University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 270

Targeted Bayesian Learning

Ivan Diaz Munoz* Alan E. Hubbard†
Mark J. van der Laan‡

*University of California, Berkeley, School of Public Health - Division of Biostatistics, idiazm@berkeley.edu

†University of California, Berkeley, hubbard@stat.berkeley.edu

‡University of California - Berkeley, laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper270

Copyright ©2010 by the authors.

Targeted Bayesian Learning

Ivan Diaz Munoz, Alan E. Hubbard, and Mark J. van der Laan

Abstract

Targeted maximum likelihood estimation (van der Laan & Rubin 2006) is a loss-based semi-parametric estimation method that yields a substitution estimator of a target parameter of the probability distribution of the data that solves the efficient influence curve estimating equation, and thereby yields a double robust locally efficient estimator of the parameter of interest, under regularity conditions. The Bayesian paradigm is concerned with including the researcher's prior uncertainty about the parameter through a prior distribution, which combined with the likelihood yields a posterior distribution for the parameter that reflects the researcher's posterior uncertainty. In this paper, we present a way to work under the Bayesian paradigm within the framework of targeted maximum likelihood estimation. In particular, we deal with the estimation of the so-called additive causal effect, but our results can be generalized to any d-dimensional parameter. For a general review of the proposed methodology, the readers referred to (van der Laan 2008, p. 178). We assess the performance of the proposed method through the asymptotic convergence of the posterior distribution to a normal limit distribution, the variance and bias of the mean of the posterior distribution, and the coverage probability of the credible interval implied by the posterior distribution.

1. Introduction

Statistical theory is concerned with deriving inferences from observations (data) of a random phenomenon about certain features of the probability mechanism that generates this phenomenon. Those features of interest are called parameters, and can usually be described as mappings between a set of possible distributions of the data, called *model*, and a d-dimensional real space. Models are in the core of statistical theory because they allow a description of the main features of the underlying probability mechanism based on prior knowledge about the phenomenon. There are three main approaches for the construction of a model: *parametric*, *semi-parametric*, and *non-parametric* models. A parametric model is one in which the data $O_1, O_2, \ldots, O_n$ is assumed to be generated by a probability distribution that belongs to a set of the form $\{P(O;\theta) : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$. In a semi-parametric model the parameter space Θ satisfies $\Theta \subset \mathbb{R}^k \times \mathbb{F}$, where $\mathbb{F}$ is an infinite dimensional space. A non-parametric model poses no restrictions on $P(O)$, and assumes that $P(O)$ belongs to the set of all possible distributions. Note that a non-parametric model is a special case of a semi-parametric model.

Statistical theory has been developed under two main paradigms: frequentist and Bayesian. In the context of inference, the main difference between these paradigms entails a conceptual distinction of the random nature of θ: in frequentist statistics θ is considered unknown but fixed, whereas Bayesian techniques treat it as a random variable.

Besides the model, whose elements are $P(O|\theta)$, Bayesian techniques incorporate to the process of inference a distribution on θ called *prior* distribution, whose density is denoted here by $\pi(\theta)$. More important than the discussion about the random nature of θ is the fact that Bayesian analysis incorporates an interpretation of the densities on θ as a way to summarize the current state of knowledge about it (Robert 2007, p. 34). Thus, $\pi(\theta)$ represents the certainty about the value of θ available prior to the recollection of $\mathbf{O}' = (O_1, O_2, \ldots, O_n)$, and $p(\theta|\mathbf{O})$ represents the certainty about it once the evidence contained in $\mathbf{O}$ is extracted and the prior information is updated. The latter is called the *posterior* density. Bayes's theorem allows the calculation of the posterior density as

$$p(\theta|\mathbf{O}) = \frac{p(\mathbf{O}|\theta)\pi(\theta)}{\int p(\mathbf{O}|\theta)\pi(\theta)d\theta}.$$

Despite the revolutionary recourse of the prior and posterior distributions, parametric Bayesian analysis suffers the same critical drawbacks as parametric frequentist analysis. First

COBRA
Research Archive

Hosted by The Berkeley Electronic Press

of all, the models used are typically very small (e.g., exponential families), and usually there is no justifiable reason to believe that the true probability distribution belongs to such small models. Choices of parametric models are often made based on the convenience of their analytical properties. Inferences about θ made according to such misspecified models are widely known to be biased.

Furthermore, the research interest usually rests in a parameter different from θ, that can be represented as a mapping from the model to a possibly multi dimensional real space. In this article we analyze the particular case of the additive causal effect. Given a full data set consisting of n independent and identically distributed copies of $O = (Y, A, W)$, where A is a binary treatment, Y is a binary or continuous outcome, and W is a vector of covariates, the additive causal effect is defined as

$$\psi_0 = \Psi(P_0) = E_W(E_0(Y|A = 1, W) - E_0(Y|A = 0, W)), \tag{1}$$

where P_0 is the distribution of O. Any possible density of O can be factorized as

$$p(O) = p(Y|A, W)p(A|W)p(W). \tag{2}$$

We define

$$\begin{aligned} Q_W(W) &\equiv P(W) \\ g(A, W) &\equiv p(A|W) \\ Q_Y(Y|A, W) &\equiv P(Y|A, W) \\ \bar{Q}(P)(A, W) &\equiv E_P(Y|A, W). \end{aligned}$$

We will occasionally use the notation $g(P)(A, W)$, to stress the dependence on P.

Standard Bayesian and frequentist techniques are aimed to do a very good job in doing inference about θ if the assumed model contains the true distribution, but substitution estimators and posterior distributions based on those techniques are not guaranteed to have optimal properties with respect to the target parameter.

Usual estimation techniques, such as maximum likelihood or mean squared error, fit densities to the data minimizing the empirical risk $\sum_i L(Q)(O_i)$ implied by some loss function $L(Q)(O)$, where Q is the relevant part of P that is needed to evaluate $\Psi(P) = \Psi(Q)$ (e.g., $Q = (\bar{Q}(P), Q_W)$). For our parameter of interest, if Y is continuous, a common choice of loss

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

function is the square loss $L(Q)(O) = (Y - \bar{Q}(P)(AW))^2$. If $\bar{Q}_n(A, W)$ is an estimator of $\bar{Q}(P_0)(A, W)$, and the empirical distribution is used as estimator of the marginal distribution of W, a substitution estimator of (1) is given by

$$\frac{1}{n}\sum_{i=1}^{n}[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)].$$

From a parametric Bayesian perspective, in order to get a posterior distribution of ψ_0, models for the distribution of W, and Y given A and W must be assumed. Let $\{Q_W(W; \theta_W) : \theta_W\}$ and $\{Q_Y(Y|A, W; \theta_Y) : \theta_Y\}$ be such models, and let the prior densities for θ_W and θ_Y be given by π_{θ_W} and π_{θ_Y}, respectively. Bayesian standard procedures can be used to compute posterior densities $\pi_{\theta_W|\mathbf{O}}$ and $\pi_{\theta_Y|\mathbf{O}}$, which using (1) can be mapped into a posterior density on ψ_0.

An important challenge one would face under a Bayesian framework is mapping a prior on ψ_0 into priors on θ_W and θ_Y. Parametric Bayesian techniques require that the prior information (usually proceeding from previous studies on the same phenomenon) be summarized in the form of prior densities on θ_W and θ_Y. Such previous studies are very likely to have used different sets of covariates W, and even different models for $Q_Y(Y|A, W)$ and $Q_W(W)$, thus providing information on different parameters θ_W^* and θ_Y^*. It is therefore more likely that information arising from such studies can be summarized (or is by nature available) in terms of a prior distribution on the parameter of interest, ψ_0, which in order to use parametric Bayesian techniques would have to be mapped into priors on θ_W and θ_Y. The Bayesian technique introduced here allows direct use of prior information on ψ_0.

