G_1] < \infty$, such that $|\ddot{L}_m(\Psi_P(X), \boldsymbol{\beta})(X)| \leq G_1$ P-almost surely in a $\boldsymbol{\beta}$-neighborhood of $B_P^\circ(\epsilon)$. Here and below the absolute value and inequalities are to be understood entrywise.*

6. *(Dominated Convergence) For all $P \in \mathcal{M}$, there exists a random variable $G_2 \in \mathbb{R}_+^p$ with $\mathbb{E}_P[G_2] < \infty$ such that $|\frac{d}{d\epsilon}\dot{L}_m(\Psi_{P_\epsilon}(X), B_P^\circ(\epsilon))(X)| \leq G_2$ P-almost surely in an ϵ-neighborhood of 0.*

7. *For all $P \in \mathcal{M}$, for all $s \in L_0^2(P)$ such that $s \neq 0$, $\|s\|_\infty < \infty$, if $\left\{P_\epsilon : |\epsilon| < \|s\|_\infty^{-1}\right\} \subset \mathcal{M}$ is characterized by $\frac{dP_\epsilon}{dP} = 1 + \epsilon s$ for all $\epsilon \in \mathbb{R}$ such that $|\epsilon| < \|s\|_\infty^{-1}$, then the mapping $\epsilon \mapsto \Psi_{P_\epsilon}(X)$ is almost surely differentiable at 0 and*

$$\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\bigg|_{\epsilon=0} = \mathbb{E}_P\left[\Delta^*(P)(O)s(O) \mid X\right]$$

for a function $\Delta^*(P) \in L_0^2(P)$.

Then the target functional $P \mapsto B(P)$ is pathwise differentiable at every $P \in \mathcal{M}$, with an efficient influence function $D^*(P)$ at P given by

$$D^*(P)(O) = M^{-1}\left[D_1^*(P)(O) + D_2^*(P)(X)\right],$$

where $D_1^*(P), D_2^*(P) \in L_0^2(P)$ are given by

$$\begin{aligned} D_1^*(P)(O) &= \nabla \dot{L}_m(\Psi_P(X), B(P))(X) \times \Delta^*(P)(O), \\ D_2^*(P)(X) &= \dot{L}_m(\Psi_P(X), B(P))(X), \end{aligned}$$

and the normalizing matrix M is given by

$$M = -\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B(P))(X)\right].$$

Proof. Set arbitrarily $P \in \mathcal{M}$ and $s \in L_0^2(P)$ such that $s \neq 0$ and $\|s\|_\infty < \infty$. Consider the submodel $\left\{P_\epsilon : |\epsilon| < \|s\|_\infty^{-1}\right\}$ given by $\frac{dP_\epsilon}{dP} = 1 + \epsilon s$. By assumption 4, $U(\epsilon, B_P^\circ(\epsilon)) = 0$ in an ϵ-neighborhood of 0. By assumption 2, U is differentiable in a neighborhood of $(0, B_P^\circ(0))$. Denote by $\frac{dU}{d\epsilon}$ and $\frac{\partial U}{\partial \boldsymbol{\beta}}$ the derivative and gradient of U with respect to its ϵ and $\boldsymbol{\beta}$ arguments, respectively. Differentiating $\epsilon \mapsto U(\epsilon, B_P^\circ(\epsilon))$ at $\epsilon = 0$ yields

$$\frac{dU}{d\epsilon}(0, B_P^\circ(0)) + \frac{\partial U}{\partial \boldsymbol{\beta}}(0, B_P^\circ(0))\frac{dB_P^\circ}{d\epsilon}(0) = 0.$$

Solving for $\frac{dB_P^\circ}{d\epsilon}(0)$, and using assumption 3, yields

$$\frac{dB_P^\circ}{d\epsilon}(0) = -\left\{\frac{\partial U}{\partial \boldsymbol{\beta}}(0, B_P^\circ(0))\right\}^{-1}\frac{dU}{d\epsilon}(0, B_P^\circ(0)). \tag{8}$$

Now we compute each of these derivatives. The first derivative becomes (the second equality is justified below)

$$\begin{aligned} -\frac{\partial U}{\partial \boldsymbol{\beta}}(0, B_P^\circ(0)) &= -\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{E}_{P_\epsilon}\left[\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)\right]\Big|_{\epsilon=0, \boldsymbol{\beta}=B_P^\circ(0)} \\ &= -\mathbb{E}_{P_\epsilon}\left[\frac{\partial}{\partial \boldsymbol{\beta}}\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)\Big|_{\epsilon=0, \boldsymbol{\beta}=B_P^\circ(0)}\right] \\ &= -\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B_P^\circ(0))(X)\right] \\ &=: M, \end{aligned}$$

where in the second line the gradient and expectation can be exchanged under the Dominated Convergence theorem (DCT) and assumptions 1 and 5. Note that M is invertible under assumption 3. The second derivative splits into two parts, first by definition of U, second by definition of P_ϵ:

$$\begin{aligned} \frac{dU}{d\epsilon}(0, B_P^\circ(0)) =& \frac{d}{d\epsilon}\mathbb{E}_{P_\epsilon}\left[\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)\right]\Big|_{\epsilon=0, \boldsymbol{\beta}=B_P^\circ(0)} \\ =& \frac{d}{d\epsilon}\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)\right]\Big|_{\epsilon=0, \boldsymbol{\beta}=B_P^\circ(0)} \\ &+ \frac{d}{d\epsilon}\epsilon\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X), \boldsymbol{\beta})(X)s(O)\right]\Big|_{\epsilon=0, \boldsymbol{\beta}=B_P^\circ(0)}. \end{aligned} \tag{9}$$