Targeted Maximum Likelihood Learning (van der Laan & Rubin 2006) provides a semi-parametric frequentist framework in which the estimation procedure is targeted do the best possible job in estimating the parameter of interest. In this article we develop a strategy that allows to do targeted Bayesian inference of the additive causal effect, using the targeting principles presented in van der Laan & Rubin (2006).

The organization of the article is as follows. In Section 2. the targeted maximum likelihood estimation (TMLE) technique is introduced. In Section 3. we develop a procedure to work with the data and a prior distribution on ψ_0 in order to get its targeted posterior distribution. Section 4. deals with the asymptotic convergence of the proposed targeted posterior distribution. In Section 5. some frequentist properties of the posterior distribution are presented, and Section 6. presents a simulation study performed in order to assess other properties for which analytical results are not available. Finally, Section 7. presents a discussion on the results and

UC Berkeley Division of Biostatistics Working Paper Series
Research Archive

Hosted by The Berkeley Electronic Press

subsequent work in this area.

2. Targeted Maximum Likelihood Estimation

In this section we provide a brief introduction to the principles and uses of the TMLE. To get a more thorough understanding of its theoretical properties and implementation we refer the reader to the original paper (van der Laan & Rubin 2006).

Targeted maximum likelihood estimation is one of the possible approaches that allows efficient and double robust estimation of our parameter of interest. The details of the TMLE for this parameter can also be found in van der Laan & Rubin (2006). The efficiency of this technique comes from the fact that the estimated distribution solves the efficient influence curve equation of (1), given by $P_n D(P) = 0$. We use here the notation $Pf \equiv \int f(o)dP(o)$. P_n is the empirical distribution function, and $D(P)$ is the efficient influence curve of (1), given by

$$D(P)(O) = (Y - \bar{Q}(P)(A,W))\frac{2A-1}{g(A,W)} + \bar{Q}(P)(1,W) - \bar{Q}(P)(0,W) - \Psi(P). \qquad (3)$$

Under the conditions stated in Theorem 1 of van der Laan & Rubin (2006), a consistent estimator of P_0 that solves the efficient influence curve equation yields a substitution estimator of (1) that is asymptotically efficient, and thereby has the lowest asymptotic variance among all regular and asymptotically linear estimators.

Frequentist and Bayesian procedures of the sort described in the introduction will always be biased if the models used do not contain the true probability distribution. Furthermore, even when the right model is used, the estimating techniques employed are targeted to accurately describe the whole density of the data, which usually leads to a poor job in finding a good trade-off between bias and variance in the estimation of the parameter of interest.

The joint use of super learner (van der Laan et al. 2007) and the TMLE is a technique that aims to do a very good job on the global fit of p_0 (super learning), but uses the targeted MLE step to target the fit towards the parameter of interest. The targeting step is done in a way such that the final fit solves the efficient influence curve equation, causing a bias reduction in the substitution estimator of the parameter of interest. The super learner estimator of the density, p_n^0, is fluctuated by means of a parametric model through p_n^0 with parameter ϵ.

Below we transcribe the definition of the TMLE given in van der Laan & Rubin (2006).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Definition 1. Let $\mathcal{M}$ be the statistical model, and P_n the empirical distribution function. Given an initial estimator p_n^0; a parametric fluctuation $\{P_n^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ satisfying $p_n^0(0) = p_n^0$, and $\frac{d}{d\epsilon} \log p_n^0(\epsilon)_{\epsilon=0} = D(P_n^0)$; and a maximum likelihood estimator

$$\epsilon(P_n|p_n^0) = \arg\max_{\epsilon} \sum_{i=1}^{n} \log p_n^0(\epsilon)(O_i)$$

of ϵ, we define the first step targeted maximum likelihood estimator as

$$p_n^1 = p_n^0(\epsilon(P_n|p_n^0)).$$

This process can be iterated to define the k-step targeted maximum likelihood density estimator as

$$p_n^{k+1} = p_n^k(\epsilon(P_n|p_n^k)).$$

The targeted maximum likelihood estimator of p is defined as

$$p_n^* = \lim_{k\to\infty} p_n^k,$$

assuming this limit exists. The corresponding targeted maximum likelihood estimator of ψ_0 is defined as $\psi_n = \Psi(P_n^*)$.

A targeted maximum likelihood estimator of (1) when the outcome is continuous can be obtained by using a normal regression model for $\{P_n^0(\epsilon) : \epsilon\}$. Consider an initial estimator p_n^0 with the marginal density of W estimated by its empirical probability distribution, an estimator $g_n(A, W)$ of $g(A, W)$, and let the conditional density of Y be

$$Q_{Y,n}(Y|A, W) = \frac{1}{\sigma(A, W)} \phi\left(\frac{Y - \bar{Q}_n^0(A, W)}{\sigma(A, W)}\right),$$

where $\bar{Q}_n^0(A, W)$ is an initial estimator of $\bar{Q}(P_0)(A, W)$, and ϕ denotes the standard normal distribution. Consider the submodel

$$Q_{Y,n}(\epsilon)(Y|A, W) = \frac{1}{\sigma(A, W)} \phi\left(\frac{Y - \bar{Q}_n^0(A, W) - \epsilon H^*(A, W)}{\sigma(A, W)}\right),$$

where $H^*(A, W) = \left(\frac{2A-1}{g_n(A,W)}\right) \sigma^2(A, W)$. Note that this submodel fulfills the conditions of Definition 1. Let ϵ_n be the MLE of ϵ in this model. $p_n^1 = p_n^0(\epsilon_n)$ is the first step TMLE of p_0, with conditional density of Y given A and W given by a normal density with mean

COBRA
BERKELEY ELECTRONIC PRESS
Collection of Biostatistics
Research Archive

$\bar{Q}_n^1(A,W) = \bar{Q}_n^0(A,W) + \epsilon_n H^*(A,W)$. It can be shown that in this case convergence of the algorithm is achieved in the first step. The TMLE of ψ_0 is therefore given by

$$\Psi(P_n^1) = \frac{1}{n}\sum_{i=1}^{n}[\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i)].$$

This estimator has been proven to be consistent if either one of $g_n(A,W)$ and $\bar{Q}_n^0(A,W)$ is consistent, and it is efficient if both $g_n(A,W)$ and $\bar{Q}_n^0(A,W)$ are consistent.

3. Prior, Likelihood and Posterior Distributions

In this section we find the posterior distribution of ψ_0 when the likelihood of the parametric submodel employed in the TMLE is adopted as likelihood of the data.

Let $\bar{Q}_A(P)(W) \equiv \bar{Q}(P)(A,W)$. The parameter in (1) can be written as a mapping between $\mathcal{M}$ and $\mathbb{R}$, defined by

$$\Psi(P) = P\{\bar{Q}_1(P) - \bar{Q}_0(P)\}. \tag{4}$$

Treating p_n^0 as fixed, the fluctuation $\{P_n^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ used in the TMLE is just a parametric model, and the likelihood under this parametric model can be used together with the prior distribution to define the posterior distribution. This posterior distribution reflects the posterior uncertainty about the parameter and can be used to do point and interval estimation. Firstly, we find a submodel $\mathcal{M}_\epsilon = \{P_n^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ such that $p_n^0(0) = p_n^0$ and $\frac{d}{d\epsilon}\log p_n^0(\epsilon)|_{\epsilon=0} = D(P_n^0)$, where p_n^0 is the initial estimator of the density p_0, and is considered as fixed. Secondly, we determine the prior distribution on ϵ yielded by the prior on the parameter ψ_0. For this purpose we define a mapping $f(P_n^0) : \epsilon \to \Psi(P_n^0(\epsilon))$. Once the prior on ϵ is found, its posterior can be computed and the mapping $f(P_n^0)$ can be used to map the posterior of ϵ into a posterior of ψ_0.

Fluctuation Model:

We restrict our discussion to the case where a normal or a binomial model (depending on the type of outcome) is used as working model $Q_{Y,n}(\epsilon)(Y|A,W)$, but the validity of this working parametric model does not affect the consistency and efficiency of the TMLE or the proposed targeted posterior distribution of ψ_0. Furthermore, the general method described

COBRA
Collection of Biostatistics
Research Archive

here can be applied using any working model satisfying the conditions of Definition 1.

Consider an initial estimator p_n^0 of p_0. Estimators $\bar{Q}_n^0(A,W)$ and $g_n(A,W)$ can be obtained through standard procedures (e.g., logit or probit regression), or through more elaborated techniques, such as machine learning techniques. It is worth to emphasize that the efficiency and consistency of the TMLE depend on the choice of those initial estimators, which must be as close as possible to the real $\bar{Q}(P_0)(A,W)$ and $g_0(A,W)$. To achieve this goal, we encourage the use of the super learner (van der Laan et al. 2007). Super learner is a machine learning technique that given a library of candidate density estimators, a loss function and a sample size big enough, performs essentially as well or better than any of the candidates in the library, in terms of the chosen loss function.

Let $Q_{W,n}(W)$ be an initial estimator of $Q_W(W)$ (e.g., the empirical probability distribution of W). We fluctuate the initial estimator p_n^0 by finding a fluctuation of $\bar{Q}_n^0(A,W)$ and $Q_{W,n}(W)$ through ϵ, such that the score of $p_n^0(\epsilon)$ at $\epsilon = 0$ equals the efficient influence curve of Ψ at P_n^0, given in (3).

We use a binomial or normal distribution with constant variance for $Q_{Y,n}^0(\epsilon)(Y|A,W)$, where only the conditional expectation $\bar{Q}_n^0(A,W)$ is fluctuated. The fluctuations adopted here are given by

$$m(\bar{Q}_n^0(\epsilon)(A,W)) = m(\bar{Q}_n^0(A,W)) + \epsilon H_1^*(A,W)$$
$$Q_{W,n}(\epsilon)(W) = \frac{\exp(\epsilon H_2^*(W))Q_{W,n}(W)}{P_n\ \exp(\epsilon H_2^*)Q_{W,n}},$$

where

$$H_1^*(A,W) = \frac{2A-1}{g_n(A,W)}, \tag{5}$$
$$H_2^*(W) = \bar{Q}(P_n^0)(1,W) - \bar{Q}(P_n^0)(0,W) - \Psi(P_n^0), \tag{6}$$

and m is the logit or identity link, depending on the type of outcome. It can be shown that the model $p_n^0(\epsilon)$ obtained by using these fluctuations has score $D(P_n^0)$ at $\epsilon = 0$. In contrast to the classic TMLE for this parameter as described in van der Laan & Rubin (2006), in which the fluctuations of $\bar{Q}_n^0(A,W)$ and $Q_{W,n}(W)$ are done independently through ϵ_1 and ϵ_2, here we fluctuate both $\bar{Q}_n^0(A,W)$ and $Q_{W,n}(W)$ through a single ϵ. This is done in order to avoid dealing with a multivariate posterior distribution for $\epsilon^* = (\epsilon_1, \epsilon_2)'$. Ensuring that all the relevant parts of p_n^0 are fluctuated so that $\frac{d}{d\epsilon} \log p_n^0(\epsilon)\big|_{\epsilon=0} = D(P_n^0)$ results in a likelihood

CORE
Provided by the author(s) and University of Westminster
Research Archive

function with the right spread, that will ultimately result in the right coverage of the credible intervals if the initial estimator p_n^0 is consistent for p_0.

Prior Distribution on ϵ:

Let $\bar{Q}_{n,A}(\epsilon)(W) \equiv \bar{Q}_n^0(\epsilon)(A,W)$. The substitution estimator based on $p_n^0(\epsilon)$ is given by

$$\Psi(P_n^0(\epsilon)) = P_n^0(\epsilon)[\bar{Q}_{n,1}(\epsilon) - \bar{Q}_{n,0}(\epsilon)]. \tag{7}$$

From the Bayesian perspective, the uncertainty in the prior knowledge of ψ_0 can be incorporated into the inference procedure through a prior distribution on the parameter, namely $\psi_0 = \Psi(P_0) \sim \Pi$.

Let π be the density of Π. Note that the prior distribution of ψ_0 defines a prior distribution on ϵ through the mapping $f(P_n^0) : \epsilon \to \Psi(P_n^0(\epsilon))$. The fluctuation $p_n^0(\epsilon)$ must be chosen in a way such that this mapping is invertible. The prior on ϵ is given by

$$\pi^*(\epsilon) = \pi[\Psi(P_n^0(\epsilon))]J(\epsilon),$$

where $J(\epsilon)$ is the jacobian of the transformation, defined as

$$J(\epsilon) = \left| \frac{d}{d\epsilon} \Psi(P_n^0(\epsilon)) \right|.$$

Based on (7), we write

$$\frac{d}{d\epsilon}\Psi(P_n^0(\epsilon)) = nP_n \left\{ \frac{d\ p_n^0(\epsilon)}{d\epsilon} \left(\bar{Q}_{n,1}(\epsilon) - \bar{Q}_{n,0}(\epsilon)\right) + p_n^0(\epsilon) \left(\frac{d\ \bar{Q}_{n,1}(\epsilon)}{d\epsilon} - \frac{d\ \bar{Q}_{n,0}(\epsilon)}{d\epsilon} \right) \right\},$$

where

$$\frac{d\, p_n^0(\epsilon)(W)}{d\epsilon} = p_n^0(\epsilon)(W) \left[H_2^*(W) - \frac{P_n^0(H_2^* \exp(\epsilon H_2^*))}{P_n^0 \exp(\epsilon H_2^*)} \right].$$

It can also be shown that

$$\frac{d\, \bar{Q}_{n,A}(\epsilon)(W)}{d\epsilon} = H_1^*(A,W),$$

and

$$\frac{d\, \bar{Q}_{n,A}(\epsilon)(W)}{d\epsilon} = H_1^*(A,W)\ \bar{Q}_{n,A}(\epsilon)(W)[1 - \bar{Q}_{n,A}(\epsilon)(W)],$$

for continuous and binary outcomes, respectively.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Targeted Posterior Distribution:

From a Bayesian perspective, the conditional density of $O_1, O_2, \ldots, O_n$ given ϵ, is given by $\prod_{i=1}^{n} p_n^0(\epsilon)(O_i)$. Therefore, in our parametric working model $\{p_n^0(\epsilon) : \epsilon\}$, the posterior density of ϵ is proportional to

$$\pi^*(\epsilon) \prod_{i=1}^{n} p_n^0(\epsilon)(O_i). \tag{8}$$

Taking into account the factorization of the likelihood given in (2), and noting that the part of (8) corresponding to $g(A, W)$ does not involve ϵ, simulating observations from (8) is equivalent to simulating observations from

$$\pi^*(\epsilon) \prod_{i=1}^{n} Q_{Y,n}(\epsilon)(Y_i|A_i, W_i) Q_{W,n}(\epsilon)(W_i). \tag{9}$$

Standard Bayesian techniques such as the Metropolis-Hastings algorithm can be used to sample a large number of draws from this posterior distribution. Once a posterior sample ϵ_i ($i = 1, 2, \ldots, m$) is drawn from (9), a sample from the targeted posterior distribution of ψ_0 can be computed as $\psi_i = \Psi(P_n^0(\epsilon_i))$. The estimated posterior mean of ψ_0 can be used as point estimator, and a 95% credible interval can be estimated as $(\psi_{2.5}, \psi_{97.5})$, where ψ_k is the k-th percentile of this posterior distribution.