The first term becomes (the first equality is justified below)

$$\begin{aligned}\frac{d}{d\epsilon}\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)\right]\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)} &= \mathbb{E}_P\left[\frac{d}{d\epsilon}\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)}\right]\\ &= \mathbb{E}_P\left\{\nabla\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)}\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\Big|_{\epsilon=0}\right\}\\ &= \mathbb{E}_P\left\{\nabla\dot{L}_m(\Psi_P(X),B_P^\circ(0))(X)\times\Delta^*(P)(O)s(O)\right\}, \quad (10)\end{aligned}$$

where the interchange of the derivative and expectation in the first line is by the DCT and assumptions 1 and 6. For the second term, apply the product rule to yield

$$\begin{aligned}\frac{d}{d\epsilon}\epsilon\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)s(O)\right]\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)} &= \left[\epsilon\frac{d}{d\epsilon}\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)s(O)\right]\right]\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)}\\ &\quad+\mathbb{E}_P\left[\dot{L}_m(\Psi_{P_\epsilon}(X),\boldsymbol{\beta})(X)\Big|_{\epsilon=0,\boldsymbol{\beta}=B_P^\circ(0)}s(O)\right]\\ &=\mathbb{E}_P\left[\dot{L}_m(\Psi_P(X),B_P^\circ(0))(X)s(O)\right]. \quad (11)\end{aligned}$$

Collecting (8), (9), (10), and (11) yields

$$\begin{aligned}\frac{dB_P^\circ}{d\epsilon}(0) =& M^{-1}\Big[\mathbb{E}_P\left[\nabla\dot{L}_m(\Psi_P(X),B_P^\circ(0))(X)\times\Delta^*(P)(O)s(O)\right]\\ &+\mathbb{E}_P\left[\dot{L}_m(\Psi_P(X),B_P^\circ(0))(X)s(O)\right]\Big],\end{aligned}$$

hence

$$D^*(P)(O) = M^{-1}\left[\nabla\dot{L}_m(\Psi_P(X),B(P))(X)\times\Delta^*(P)(O)+\dot{L}_m(\Psi_P(X),B(P))(X)\right].$$

This completes the proof. □

11.2 Proof of Lemma 1

Proof. The derivative of Ψ_{P_ϵ} evaluated at $\epsilon=0$ decomposes into two parts:

$$\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\bar{Q}_{P_\epsilon}^{(1)}(X)-\bar{Q}_{P_\epsilon}^{(0)}(X)\Big|_{\epsilon=0}.$$

For any $a\in\{0,1\}$, write $\bar{Q}_{P_\epsilon}^{(a)}(X)$ and apply the definition of P_ϵ to arrive at

$$\begin{aligned}\bar{Q}_{P_\epsilon}^{(a)}(X) &= \mathbb{E}_{P_\epsilon}\left[Y|A=a,X\right]\\ &= \frac{\mathbb{E}_P\left[Y|A=a,X\right]+\epsilon\mathbb{E}_P\left[Ys(O)|A=a,X\right]}{1+\epsilon\mathbb{E}_P[s(O)|A=a,X]} \quad \text{(Chambaz et al., 2012, Lemma 1)}\\ &= \frac{\mathbb{E}_P[Y|A=a,X]+\epsilon\mathbb{E}_P[s(O)(Y-\mathbb{E}_P[Y|A=a,X])|A=a,X]}{1+\epsilon\mathbb{E}_P[s(O)|A=a,X]}\\ &\quad+\frac{\epsilon\mathbb{E}_P[Y|A=a,X]\mathbb{E}_P[s(O)|A=a,X]}{1+\epsilon\mathbb{E}_P[s(O)|A=a,X]}\\ &= \frac{\mathbb{E}_P[Y|A=a,X](1+\epsilon\mathbb{E}_P[s(O)|A=a,X])+\epsilon\mathbb{E}_P[s(O)(Y-\mathbb{E}_P[Y|A=a,X]]|A=a,X]}{1+\epsilon\mathbb{E}_P[s(O)|A=a,X]}\\ &= \mathbb{E}_P[Y|A=a,X]+\frac{\epsilon\mathbb{E}_P[s(O)(Y-\mathbb{E}_P[Y|A=a,X]]|A=a,X]}{1+\epsilon\mathbb{E}_P[s(O)|A=a,X]}.\end{aligned}$$