Note that simulating observations from this posterior distribution is just one possible way of computing the quantities of interest. In particular, one can use the posterior of ϵ and the mapping $f(P_n^0)$ to find the analytical form of the posterior distribution of ψ. Denote $\epsilon = f^{-1}(P_n^0)(\psi) = m(\psi)$, we have that

$$P(\psi|O_1, \ldots, O_n) \propto \left| \frac{d\ m(\psi)}{d\psi} \right| \pi^*(m(\psi)) \prod_{i=1}^{n} Q_{Y,n}(m(\psi))(Y_i|A_i, W_i) Q_{W,n}((\psi))(W_i),$$

where the constant of proportionality can be computed by using numerical integration. We can now calculate the value of the posterior distribution for any value ψ, plot the posterior distribution, or use numerical integration to find the analytical posterior mean or the posterior percentiles.

As a particular interesting case, the targeted posterior distribution when the TMLE procedure is implemented as in (van der Laan & Rubin 2006, p. 21) is presented in Appendix 2.

COBRA
Collection of Biostatistics
Research Archive

In this posterior distribution, if the TMLE of p_0 is used as initial estimator p_n^0, the posterior mean is equal to

$$\mu_{\psi_0|O} = \frac{w_1\psi_n + w_2\mu_{\psi_0}}{w_1 + w_2},$$

where ψ_n is the targeted maximum likelihood estimator, μ_{ψ_0} is the prior mean and w_1 and w_2 are weights given in Appendix 2. It is important to note that $w_2/w_1 \to 0$ when either the sample size increases or the variance of the prior distribution is very large. This means that in those situations the posterior mean reduces to the TMLE, acquiring its double robustness and efficiency.

4. Asymptotic Convergence of the Targeted Posterior Distribution

In standard Bayesian analysis, if X is a random variable distributed as the posterior, and θ_n is the maximum likelihood estimator of the parameter of the distribution of X, the variable $\sqrt{n}(X - \theta_n)$ can be shown to converge to a normal distribution with mean zero, and variance given by the inverse of the fisher information, whenever the model is correct (Lindley 1965). This result is analogue to the central limit theorem, and is very useful in establishing asymptotic properties of the Bayesian point and interval estimators, such as their asymptotic bias and coverage probability. It also implies that as the sample size increases, the information given by the prior is neglected, and only the data is used to make inferences.

An analogue result, presented in the next theorem, is valid in the case of the targeted posterior distribution when the TMLE p_n^* itself is used as initial estimator of p_0.

Theorem 1. Let p_n^* be the targeted maximum likelihood estimator of p_0, and let $\{P_n^*(\epsilon) : \epsilon\} \subset \mathcal{M}$ be a parametric fluctuation satisfying $p_n^*(0) = p_n^*$ and $\frac{d}{d\epsilon} \log p_n^*(\epsilon)|_{\epsilon=0} = D(P_n^*)$, where $D(P)$ is the efficient influence curve of $\Psi(P)$, defined in (3). Assume that there exists a distribution P^* such that $P_0[h(\psi_n, P_n^*) - h(\psi_0, P^*)]^2$ converges to zero, where

$$h(\psi, P)(O) = \frac{d^2}{d\psi^2} \log p(f^{-1}(P)(\psi))(O);$$

and that $h(\psi_n, P_n^*) - h(\psi_0, P^*)$ falls in a Glivenko-Cantelli class $\mathcal{F}$. Define $\psi_n = \Psi(P_n^*)$ (i.e.,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

ψ_n is the TMLE of ψ_0). Note that $S(\psi_n) = 0$, where

$$S(\psi) = \sum_{i=1}^{n} \frac{d}{d\psi} \log p_n^*(f^{-1}(P_n^*)(\psi))(O_i).$$

Assume that $\pi(\psi)$ is a prior density on ψ_0 such that $\pi(\psi) > 0$ for every possible value of ψ. Let $\tilde{\psi}_n$ be a random variable with posterior density proportional to (9). Provided that the mapping $f : \epsilon \to \Psi(P(\epsilon))$ is invertible, the sequence $\sqrt{n}(\tilde{\psi}_n - \psi_n)$ converges in distribution to T, where $T \sim N(0, \sigma^2)$ and

$$\begin{aligned}\sigma^2 &= -\left(P_0 \frac{d^2}{d\psi_0^2} \log p^*(f^{-1}(P^*)(\psi_0))\right)^{-1} \\ &= \frac{\left[P^*\left(\frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2\right)\right]^2}{P_0\left(\frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2\right)},\end{aligned}$$

with $\sigma^2(P^*)(A, W) = Var_{P^*}(Y|A, W)$ and $\bar{Q}_A^*(W) = \bar{Q}(P^*)(A, W)$.

A proof is provided in Appendix 1. Since ψ_n is double robust, this theorem teaches us that the targeted posterior distribution is also double robust in the sense that it will be centered at ψ_0 if either g_n or $\bar{Q}_n^0$ are correctly specified. Another important consequence is that if the limit P^* equals the true P_0, then the asymptotic variance of the posterior distribution is equal to

$$\sigma^2 = P_0\left(\frac{\sigma^2(P_0)}{g^2(P_0)} + (\bar{Q}_1(P_0) - \bar{Q}_0(P_0) - \Psi(P_0))^2\right),$$

where $\bar{Q}_A(P_0) = \bar{Q}(P_0)(A, W)$. This asymptotic variance equals the variance of the efficient influence curve at P_0. This means that asymptotic credible intervals are also confidence intervals (i.e., they have coverage probability $1 - \alpha$). A correction for the cases in which $p^* \neq p_0$ will be provided in the next section.

5. Frequentist Properties of the Targeted Posterior Distribution

Once the posterior sample ψ_i $(i = 1, 2, \ldots, m)$ is obtained, point estimates and $(1-\alpha)100\%$ credible intervals for ψ_0 can be computed as $\bar{\psi} = \frac{1}{m}\sum_{i=1}^{m} \psi_i$ and $(\psi_{[m\frac{\alpha}{2}]}, \psi_{[m(1-\frac{\alpha}{2})]})$, where the limits of the interval are given by order statistics and [ ] indicates rounding to the nearest integer.

COBRA
Collection of Biostatistics
Research Archive

Recall that the TMLE is double robust under certain conditions. Assume that those conditions and the conditions of Theorem 1 hold. Then, we have that $E(\tilde{\psi}_n - \psi_0) = E(\tilde{\psi}_n - \psi_n) + E(\psi_n - \psi_0)$ converges to zero. This means that the estimated posterior mean is also double robust.

As mentioned in the previous section, $(1-\alpha)100\%$ credible intervals only are guaranteed to have $(1-\alpha)100\%$ asymptotic coverage if the initial estimator p_n^0 converges to the true p_0. This is a very strong assumption in which we cannot usually rely. The next subsection provides a correction factor that can be applied to the credible intervals if they are required to have $1-\alpha$ asymptotic coverage probability.

5.1. Correction to the Credible Intervals

The TMLE can be written as

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o\left(\frac{1}{\sqrt{n}}\right),$$

where IC denotes the influence curve of ψ_n. Assume that the conditions of Theorem 1 hold, then we have that

$$\sqrt{n}(\tilde{\psi}_n - \psi_n) \to N(0, \sigma^2),$$
$$\sqrt{n}(\psi_n - \psi_0) \to N(0, \sigma^{2*}),$$

where σ^2 is given in Theorem 1 and $\sigma^{2*} = Var(IC(O))$. Denote by q_β the β percentile of the distribution of $\tilde{\psi}_n$. Then

$$q_\beta \simeq \psi_n + z_\beta \frac{\sigma}{\sqrt{n}},$$

where z_β is the β percentile of a standard normal distribution. This means that

$$\begin{aligned} P\left[(q_\beta, q_{1-\beta}) \ni \psi_0\right] &\simeq P\left(\psi_n - z_{1-\beta}\ \frac{\sigma}{\sqrt{n}} < \psi_0 < \psi_n + z_{1-\beta}\ \frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-z_{1-\beta}\ \frac{\sigma}{\sigma^*} < \frac{\sqrt{n}(\psi_n - \psi_0)}{\sigma^*} < z_{1-\beta}\ \frac{\sigma}{\sigma^*}\right). \end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Therefore, for the credible interval $(q_\beta,\ q_{1-\beta})$ to have coverage probability $1-\alpha$, the value of β must be chosen such that

$$z_{1-\beta}\ \frac{\sigma}{\sigma^*} = z_{1-\alpha/2}, \tag{10}$$

which means that $\beta = 1 - \Phi^{-1}\left(z_{1-\alpha/2}\ \frac{\sigma^*}{\sigma}\right)$, where $\Phi(x)$ is the $N(0,1)$ cumulative density function. Since P_0 and P^* are not known, the value of σ^2 cannot be computed. However an estimate can be obtained by replacing P_0 by P_n and P^* by P_n^0. The variance σ^{2*} can also be estimated by the empirical variance of $IC_n(O)$ (estimated influence curve).

6. Simulation

In order to explore some other frequentist properties of the targeted posterior distribution, and compare the Bayesian estimators with the classic TMLE, a simulation study was performed. In this section we describe the simulation scheme used, introduce the frequentist criteria used, and finally present the results.

The data was generated based on the following scheme:

Simulate W from $N_2\left(\begin{pmatrix}.5\\2\end{pmatrix}, \begin{pmatrix}2 & .3\\.3 & .8\end{pmatrix}\right)$.

Given $W = w$, simulate A from a bernoulli distribution with probability $expit(-.2 + .1w_1 - .2w_2 + .05w_1w_2)$, where *expit* is the inverse of the *logit* function.

Given $W = w$ and $A = a$, draw Y from a bernoulli distribution with probability $expit(-.2 + .07a - .2w_1 + .02w_2 + .2aw_1 - .5aw_2 - .01w_1w_2 - .003aw_1w_2)$.

This probability distribution yields a parameter value of $\psi_0 = -.1764$. For each of the sample sizes 30, 50, 100, 150, 200 and 250, one thousand data sets were generated. A beta distribution in the interval $(-1, 1)$ was used as prior, and three different sets of parameters were used, corresponding to a uniform prior, a beta density with mean ψ_0 and variance 0.1, and a beta density with mean ψ_0 and variance 0.25. The uniform prior corresponds to the situation in which no prior information is available, and the other two correspond to situations in which there are different levels of certainty about the prior information. These three priors are plotted in Figure 1.

CORA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

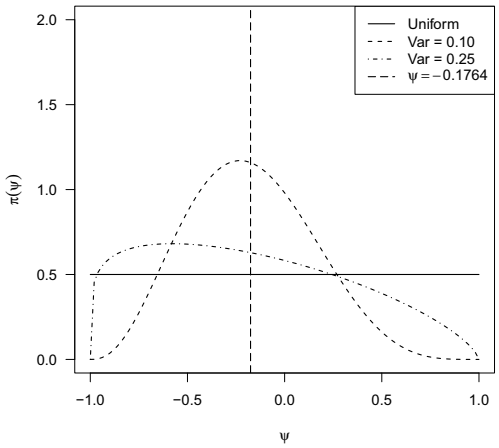

FIGURE 1: **Prior densities of ψ_0.**

Consider the following model

$$W \sim N_2(\mu, \Sigma);\ A|W \sim Ber(expit(X'\beta_1));\ Y|A, W \sim Ber(expit(M'\beta_2)),$$

where $X' = (1, W_1, W_2, W_1 \times W_2)$ and $M' = (X', A, A \times X')$. Note that this model contains the real data generating distribution.

A misspecified model (i.e., a model that does not include true Q_0) was also considered by not including interaction terms in M'. The TMLE estimator based on these two models was used as initial estimator p_n^0, and the Metropolis Hastings algorithm was used to draw 1000 observations from the posterior distribution given by (9). A brief description of this algorithm is presented in Appendix 3. The mean and variance of the posterior distribution were computed numerically, and a normal distribution was used as proposal density for the Metropolis Hastings algorithm. The average acceptance rate of this procedure was 70%.

The estimated posterior mean was used as estimator of ψ_0. Its variance and bias were estimated for each sample size. The 2.5th and 97.5th percentiles of the posterior sample were used as estimators of the limits of the 95% credible intervals; corrected credible intervals based on (10) were also computed. The performance of these intervals was assessed through their average length and coverage probability, estimated by the percentage of times that the interval contained the parameter. Bias, variance, coverage probability and average length were also computed for the classic TMLE and its confidence interval. The results are shown in Figure 2 and 3.

COBRA
Collection of Biostatistics
Research Archive

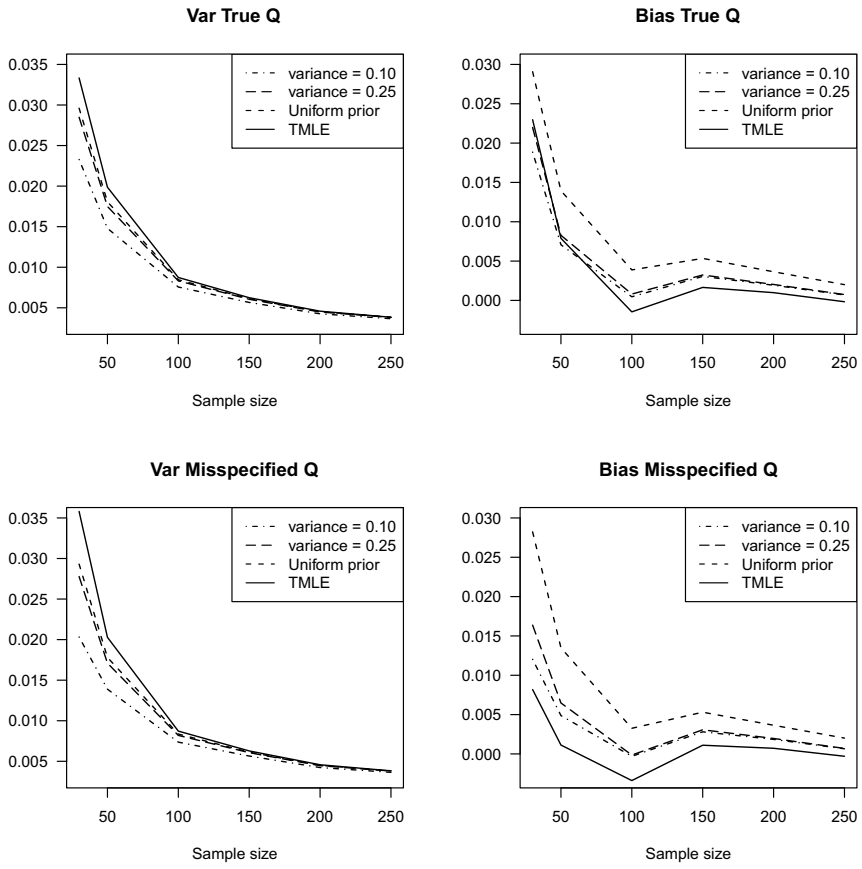

FIGURE 2: **Variance and bias of the posterior mean for different sample sizes when the model for $\bar{Q}$ is correctly and incorrectly specified.** (a) shows the variance for correctly specified $\bar{Q}$, (b) shows the bias for correctly specified $\bar{Q}$, (c) show the variance for misspecified $\bar{Q}$, and (d) shows the bias for misspecified $\bar{Q}$.