Taking the derivative of $\bar{Q}_{P_\epsilon}^{(a)}$ at $\epsilon = 0$ yields

$$\begin{aligned}\frac{d}{d\epsilon}\bar{Q}_{P_\epsilon}^{(a)}(X)\bigg|_{\epsilon=0} &= \mathbb{E}_P\left[s(O)(Y - \mathbb{E}_P[Y|A=a,X])|A=a,X\right] \\ &= \mathbb{E}_P\left[s(O)\frac{\mathbb{I}(A=a)}{P[A=a \mid X]}\left(Y - \mathbb{E}_P[Y|A=a,X]\right)\bigg|X\right] \\ &= \mathbb{E}_P\left[s(O)\frac{\mathbb{I}(A=a)}{P[A=a \mid X]}(Y - \bar{Q}_P^{(A)}(X))\bigg|X\right].\end{aligned}$$

Therefore

$$\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\bigg|_{\epsilon=0} = \mathbb{E}_P\left[s(O)\left\{\frac{\mathbb{I}(A=1)}{P\left[A=1|X\right]} - \frac{\mathbb{I}(A=0)}{P\left[A=0|X\right]}\right\}(Y - \bar{Q}_P^{(A)}(X))\bigg|X\right],$$

which completes the proof. □

11.3 Proof of Theorem 1

Proof. The result follows directly from Theorem 3. We start by checking each of the assumptions required therein. For every $P \in \mathcal{M}$, $t \in \mathcal{T}$, the derivatives $\dot{L}_m(t,\boldsymbol{\beta})(X) = -2(t - \boldsymbol{\beta}^\top(1,V)^\top)(1,V)^\top$ and $\ddot{L}_m(t,\boldsymbol{\beta})(X) = 2(1,V)^\top(1,V)$ exist at every $\boldsymbol{\beta} \in \mathcal{B}$ P-almost surely. For every $P \in \mathcal{M}$, $\boldsymbol{\beta} \in \mathcal{B}$, the derivative $\nabla\dot{L}_m(t,\boldsymbol{\beta})(X) = -2(1,V)^\top$ exists at every $t \in \mathcal{T}$ P-almost surely. Therefore Assumption (1) is satisfied. Assumptions (2)-(6) are satisfied by assumption. Lemma (1) shows that Assumption (7) holds, and gives the form of $\Delta^*(P)$. Plugging in the calculated derivatives $\ddot{L}_m, \ddot{L}_m, \nabla\dot{L}_m$ into the form of $D^*(P)$ of Theorem 3 and simplifying gives the stated result. □

11.4 Proof of Theorem 2

Proof. To begin, write

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0 = &- P_n[D^*(P_n^*)] && \text{(bias term)} \\ &+ (P_n - P_0)[D^*(P_0)] && \text{(CLT term)} \\ &+ (P_n - P_0)[D^*(P_n^*) - D^*(P_0)] && \text{(empirical process term)} \\ &+ \hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0 + P_0[D^*(P_n^*)] && \text{(second-order remainder).}\end{aligned}$$

We take each of these terms in turn:

bias term: $P_n[D^*(P_n^*)] = o_P(n^{-1/2})$ by construction of the TMLE fluctuation;

CLT term: $\sqrt{n}(P_n - P_0)[D^*(P_0)] \rightsquigarrow N(0, P_0[D^*(P_0)D^*(P_0)^\top)]$ by the Central Limit theorem;

empirical process term: $(P_n - P_0)[D^*(P_n^*) - D^*(P_0)] = o_P(n^{-1/2})$ by the assumed Donsker conditions (van der Vaart, 1998, Lemma 19.24) and assumed convergence of the nuisance parameters;

second-order remainder: $\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0 + P_0[D^*(P_n^*)] = o_P(n^{-1/2})$ by assumption.

This completes the proof. □

11.5 Supplementary Example

Here we present an adaptation of the running example introduced in Section 2.1 where the outcome Y is continuous. The form of the EIF for B is the same as in the main example. The form of the fluctuation models for TMLE, however, requires more work.

Let $Z = (A, Y)$, where A is a binary treatment indicator and $Y \in \mathbb{R}$ is a *continuous* outcome. Set arbitrarily $P \in \mathcal{M}$. Let $g_P(a,x) := P\left(A = a|X = x\right)$ for any $(a,x) \in \{0,1\} \times \mathcal{X}$, and suppose that $g_P(1,X) > 0$ holds P-almost-surely. For both $a \in \{0,1\}$ let $\bar{Q}_P^{(a)}(X) = \mathbb{E}_P\left[Y|A=a,X\right]$, which is defined almost surely on $\mathcal{X}$. Define $\Psi_P^\flat(X) = \bar{Q}_P^{(1)}(X) - \bar{Q}_P^{(0)}(X)$. Following the main example, let $B^\flat$ be given by

$$B^\flat(P) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}_P\left[\left(\Psi_P(X) - \boldsymbol{\beta}^\top(1,V)^\top\right)^2\right].$$