As expected, the inclusion of additional unbiased information reduces the variance of the estimators for small sample sizes, causing a bigger impact when the certainty about that additional knowledge is high. It is important to note that the variance of the posterior mean seems to be unaffected by the misspecification of the parametric model for $\bar{Q}_0$, though this simulation is not enough to believe that this type of robustness applies in general. Bayesian estimators appear to be more biased than the TMLE, specially if $\bar{Q}_0$ is misspecified and a uniform distribution is used as prior for ψ_0. However, all the estimators seem to be asymptotically unbiased.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

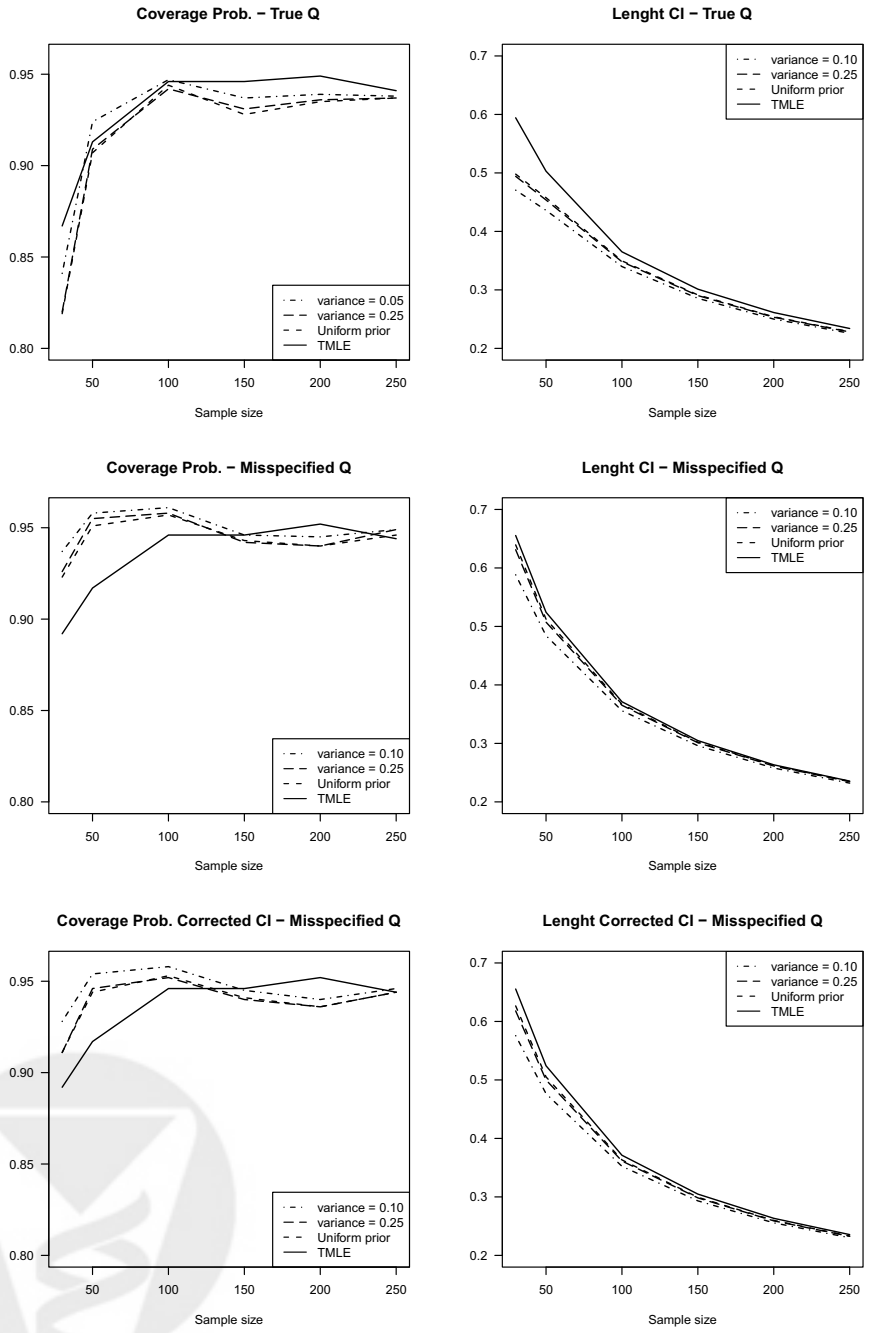

FIGURE 3: **Coverage probability and length of credible intervals for different sample sizes when the model for $\bar{Q}$ is correctly and incorrectly specified.** (a) shows the coverage probability for correctly specified $\bar{Q}$, (b) shows the length for correctly specified $\bar{Q}$ (c) show the coverage probability for misspecified $\bar{Q}$, (d) shows the length for misspecified $\bar{Q}$, (e) shows the coverage probability of the corrected intervals, and (e) shows the length of the corrected intervals.

Collection of Biostatistics
Research Archive

Figure 3 shows the coverage probability and length of corrected and uncorrected credible intervals for cases in which the true and misspecified $\bar{Q}_0$ are used. Although all the intervals have asymptotic right coverage, credible intervals with misspecified Q are somewhat conservative for some small sample sizes, having wider lengths and a coverage probability that is barely greater than the pre-specified level (95%). This means that the variance of the posterior distribution is greater if $\bar{Q}_0$ is misspecified, therefore reflecting some kind of "inefficiency" of the posterior distribution due to misspecification of $\bar{Q}_0$. The correction to the credible intervals proposed in (10) operates causing (above discussion) a slight and almost imperceptible decrease in the coverage probability and length of the intervals for all sample sizes, thereby providing an adjustment for the conservativeness of the intervals.

7. Discussion

A methodology to do targeted inference for the additive causal effect under the Bayesian paradigm is now available. Prior information on the effect of a binary treatment on an outcome can be directly used jointly with new data to update the knowledge about such effect. This update involves the computation of a targeted posterior distribution of the parameter of interest, whose mean has been found to be asymptotically double robust in the same sense as the targeted maximum likelihood estimator: it is a consistent estimator of the parameter of interest if either the model for the conditional expectation of the outcome or the treatment mechanism is misspecified. The asymptotic variance of the targeted posterior distribution has been proven to be equal to the variance of the efficient influence curve when the initial estimator of the density p_0 is consistent. This implies, amongst other characteristics, that credible intervals will also be confidence intervals in the sense that their credibility level will also be equal to their coverage probability. If consistency of the initial estimator is not a sensible assumption, but credible intervals are desired to have a specified coverage probability, a methodology to choose the right percentiles was provided.

A simulation study showed that misspecification of the model for the expectation of the outcome leads to wider credible intervals. Moreover, it showed that in the particular case studied, the uncorrected credible intervals based on misspecified $\bar{Q}_0$ also have the right asymptotic coverage probability, suggesting the possibility that for some cases, even if p_n^0 is not consistent, $1-\alpha$ credible intervals also have $1-\alpha$ coverage probability. It is therefore needed a more thorough understanding of the properties of the targeted posterior distribution, that allows us to identify such cases. The simulation also showed that the credible intervals for a misspeci-

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

fied $\bar{Q}_0$ were conservative for small sample sizes. The correction provided generated a slight correction of that conservativeness.

The methodology presented here is completely general, and is directly applicable to allow the computation of targeted posterior distributions for any pathwise differentiable parameter for which a TMLE can be computed. Future work in this area includes the determination of the analytical form of targeted posterior distributions for other interesting parameters, as well as simulations and theoretical studies that provide a comprehensive understanding of those targeted posterior distribution.

Appendix 1

Theorem 1. Let $u(P)(\psi)(O_i) \equiv p(f^{-1}(P)(\psi))(O_i)$. Let $\tilde{\psi}_n$ be a random variable with distribution given by the targeted posterior distribution of ψ_0.

$$\tilde{\psi}_n \sim p_{\tilde{\psi}_n}(\psi) \propto \pi(\psi) \prod_{i=1}^{n} u(P_n^*)(\psi)(O_i)$$

we define $T = \sqrt{n}(\tilde{\psi}_n - \psi_n)$. The density of T is given by

$$\begin{aligned} p_T(t) &= \frac{1}{\sqrt{n}}\, p_{\tilde{\psi}_n}\left(\psi_n + \frac{t}{\sqrt{n}}\right) \\ &\propto \pi\left(\psi_n + \frac{t}{\sqrt{n}}\right) \prod_{i=1}^{n} u(P_n^*)\left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i) \end{aligned}$$

We have that

$$\log p_T(t) = \log c + \log \pi\left(\psi_n + \frac{t}{\sqrt{n}}\right) + \sum_{i=1}^{n} \log u(P_n^*)\left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i)$$

A Taylor series expansion in t around zero yields

$$\begin{aligned} \sum_{i=1}^{n} \log u(P_n^*)\left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i) &= \frac{t^2}{2}\, \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\psi_n^2} \log u(P_n^*)(\psi_n)(O_i) + R_n \\ &= \frac{t^2}{2}\, P_n \frac{d^2}{d\hat{\psi_n}^2} \log u(P_n^*)(\psi_n) + R_n. \end{aligned} \tag{11}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Note that the linear term of this expansion vanishes because

$$\begin{aligned} S(\psi_n) &= \sum_{i=1}^{n} \frac{d}{d\psi_n} \log p_n^*(f^{-1}(P_n^*)(\psi_n))(O_i) \\ &= \frac{d}{d\psi} f^{-1}(P_n^*)(\psi)\Big|_{\psi=\psi_n} \sum_{i=1}^{n} \frac{d}{d\epsilon} \log p_n^*(\epsilon)(O_i)\Big|_{\epsilon=0} = 0, \end{aligned}$$

and $\epsilon_n = 0$ is the MLE of ϵ in the model $\{P_n^*(\epsilon) : \epsilon\}$. The remainder term R_n, which can be written as

$$\frac{t^3}{6} \frac{1}{n^{3/2}} \sum_{i=1}^{n} \frac{d^3}{d\psi_1^3} \log u(P_n^*)(\psi_1)(O_i)$$

for some ψ_1 between zero and ψ_n, is of order $n^{-\frac{1}{2}}$, and is therefore negligible compared with the other term in (11) which is of order 1.

Define $h_n = h(\psi_n, P_n^*)$ and $h_0 = h(\psi_0, P^*)$, and note that

$$P_n h_n - P_0 h_0 = (P_n - P_0)h_0 + (P_n - P_0)(h_n - h_0) + P_0(h_n - h_0).$$

The first term in this sum converges to zero by the law of the large numbers, the second term converges to zero because $h_n - h_0$ falls in a Glivenko-Cantelli class, and the last term converges to zero because it is bounded by $P_0(h_n - h_0)^2$, which converges to zero. This proves that

$$P_n \frac{d^2}{d\psi_n^2} \log u(P_n^*)(\psi_n) \longrightarrow P_0 \frac{d^2}{d\psi_0^2} \log u(P^*)(\psi_0)$$

which in turn proves that $p_T(t)$ converges, up to a constant, to

$$\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

where

$$\sigma^2 = \left(P_0 \frac{d^2}{d\psi_0^2} \log p^*(f^{-1}(P^*)(\psi_0))\right)^{-1}.$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

This asymptotic variance can be written as

$$\begin{aligned}
-\sigma^{-2} &= P_0 \frac{d^2}{d\psi_0^2} \log p_n^0(f^{-1}(P^*)(\psi_0)) \\
&= P_0 \left[\frac{d^2}{d\epsilon^2} \log p^*(\epsilon)\Big|_{\epsilon=0} \left(\frac{d}{d\psi_0} f^{-1}(P^*)(\psi_0) \right)^2 + \frac{d}{d\epsilon} \log p^*(\epsilon)\Big|_{\epsilon=0} \frac{d^2}{d\psi_0^2} f^{-1}(P^*)(\psi_0) \right] \\
&= P_0 \left[\frac{d^2}{d\epsilon^2} \log p^*(\epsilon)\Big|_{\epsilon=0} \left(\frac{d}{d\psi_0} f^{-1}(P^*)(\psi_0) \right)^2 \right]
\end{aligned}$$

Note that

$$-\frac{d^2}{d\epsilon^2} \log p^*(\epsilon)\Big|_{\epsilon=0} = \frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2,$$

where $\sigma^2(P^*)(A,W) = Var_{P^*}(Y|A,W)$ and $\bar{Q}_A^* = \bar{Q}(P^*)(A,W)$. On the other hand, since Ψ is pathwise differentiable we know that

$$\frac{d}{d\epsilon} \Psi(p^*(\epsilon))\Big|_{\epsilon=0} = P^*[D(P^*)s(P^*)]$$

where $D(P^*)$ is the canonical gradient given by the efficient influence curve at P^* and $s(P^*)$ is the score of $P^*(\epsilon)$ at $\epsilon = 0$ which is precisely $D(P^*)$. Therefore

$$\begin{aligned}
\left(\frac{d}{d\psi_0} f^{-1}(P^*)(\psi_0) \right)^2 &= (P^* D^2(P^*))^{-2} \\
&= \left[P^* \left(\frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2 \right) \right]^{-2},
\end{aligned}$$

and we conclude that

$$\sigma^2 = \frac{\left[P^* \left(\frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2 \right) \right]^2}{P_0 \left(\frac{\sigma^2(P^*)}{g^2(P^*)} + (\bar{Q}_1^* - \bar{Q}_0^* - \Psi(P^*))^2 \right)}$$

□

Appendix 2

Posterior distribution if only Q is fluctuated.

COBRA
Collection of Biostatistics
Research Archive

If the outcome is continuous, we can consider a linear model for $\bar{Q}_n^0(\epsilon)(A,W)$, this is

$$\bar{Q}_n^0(\epsilon)(A,W) = \bar{Q}_n^0(A,W) + \epsilon H_1^*(A,W), \tag{12}$$

where H_1^* is defined in (5). In this case, the mapping $\Psi(p_n^0(\epsilon))$ can be written as

$$\begin{aligned}\Psi(P_n^0(\epsilon)) =& P_n\left(\bar{Q}_{n,1} - \bar{Q}_{n,0}\right) + \epsilon P_n\left(H_{1,1}^* - H_{1,0}^*\right)\\ =& \Psi\left(P_n^0\right) + \epsilon P_n\left(H_{1,1}^* - H_{1,0}^*\right)\end{aligned} \tag{13}$$

where $Q_{n,A}(W) \equiv \bar{Q}_n^0(A,W)$ and $H_{1,A}^* \equiv H_1^*(A,W)$. The jacobian of this transformation is

$$J(\epsilon) = |P_n(H_{1,1}^* - H_{1,0}^*)|.$$

If a normal distribution with mean μ_{ψ_0} and variance $\sigma_{\psi_0}^2$ is considered as prior for ψ_0, the prior distribution for ϵ is given by

$$f_{\Pi^*}(\epsilon) = \frac{1}{\sigma_{\psi_0}}\phi\left(\frac{\Psi(P_n^0(\epsilon)) - \mu_{\psi_0}}{\sigma_{\psi_0}}\right)|P_n(H_{1,1}^* - H_{1,0}^*)|.$$

This implies that the prior of ϵ is a normal distribution with mean μ_ϵ and variance σ_ϵ^2, where

$$\mu_\epsilon = \frac{\mu_{\psi_0} - \Psi(P_n^0)}{P_n(H_{1,1}^* - H_{1,0}^*)}; \qquad \sigma_\epsilon = \frac{\sigma_{\psi_0}}{|P_n(H_{1,1}^* - H_{1,0}^*)|}.$$