11.5.1 Frequentist TMLE

For any $P \in \mathcal{M}$ with corresponding $\Psi_P^\flat$, $\bar{Q}_P$, Q_P and η_P, characterize the fluctuation model with transformation $x \mapsto \phi(x) = x$ as follows: for all $\boldsymbol{\epsilon} \in \mathbb{R}^p$,

$$\begin{aligned}
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(0)}(O) &= \bar{Q}_P^{(0)}(X) + H_0(O)\boldsymbol{\epsilon}^\top (1,V)^\top, \\
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(1)}(O) &= \bar{Q}_P^{(1)}(X) + H_1(O)\boldsymbol{\epsilon}^\top (1,V)^\top, \\
Q_{P,\boldsymbol{\epsilon}}(X) &= C(\boldsymbol{\epsilon}) \exp\left(\boldsymbol{\epsilon}^\top (\Psi_P^\flat(X) - B^\flat(P)^\top (1,V)^\top)(1,V)^\top\right) Q_X(X).
\end{aligned}$$

The clever covariates are given by

$$\begin{aligned}
H_0(O) &= -\frac{I(A=0)}{g_P(0,X)}, \\
H_1(O) &= \quad \frac{I(A=1)}{g_P(1,X)},
\end{aligned}$$

and the loss functions by

$$\begin{aligned}
\mathcal{L}_0(t,O) &= \frac{1}{2} I(A=0)(Y-t)^2, \\
\mathcal{L}_1(t,O) &= \frac{1}{2} I(A=1)(Y-t)^2.
\end{aligned}$$

To ensure this is a valid setup, we need to check conditions (L1), (L2), and (M1). Condition (L1) is satisfied for this choice of loss function (van der Laan and Rose, 2011). The derivatives of the loss functions are

$$\begin{aligned}
\dot{\mathcal{L}}_0(t,O) &= -I(A=0)(Y-t), \\
\dot{\mathcal{L}}_1(t,O) &= -I(A=1)(Y-t).
\end{aligned}$$

Therefore, (L2) is satisfied. To check (M1), see that $(\phi^{-1})'(x) = 1$, so

$$\begin{aligned}
&\dot{\mathcal{L}}_0(\bar{Q}_P^{(0)}(X),O)(\phi^{-1})'(\phi(\bar{Q}_P^{(0)}(X)))H_0(O) + \dot{\mathcal{L}}_1(\bar{Q}_P^{(1)}(X),O)(\phi^{-1})'(\phi(\bar{Q}_P^{(1)}(X)))H_1(O) \\
&= \dot{\mathcal{L}}_0(\bar{Q}_P^{(0)}(X),O)H_0(O) + \dot{\mathcal{L}}_1(\bar{Q}_P^{(1)}(X),O)H_1(O) \\
&= -\left\{\frac{I(A=1)}{g_P(1,X)} - \frac{I(A=0)}{g_P(0,X)}\right\}(Y - \bar{Q}_P^{(A)}(X)) \\
&= -\Delta_P^*(X).
\end{aligned}$$

11.5.2 Bayesian TMLE

The TMLE loss functions $\mathcal{L}_0$ and $\mathcal{L}_1$ in this example can be interpreted as the conditional log-likelihoods of Y given either $A = 0$ or $A = 1$ and X under Gaussian distributions with fixed unit variance. This suggests constructing each $F_{\boldsymbol{\epsilon}}$ such that Y given A and X follows a Gaussian distribution. Fix arbitrarily $P \in \mathcal{M}$, $\boldsymbol{\epsilon} \in \mathbb{R}^p$, introduce $\sigma_P^2(A,X) = \mathrm{Var}_P[Y|A,X]$, and let $F_{\boldsymbol{\epsilon}}$ be characterized by

$$\begin{aligned}
X &\sim Q_{P,\boldsymbol{\epsilon}}, \\
A \mid X &\sim \text{Bernoulli}(1/2), \\
Y \mid X, A &\sim \mathrm{N}\left(\bar{Q}_{P,\boldsymbol{\epsilon}}^{(A)}(X), \sigma_P^2(A,X)\right)
\end{aligned}$$

where

$$\begin{aligned}
Q_{P,\boldsymbol{\epsilon}}(X) &= C(\boldsymbol{\epsilon}) \exp\left(M^{-1}\boldsymbol{\epsilon}^\top (\Psi_P^\flat(X) - B^\flat(P)^\top (1,V)^\top)(1,V)^\top\right) Q_X(X), \\
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(0)}(X) &= \bar{Q}_P^{(0)}(X) + \sigma_P^2(0,X) H_0(O)\boldsymbol{\epsilon}^\top M^{-1}(1,V)^\top, \\
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(1)}(X) &= \bar{Q}_P^{(1)}(X) + \sigma_P^2(1,X) H_1(O)\boldsymbol{\epsilon}^\top M^{-1}(1,V)^\top.
\end{aligned}$$

Then the conditional log-likelihood of O under $F_{\boldsymbol{\epsilon}}$ is given by

$$\begin{aligned}\log f(O \mid \boldsymbol{\epsilon}) &= -\log(\sigma_P^2(A,X)) - \frac{1}{2\sigma_P^2(A,X)}\left(Y - \bar{Q}_{P,\boldsymbol{\epsilon}}^{(A)}(X)\right)^2 + \log Q_{P,\boldsymbol{\epsilon}}(X) + \text{constant} \\ &= -\log(\sigma_P^2(A,X)) - \frac{1}{\sigma_P^2(A,X)}\left(\mathcal{L}_0(\bar{Q}_{P,\boldsymbol{\epsilon}}^{(0)}(X), O) + \mathcal{L}_1(\bar{Q}_{P,\boldsymbol{\epsilon}}^{(1)}(X), O)\right) \\ &\quad + \log Q_{P,\boldsymbol{\epsilon}}(X) + \text{constant}.\end{aligned}$$

The gradient of the log-likelihood evaluated at $\boldsymbol{\epsilon} = \mathbf{0}$ is then

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\epsilon}} \log f(O|\boldsymbol{\epsilon})\Big|_{\boldsymbol{\epsilon}=\mathbf{0}} &= \left\{\frac{I(A=1)}{g_P(1,X)} - \frac{I(A=0)}{g_P(0,X)}\right\}(Y - \bar{Q}_P^{(A)}(X))M^{-1}(1,V)^\top \\ &\quad + (\Psi_P^\flat(X) - B^\flat(P)^\top(1,V)^\top)M^{-1}(1,V)^\top,\end{aligned}$$

which is equal to $D^*(P)(O)$. The Hessian of the log-likelihood evaluated at $\boldsymbol{\epsilon} = \mathbf{0}$ is given by

$$\begin{aligned}\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log f(O|\boldsymbol{\epsilon})\Big|_{\boldsymbol{\epsilon}=\mathbf{0}} &= \left(\frac{\sigma_P^2(A,X)}{g_P(A,X)^2}\right) M^{-1}(1,V)^\top(1,V)M^{-1} \\ &\quad + (\Psi_P(X)^\flat - (B(P)^\flat)^\top(1,V)^\top)^2 M^{-1}(1,V)^\top(1,V)M^{-1}.\end{aligned}$$

Note that

$$P_0\left[\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log f(O|\boldsymbol{\epsilon})\Big|_{\boldsymbol{\epsilon}=\mathbf{0}}\right] = P_0[\lambda^*(P_0)].$$

Therefore conditions 5 and 6 of Theorem 4 are satisfied for this fluctuation model.

11.6 Statement and Proof of Theorem 4

Definition 1 (Local Asymptotic Normality (van der Vaart, 1998, Definition 7.14))**.** *The sequence of statistical models $\{P_{n,\theta} : \theta \in \Theta\}$ is locally asymptotically normal (LAN) at θ if there exist matrices r_n and I_θ and random vectors $\Delta_{n,\theta}$ such that $\Delta_{n,\theta} \rightsquigarrow N(0, I_\theta)$ and for every converging sequence $h_n \to h$*

$$\log \frac{dP_{n,\theta + r_n^{-1}h_n}}{dP_{n,\theta}} = h^\top \Delta_{n,\theta} - \frac{1}{2}h^\top I_\theta h + o_{P_{n,\theta}}(1).$$

It is sometimes more practical to show that $\{P_{n,\theta} : \theta \in \Theta\}$ is differentiable in quadratic mean, which implies local asympotic normality (van der Vaart, 1998, Chapter 7).

Let

$$\begin{aligned}f_n^0(O_{1:n}|\boldsymbol{\epsilon}) &:= \prod_{i=1}^n f\left(O_i\middle|\boldsymbol{\epsilon}, \bar{Q}_0, Q_0, \eta_0, \sigma_0\right), \\ \Pi_\epsilon^0(\boldsymbol{\epsilon}|O_{1:n}) &:= \Pi_\epsilon\left(\boldsymbol{\epsilon}\middle|O_{1:n}, \bar{Q}_0, Q_0, \eta_0, \sigma_0\right), \\ \Pi_\beta^0(\boldsymbol{\beta}|O_{1:n}) &:= \Pi_\beta\left(\boldsymbol{\beta}\middle|O_{1:n}, \bar{Q}_0, Q_0, \eta_0, \sigma_0\right),\end{aligned}$$

be the likelihood, posterior for $\boldsymbol{\epsilon}$, and posterior for $\boldsymbol{\beta}$ corresponding to the submodel $\mathcal{F}^0 = \{F_{\boldsymbol{\epsilon}}^0 : \boldsymbol{\epsilon} \in \mathbb{R}^p\}$ – that is, the submodel through P_0 satisfying (6). Introduce $\vartheta_0 : \boldsymbol{\epsilon} \mapsto B(F_{\boldsymbol{\epsilon}}^0)$.

Theorem 4 (Oracular Bernstein von-Mises)**.** *Let us assume that:*

1. *$\mathcal{F}^0$ is locally asymptotically normal.*
2. *The mapping ϑ_0 is invertible and its inverse ϑ_0^{-1} is twice differentiable at $\boldsymbol{\beta}_0 = B(P_0)$.*
3. *For every $\delta > 0$, there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that*

$$P_0[\phi_n] \to 0 \text{ and } \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq \delta} F_{\boldsymbol{\epsilon}}^0[(1 - \phi_n)] \to 0.$$

4. *The following holds P_0-almost surely: the map $\boldsymbol{\epsilon} \mapsto \log f_n^0(O|\boldsymbol{\epsilon})$ is twice differentiable at every $\boldsymbol{\epsilon} \in \mathbb{R}^p$ with gradient $\frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O \mid \boldsymbol{\epsilon}) \in \mathbb{R}^p$ and Hessian $\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log f_n^0(O|\boldsymbol{\epsilon}) \in \mathbb{R}^{p\times p}$.*

5. *The gradient satisfies*

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} = D^*(P_0)(O).$$

6. *Define*

$$\Lambda_n^0(O) := \frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log f_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}}.$$

Then

$$P_0[\Lambda_n^0] = P_0[\lambda^*(P_0)],$$

where $P_0[\lambda^*(P_0)]$ is assumed to be non-singular.

7. *The joint prior density π_β is absolutely continuous in a neighborhood of $\boldsymbol{\beta}_0$ and $\pi_\beta(\boldsymbol{\beta}_0) > 0$.*

Then

$$\|\Pi^0_{\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}(\cdot \mid O_{1:n}) - N\left(\Delta_n^0, P_0[\lambda^*(P_0)]\right)\|_1 = o_P(1)$$

where

$$\Delta_n^0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n P_0[\lambda^*(P_0)]^{-1} D^*(P_0)(O_i).$$

Note that condition 3 can be difficult to verify. However, if $\boldsymbol{\epsilon}$ belongs to a compact set E, if $\mathcal{F}^0 = \{F^0_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E\}$ is identifiable (meaning $F^0_{\boldsymbol{\epsilon}} = F^0_{\boldsymbol{\epsilon}'}$ implies $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'$) and if $\boldsymbol{\epsilon} \mapsto F^0_{\boldsymbol{\epsilon}}$ is continuous, then condition 3 is not necessary (van der Vaart, 1998, Chapter 10).

Proof. First, we derive the Fisher information matrix of $\boldsymbol{\beta}$. The log-density of $\boldsymbol{\beta}$ is given by

$$\log f_n^0(O_{1:n} \mid \boldsymbol{\beta}) = \log f_n^0(O_{1:n} \mid \boldsymbol{\epsilon} = \vartheta_0^{-1}(\boldsymbol{\beta})).$$

Next, calculate

$$\begin{aligned}
&\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log f_n^0\left(O \middle| \boldsymbol{\epsilon} = \vartheta_0^{-1}(\boldsymbol{\beta})\right)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\
&\quad = \left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \vartheta_0^{-1}(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} \frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}}^\top \\
&\qquad + \left[\frac{\partial}{\partial \boldsymbol{\beta}} \vartheta_0^{-1}(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log f_n^0(O_i|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \vartheta_0^{-1}(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right].
\end{aligned}$$

The first term (the product of a 3D tensor and a matrix) equals zero because $\boldsymbol{\epsilon} = \mathbf{0}$ maximizes $\boldsymbol{\epsilon} \mapsto \log f_n^0(O|\boldsymbol{\epsilon})$, so

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}} = \mathbf{0}.$$

The inner Hessian matrix of the second term equals $\Lambda_n^0(O)$. The surrounding matrices of the second term satisfy

$$\frac{\partial}{\partial \boldsymbol{\beta}} \vartheta_0^{-1}(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \left[\frac{\partial}{\partial \boldsymbol{\epsilon}} \vartheta_0(\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=\mathbf{0}}\right]^{-1}.$$

Recall that $\vartheta_0(\boldsymbol{\epsilon}) = B(F^0_{\boldsymbol{\epsilon}})$. Therefore, since B is pathwise differentiable,

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} \vartheta_0(\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=0} = P_0\left[D^*(P_0)s^\top\right],$$

where $s \in (L_0^2(P_0))^p$ is the score of F_n^0 at $\boldsymbol{\epsilon} = \mathbf{0}$ that is, by assumption on the submodel $\mathcal{F}^0$, $s = D^*(P_0)$. Therefore

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \vartheta_0^{-1}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &= \left\{P_0\left[D^*(P_0)D^*(P_0)^\top\right]\right\}^{-1} \\ &= P_0[\lambda^*(P_0)]^{-1},\end{aligned}$$

and

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log f_n^0\left(O \middle| \boldsymbol{\epsilon} = \vartheta_0^{-1}(\boldsymbol{\beta})\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = P_0[\lambda^*(P_0)]^{-1}\Lambda_n^0(O)P_0[\lambda^*(P_0)]^{-1}.$$

The Fisher Information is then

$$\begin{aligned}I_{\boldsymbol{\beta}_0} &:= \mathbb{E}_{P_0}\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log f_n^0\left(O \middle| \boldsymbol{\epsilon} = \vartheta_0^{-1}(\boldsymbol{\beta})\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \\ &= \mathbb{E}_{P_0}\left[P_0[\lambda^*(P_0)]^{-1}\Lambda_n^0(O)P_0[\lambda^*(P_0)]^{-1}\right] \\ &= P_0[\lambda^*(P_0)]^{-1}P_0[\Lambda_n^0(O)]P_0[\lambda^*(P_0)]^{-1} \\ &= P_0[\lambda^*(P_0)]^{-1}P_0[\lambda^*(P_0)]P_0[\lambda^*(P_0)]^{-1} \\ &= P_0[\lambda^*(P_0)]^{-1}.\end{aligned}$$

The result follows from van der Vaart (1998, Theorem 10.1). □

Example (cont'd). Applying Theorem 4 requires checking each of the assumptions in the context of the example.

- Assumption 1: the likelihood f_n^0 is that of an exponential family satisfying the conditions of (van der Vaart, 1998, Example 7.7), therefore $\{f_n^0(\cdot|\boldsymbol{\epsilon}) : \boldsymbol{\epsilon} \in \mathbb{R}^p\}$ is locally asymptotically normal.
- Assumption 2: it is impossible to directly check this condition; we propose a diagnostic in Section 6.3 that an oracle knowing P_0, and hence $\mathcal{F}^0$, could rely on to check the invertibility of ϑ_0.
- Assumption 3: we restrict $F_{\boldsymbol{\epsilon}}^0$ to be defined for $\boldsymbol{\epsilon} \in E := [-C, C]$ for an arbitrary (large) constant C. Because E is compact, assumption 3 is no longer necessary.
- Assumption 4: $\boldsymbol{\epsilon} \mapsto \log f_n^0(O \mid \boldsymbol{\epsilon})$ is infinitely differentiable.
- Assumption 5: it is easily checked that

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} \log f_n^0(O \mid \boldsymbol{\epsilon})\Big|_{\boldsymbol{\epsilon}=\mathbf{0}} = D_\sharp^*(P_0)(O).$$

- Assumption 6: the Hessian of $\boldsymbol{\epsilon} \mapsto \log f_n^0(O \mid \boldsymbol{\epsilon})$ evaluated at $\boldsymbol{\epsilon} = \mathbf{0}$ is

$$\begin{aligned}\Lambda_n^*(O) =& \frac{1}{g_0(A,X)^2}\left(\bar{Q}_0^{(A)}(X) - 1\right)\bar{Q}_0^{(A)}(X)M^{-1}(1,V)(1,V)^\top M^{-1} \\ &+ (\psi_0^\sharp(X) - B^\sharp(P_0)(1,V)^\top)^2 M^{-1}(1,V)(1,V)^\top M^{-1},\end{aligned}$$

  and its expectation is given by

$$P_0\left[\Lambda_n^*(O)\right] = M^{-1}P_0\left[\left(\frac{\mathrm{Var}_P(Y \mid A, X)}{g_0(A,X)^2} + \left(\psi_0^\sharp - B^\sharp(P_0)(1,V)^\top\right)^2\right)(1,V)(1,V)^\top\right]M^{-1},$$

  which is the variance of the EIF $D^*(P_0)$.
- Assumption 7: this condition can be easily satisfied by proper choice of the prior density π_β.

11.7 Metropolis-Hastings Algorithm

Let

$$f_n^*(O_{1:n} \mid \boldsymbol{\epsilon}) = \prod_{i=1}^{n} f(O_i|\boldsymbol{\epsilon}, \bar{Q}_n^*, Q_n^*, \eta_n, \sigma_n)$$

be the targeted likelihood akin to $f_n^0(O_{1:n} \mid \boldsymbol{\epsilon})$ that we defined and used in Section 11.6 (we simply substitute the estimated features for the true features). Let π_ϵ be a prior distribution for $\boldsymbol{\epsilon}$ as in (7). Let $\boldsymbol{\epsilon}^{(0)}$ be a starting value for the Markov Chain specified by the user, τ a tuning parameter, and T the number of iterations. The algorithm is implemented on the log-scale to avoid numerical precision issues. For $t = 1, \ldots, T$:

1. Draw $\boldsymbol{\epsilon}' \sim \mathrm{N}(\boldsymbol{\epsilon}^{(t-1)}, \tau \boldsymbol{I})$.
2. Draw $u \sim \mathrm{Uniform}(0, 1)$.
3. Let

$$\ell^{t-1} = \log f_n^*(O_{1:n} \mid \boldsymbol{\epsilon}^{(t-1)}) + \log \pi_\epsilon(\boldsymbol{\epsilon}^{(t-1)}),$$
$$\ell' = \log f_n^*(O_{1:n} \mid \boldsymbol{\epsilon}') + \log \pi_\epsilon(\boldsymbol{\epsilon}').$$

4. Let $A^{(t)} = \mathbb{I}(\ell' - \ell^{t-1} > \log(u))$. Then

$$\boldsymbol{\epsilon}^{(t)} = \begin{cases} \boldsymbol{\epsilon}' & \text{if } A^{(t)} = 1, \\ \boldsymbol{\epsilon}^{(t-1)} & \text{otherwise.} \end{cases}$$

5. Set $\boldsymbol{\beta}^{(t)} = B(\boldsymbol{\epsilon}^{(t)})$.

The mean $\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A^{(t)}$ is called the *acceptance ratio*.

The value of τ, the standard deviation of the proposal distribution, is found by a binary search procedure. Let $\tau_{\min}$ and $\tau_{\max}$ define the upper and lower search bounds for τ. Let K be the maximum number of iterations in the search. Let $\bar{A}(\tau)$ be the acceptance ratio from the Metropolis-Hastings algorithm given above with proposal distribution τ run for 1000 iterations. Then iterate:

1. Let $\tau_l = \tau_{\min}$ and $\tau_u = \tau_{\max}$.
2. For $k = 1, \ldots, K$:
   (a) Let $\tau_k = (\tau_l + \tau_u)/2$.
   (b) Let $\bar{A}_k = \bar{A}(\tau_k)$.
   (c) If $\bar{A}_k \in (0.3, 0.4)$, then break and return τ_k.
   (d) If $\bar{A}_k \leq 0.3$, set $\tau_u = \tau_k$.
   (e) If $\bar{A}_k \geq 0.4$, set $\tau_l = \tau_k$.
3. Return τ_K.

This procedure targets an acceptance ratio in between 30% and 40%.

11.8 Simulation Study Results

		β_1		β_2	
N	Estimator	95% Coverage	Absolute Bias	95% Coverage	Absolute Bias
(a): g_P correctly specified, $\bar{Q}_P$ correctly specified					
50	Frequentist	0.87	0.099	0.78	0.110
	Bayesian	0.90	0.100	0.85	0.110
100	Frequentist	0.92	0.069	0.91	0.066
	Bayesian	0.94	0.069	0.95	0.067
250	Frequentist	0.93	0.043	0.94	0.041
	Bayesian	0.94	0.043	0.95	0.042
500	Frequentist	0.96	0.030	0.95	0.030
	Bayesian	0.97	0.030	0.97	0.030
750	Frequentist	0.95	0.025	0.96	0.022
	Bayesian	0.96	0.024	0.97	0.023
1000	Frequentist	0.92	0.022	0.95	0.020
	Bayesian	0.92	0.022	0.95	0.020
(b): g_P correctly specified, $\bar{Q}_P$ incorrectly specified					
50	Frequentist	0.93	0.100	0.88	0.110
	Bayesian	0.95	0.100	0.95	0.110
100	Frequentist	0.96	0.068	0.90	0.080
	Bayesian	0.98	0.068	0.96	0.081
250	Frequentist	0.96	0.042	0.92	0.048
	Bayesian	0.98	0.042	0.95	0.049
500	Frequentist	0.95	0.032	0.94	0.035
	Bayesian	0.97	0.032	0.95	0.035
750	Frequentist	0.96	0.025	0.94	0.028
	Bayesian	0.97	0.025	0.95	0.028
1000	Frequentist	0.96	0.021	0.94	0.024
	Bayesian	0.97	0.021	0.95	0.025
(c): g_P incorrectly specified, $\bar{Q}_P$ correctly specified					
50	Frequentist	0.88	0.100	0.86	0.099
	Bayesian	0.89	0.100	0.88	0.100
100	Frequentist	0.92	0.068	0.92	0.067
	Bayesian	0.91	0.068	0.93	0.068
250	Frequentist	0.94	0.040	0.92	0.042
	Bayesian	0.95	0.040	0.94	0.042
500	Frequentist	0.95	0.029	0.96	0.029
	Bayesian	0.95	0.029	0.96	0.029
750	Frequentist	0.96	0.024	0.95	0.024
	Bayesian	0.96	0.024	0.94	0.024
1000	Frequentist	0.95	0.020	0.96	0.019
	Bayesian	0.95	0.020	0.96	0.019
(c): g_P incorrectly specified, $\bar{Q}_P$ incorrectly specified					
50	Frequentist	0.89	0.120	0.85	0.120
	Bayesian	0.90	0.120	0.89	0.120
100	Frequentist	0.88	0.087	0.92	0.076
	Bayesian	0.88	0.087	0.94	0.077
250	Frequentist	0.90	0.057	0.94	0.047
	Bayesian	0.89	0.057	0.96	0.047
500	Frequentist	0.82	0.049	0.95	0.032
	Bayesian	0.81	0.049	0.96	0.032
750	Frequentist	0.77	0.046	0.96	0.028
	Bayesian	0.76	0.046	0.97	0.028
1000	Frequentist	0.74	0.042	0.95	0.023
	Bayesian	0.71	0.042	0.95	0.023

Table 1: Empirical coverage of 95% credible (confidence) intervals and absolute bias of estimators in the simulation study.