Let us consider $Q_{Y,n}(\epsilon)(Y|A,W)$ to be a normal distribution with mean $\bar{Q}_n^0(A,W) + \epsilon H_1^*(A,W)$ and variance $\sigma^2(\bar{Q}_n^0)(A,W)$, and denote $\bar{Q}_n^0 \equiv \bar{Q}_n^0(A,W)$, $H_1^* \equiv H_1^*(A,W)$ and $\sigma_{\bar{Q}_n^0}^2 \equiv \sigma^2(\bar{Q}_n^0)(A,W)$. The part of the likelihood corresponding to $Q_{Y,n}(\epsilon)(Y|A,W)$ can be written as follows

$$\prod_{i=1}^{n} Q_{Y,n}(\epsilon)(Y_i|A_i,W_i) \propto \exp\left(-nP_n\frac{\left(Y - \bar{Q}_n^0 - \epsilon H_1^*\right)^2}{\sigma_{\bar{Q}_n^0}^2}\right).$$

Then, the posterior for ϵ is

$$\begin{aligned}p(\epsilon|O_1,\ldots O_n) &\propto \exp\left(-nP_n\frac{\left(Y - \bar{Q}_n^0 - \epsilon H_1^*\right)^2}{2\sigma_{\bar{Q}_n^0}^2} - \frac{(\epsilon - \mu_\epsilon)^2}{2\sigma_\epsilon^2}\right)\\ &\propto \exp\left(-\epsilon^2\left(nP_n\frac{(H_1^*)^2}{2\sigma_{\bar{Q}_n^0}^2} + \frac{1}{2\sigma_\epsilon^2}\right) + \epsilon\left(nP_n\frac{H_1^*(Y - \bar{Q}_n^0)}{\sigma_{\bar{Q}_n^0}^2} + \frac{\mu_\epsilon}{\sigma_\epsilon^2}\right)\right).\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Now let

$$\sigma^2_{\epsilon|O} = \left(nP_n \frac{(H_1^*)^2}{\sigma^2_{\bar{Q}_n^0}} + \frac{1}{\sigma^2_\epsilon} \right)^{-1}; \text{ and } \mu_{\epsilon|O} = \left(nP_n \frac{H_1^*(Y - \bar{Q}_n^0)}{\sigma^2_{\bar{Q}_n^0}} + \frac{\mu_\epsilon}{\sigma^2_\epsilon} \right) \sigma^2_{\epsilon|O}$$

Then,

$$p(\epsilon|O_1, \ldots O_n) \propto \exp\left(-\frac{(\epsilon - \mu_{\epsilon|O})^2}{2\sigma^2_{\epsilon|O}} \right),$$

which is a normal distribution with mean $\mu_{\epsilon|O}$ and variance $\sigma^2_{\epsilon|O}$.

Note that the maximum likelihood estimator of ϵ in the model (12), under a normal distribution, is given by

$$\epsilon_n = \frac{P_n \frac{H_1^*(Y - \bar{Q}_n^0)}{\sigma^2_{\bar{Q}_n^0}}}{P_n \frac{(H_1^*)^2}{\sigma^2_{\bar{Q}_n^0}}},$$

So that the posterior mean $\mu_{\epsilon|O}$ is, as expected, a weighted average of the maximum likelihood estimator and μ_ϵ, the prior mean of ϵ.

The posterior distribution of ψ_0 is also normal with mean

$$\mu_{\psi_0|O} = \Psi(P_n^0) + \mu_{\epsilon|O} P_n \left(H_{1,1}^* - H_{1,0}^* \right),$$

and variance

$$\sigma^2_{\psi_0|O} = \sigma^2_{\epsilon|O} [P_n \left(H_{1,1}^* - H_{1,0}^* \right)]^2.$$

By plugging in $\mu_{\epsilon|O}$ and $\sigma^2_{\epsilon|O}$, and working out the algebraic details, we get that

$$\mu_{\psi_0|O} = \frac{w_1 \left[\Psi(p_n^0) + \epsilon_n P_n (H_{1,1}^* - H_{1,0}^*) \right] + w_2 \mu_{\psi_0}}{w_1 + w_2} = \frac{w_1 \hat{\psi}_n + w_2 \mu_{\psi_0}}{w_1 + w_2},$$

$$\sigma^2_{\psi_0|O} = \frac{w_2}{w_1 + w_2} \sigma^2_{\psi_0},$$

where

$$w_1 = nP_n \frac{(H_1^*)^2}{\sigma^2_{Q_0}}; \text{ and } w_2 = \frac{[P_n (H_{1,1}^* - H_{1,0}^*)]^2}{\sigma^2_{\psi_0}}.$$

Note the posterior mean of ψ_0 is just a weighted average of the T-MLE of ψ_0 and its prior mean. Also, if the variance of the prior is very large compared to $[P_n(H_{1,1}^* - H_{1,0}^*)]^2$, the weight of the prior mean is very small, and the posterior mean of ψ_0 is just its TMLE.

COBRA Collection of Biostatistics Research Archive

Appendix 3

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method for sampling observations from a probability distribution whose analytic form is not easy to handle. Assume that $p(x)$ is the density from which observations are going to be drawn. The Metropolis-Hastings algorithm requires only that a function proportional to this density can be calculated. This is one of the most important aspects of the algorithm, since the constants of proportionality that arise in Bayesian applications are usually very difficult to compute. The algorithm generates a chain $x_1, x_2, \ldots, x_n$ by using a proposal density $q(x', x^i)$ at each step to generate a new proposed observation, x', that depends only on the previous state of the chain, x^i. This proposal is accepted as x^{i+1} if

$$\alpha < \min\left\{\frac{p(x')q(x^i, x')}{p(x^i)q(x', x^i)}, 1\right\},$$

where α is drawn from a uniform distribution in the interval $(0, 1)$. If the proposal is not accepted, the previous value is preserved in the chain, $x^{i+1} = x^i$. For additional references on the Metropolis-Hastings algorithm, the reader is referred to (Robert 2007, p.303).

For the sake of simulating observations from the targeted posterior distribution of ϵ, a normal distribution was used as proposal density. The mean and variance of the posterior were computed numerically, and used as parameters of this normal distribution. The starting value of the chain was set to zero. The acceptance rate was computed as the proportion of times that the proposal was accepted.

The R function used to draw samples of size n from the posterior distribution of ϵ is described below.

```
mh.epsilon <- function (n, posterior, e0, sd0){
  n = n + 1
  e <- cand <- pr <- numeric(n); e[1] <- e0
  for(i in 2:n){
      cand[i] <- rnorm(1, mean = e[i-1], sd = sd0)
      p <- (posterior(cand[i]) * dnorm(e[i-1],mean = cand[i],
           sd = sd0))/(posterior(e[i-1]) * dnorm(cand[i],
           mean = e[i-1], sd = sd0))
      pr[i] <- min(p, 1)
      e[i] <- sample(c(cand[i], e[i-1]), 1, prob=c(pr[i],
           1-pr[i]))
```

COBRA
Collection of Biostatistics
Research Archive

```
}
return(e[-1])}
```

References

Lindley, D. V. (1965), *Introduction to Probability and Statistics from a Bayesian Point of View, Part 2*, Cambridge University Press, Cambridge.

Robert, P. C. (2007), *The Bayesian Choice: From Decision-Thoeretic Foundations to Computational Implementation*, Springer, New York.

van der Laan, M., Polley, E. & Hubbard, A. (2007), 'Super learner', *Statistical Applications in Genetics and Molecular Biology* **6**(25).

van der Laan, M. & Rubin, D. (2006), 'Targeted maximum likelihood learning', *The International Journal of Biostatistics* **2**(1).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive