Chapter 28
Targeted Bayesian Learning

Iván Díaz Muñoz, Alan E. Hubbard, Mark J. van der Laan

TMLE is a loss-based semiparametric 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 probability distribution through a prior distribution on a statistical model for the probability distribution, which combined with the likelihood yields a posterior distribution of the probability distribution that reflects the researcher's posterior uncertainty. Just like model-based maximum likelihood learning, Bayesian learning is intrinsically nontargeted by working with the prior and posterior distributions of the whole probability distribution of the observed data structure and is thereby very susceptible to bias due to model misspecification or nontargeted model selection.

In this chapter, we present a targeted Bayesian learning methodology mapping a prior distribution on the target parameter of interest into a valid posterior distribution of this target parameter. It relies on a marriage with TMLE, and we show that the posterior distribution of the target parameter inherits the double robust properties of the TMLE. In particular, we will apply this targeted Bayesian learning methodology to the additive causal effect, but our results can be generalized to any d-dimensional target parameter. For a general review of the proposed methodology, we refer the interested reader to van der Laan (2008b), p. 178.

Statistical theory is concerned with deriving inferences from observations (data) on a random variable about certain features of the probability mechanism that generates this random variable. Those features of interest are called parameters and can be described as mappings from a set of possible distributions of the data, called a *model*, to a d-dimensional real space. Models are at 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 experiment that generated the random variable. A model can be classified in three main categories: *parametric*, *semiparametric*, and *nonparametric* models. A parametric model is one in which

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_28,

© Springer Science+Business Media, LLC 2011

the i.i.d. random variables $O_1, O_2, \ldots, O_n$ are assumed to be generated by a probability distribution P_0 that belongs to a set of the form $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$. In a semiparametric model the parameter space Θ satisfies $\Theta \subset \mathbb{R}^k \times \mathbb{F}$, where $\mathbb{F}$ is an infinite-dimensional space. A nonparametric model poses no restrictions on P_0 and assumes that P_0 belongs to the set of all possible distributions. Note that a nonparametric model is a special case of a semiparametric 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_θ, Bayesian techniques incorporate a *prior* distribution on θ in the statistical inference, whose density is denoted here by π. More important than the randomness 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 collection of $\mathbf{O}' = (O_1, O_2, \ldots, O_n)$, and $p(\theta \mid \mathbf{O})$ 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

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

Despite the revolutionary recourse of the prior and posterior distributions, parametric Bayesian analysis suffers from the same critical drawbacks as parametric frequentist analysis. First 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 multidimensional real space. In this article we analyze the particular case of the additive causal effect whose definition we now recall. Given a data set consisting of n identically distributed copies of $O = (W, A, Y)$, 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

$$\psi_0 = \Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)], \qquad (28.1)$$

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

$$P(O) = P(Y \mid A, W)P(A \mid W)P(W). \qquad (28.2)$$

We define: $Q_W(W) \equiv P(W)$, $g(A \mid W) \equiv P(A \mid W)$, $Q_Y(Y \mid A, W) \equiv P(Y \mid A, W)$, and $\bar{Q}(P)(A, W) \equiv E_P(Y \mid A, W)$. We will occasionally use the notation $g(P)(A \mid W)$, to stress the dependence on P.

Standard Bayesian and frequentist techniques do well regarding inference for θ if the assumed model is small enough and contains the true distribution (consistency, efficiency, and central limit theorem), but, for general semiparametric models, substitution estimators and posterior distributions of the parameter of interest (i.e., additive causal effect) based on those techniques are not guaranteed to have optimal properties with respect to the target parameter.

Classical estimation techniques, such as maximum likelihood estimation or least squares estimation, fit densities to the data by minimizing the empirical risk $\sum_i L(Q)(O_i)$ implied by some loss function $L(Q)$. Here Q is the relevant part of P that is required to evaluate $\Psi(P) = \Psi(Q)$ [e.g., in the additive causal effect example, $Q = (\bar{Q}(P), Q_W)$]. For our parameter of interest, if Y is continuous, a common choice of loss function for the conditional mean $\bar{Q}_0$ is the square loss $L(\bar{Q})(O) = (Y - \bar{Q}(W, A))^2$. If one estimates $\bar{Q}(P_0)$ with $\bar{Q}_n$, and the marginal distribution of W by its empirical counterpart, then the substitution estimator of ψ_0 is given by

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

In the parametric Bayesian paradigm, models for the marginal distribution of W and conditional distribution of Y given (A, W) must be assumed in order to get a posterior distribution of the parameter ψ_0. Let $\{Q_W(W; \theta_W) : \theta_W\}$ and $\{Q_Y(Y \mid A, W; \theta_Y) : \theta_Y\}$ 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 can be mapped into a posterior density on ψ_0 by (28.1).

Indeed, it is very likely that (1) prior information on the treatment effect ψ_0 itself is readily available and that (2) previous studies of the same treatment were analyzed based on different sets of covariates W, and different models for $Q_Y(Y \mid A, W)$ and $Q_W(W)$, thus providing information on different parameters θ'_W and θ'_Y. The Bayesian technique introduced here only requires a prior distribution on ψ_0, allows for realistic semiparametric models, and it maps it into a posterior distribution on ψ_0 with frequentist properties analogous to the TMLE.

28.1 Prior, Likelihood, and Posterior Distributions

In this section we determine the posterior distribution of ψ_0 when the likelihood of the parametric submodel employed in the TMLE is adopted as the likelihood of the data. For notational convenience, let $\bar{Q}_A(P)(W) \equiv \bar{Q}(P)(A, W)$. The parameter in (28.1) can be written as a mapping from $\mathcal{M}$ to $\mathbb{R}$, defined by

$$\Psi(P) = P\{\bar{Q}_1(P) - \bar{Q}_0(P)\}. \tag{28.3}$$

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 a prior distribution on ϵ to define the posterior distribution of ϵ. The corresponding posterior distribution of $\Psi(P_n^0(\epsilon))$ reflects the posterior uncertainty about the target parameter. It can be used to proceed to point and interval estimation of the target parameter of interest.

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 $\langle D^*(P_n^0)\rangle \subset \langle \frac{d}{d\epsilon} \log P_n^0(\epsilon)|_{\epsilon=0}\rangle$, where P_n^0 is the initial estimator of P_0 and is considered fixed. Here $D^*(P)$ is the efficient influence curve of Ψ at P, and we used the notation $\langle S\rangle$ for the linear span generated by the components of the function S of O. Secondly, we determine the prior distribution on ϵ yielded by the prior on the parameter ψ_0. For this purpose we define a mapping $f_n : \epsilon \rightarrow \Psi(P_n^0(\epsilon))$. Once the prior on ϵ is determined, its posterior can be computed and the mapping f_n can be used to map the posterior on ϵ into a posterior on ψ_0.

Fluctuation model. We consider a normal working model for $Q_{Y,n}(\epsilon)$ when Y is continuous and a logit regression model when Y is binary. If Y is continuous, the loss function $-\log Q_{Y,n}(\epsilon)$ corresponds with the squared error loss for the conditional mean $\bar{Q}_n(\epsilon)$. As a consequence, the TMLE and the proposed targeted posterior distribution of ψ_0 are not affected by the validity of this normal working model.

Consider an initial estimator P_n^0 of P_0: estimators $\bar{Q}_n^0$ and g_n 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 emphasizing 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 $\bar{Q}(P_0)$ and g_0. To achieve this goal, we encourage the use of the super learner. Let $Q_{W,n}$ be an initial estimator of Q_W (e.g., the empirical probability distribution of W). We fluctuate the initial estimator P_n^0 by finding a fluctuation of $\bar{Q}_n^0$ and $Q_{W,n}$ 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 by

$$D^*(P)(O) = (Y - \bar{Q}(P)(A, W))\frac{2A-1}{g(A \mid W)} + \bar{Q}(P)(1, W) - \bar{Q}(P)(0, W) - \Psi(P). \quad (28.4)$$

We use either a binomial working model (case Y binary) or a constant variance normal working model (case Y continuous) for $Q_{Y,n}^0(\epsilon)$. The fluctuations adopted here are given by

$$\begin{aligned} m(\bar{Q}_n^0(\epsilon)) &= m(\bar{Q}_n^0) + \epsilon H_1^*, \\ Q_{W,n}(\epsilon) &= \frac{\exp(\epsilon H_2^*)}{Q_{W,n}\ \exp(\epsilon H_2^*)} Q_{W,n}, \end{aligned} \quad (28.5)$$

where

$$H_1^*(A, W) = \frac{2A-1}{g_n(A, W)}, \quad (28.6)$$

$$H_2^*(W) = \bar{Q}(P_n^0)(1, W) - \bar{Q}(P_n^0)(0, W) - \Psi(P_n^0), \quad (28.7)$$

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, in which the fluctuations of $\bar{Q}_n^0$ and $Q_{W,n}$ are done independently through ϵ_1 and ϵ_2, and the maximum likelihood estimator of ϵ_2 happens to be zero, here we fluctuate both $\bar{Q}_n^0$ and $Q_{W,n}$ 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 $d/d\epsilon \log P_n^0(\epsilon)\big|_{\epsilon=0} = D^*(P_n^0)$ results in a likelihood function with the right spread, which 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. For notational convenience, 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

$$\begin{aligned}\Psi(P_n^0(\epsilon)) &= Q_{W,n}(\epsilon)[\bar{Q}_{n,1}(\epsilon) - \bar{Q}_{n,0}(\epsilon)] \qquad (28.8)\\ &= \sum_{i=1}^{n} \frac{\exp(\epsilon H_2^*(W_i))Q_{W,n}(W_i)}{\sum_{j=1}^{n}\exp(\epsilon H_2^*(W_j))Q_{W,n}(W_j)}[\bar{Q}_{n,1}(\epsilon)(W_i) - \bar{Q}_{n,0}(\epsilon)(W_i)].\end{aligned}$$

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

Let π be the density of Π. Note that the prior distribution of ψ_0 defines a prior distribution on ϵ through the mapping $f_n : \epsilon \rightarrow \Psi(P_n^0(\epsilon))$. The fluctuation $p_n^0(\epsilon)$ must be chosen in a such way 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 (28.8), we obtain

$$\begin{aligned}\frac{d}{d\epsilon}\Psi(P_n^0(\epsilon)) = \sum_{i=1}^{n}\Bigg\{&\frac{d\,Q_{W,n}(\epsilon)(W_i)}{d\epsilon}(\bar{Q}_{n,1}(\epsilon)(W_i) - \bar{Q}_{n,0}(\epsilon)(W_i))\\ &+ Q_{W,n}(\epsilon)(W_i)\frac{d}{d\epsilon}(\bar{Q}_{n,1}(\epsilon)(W_i) - \bar{Q}_{n,0}(\epsilon)(W_i))\Bigg\}, \qquad (28.9)\end{aligned}$$

where

$$\frac{d\,Q_{W,n}(\epsilon)(W)}{d\epsilon} = Q_{W,n}(\epsilon)(W)\left[H_2^*(W) - \frac{Q_{W,n}(H_2^*\exp(\epsilon H_2^*))}{Q_{W,n}\exp(\epsilon H_2^*)}\right],$$

and $Q_{W,n}(\epsilon)$ is defined in (28.5). 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.

Targeted posterior distribution. Operating from a Bayesian perspective under the working fluctuation model, the conditional density of $O_1, O_2, \ldots, O_n$ given ϵ equals $\prod_{i=1}^n P_n^0(\epsilon)(O_i)$. Therefore, in our parametric working model $\{P_n^0(\epsilon) : \epsilon\}$, the posterior density of ϵ is proportional to

$$\pi^*(\epsilon)\prod_{i=1}^n P_n^0(\epsilon)(O_i). \tag{28.10}$$

Taking into account the factorization of the likelihood given in (28.2), and noting that the part of (28.10) corresponding to $g_n(A \mid W)$ does not involve ϵ, simulating from (28.10) is equivalent to simulating from the density proportional to

$$\pi^*(\epsilon)\prod_{i=1}^n Q_{Y,n}(\epsilon)(Y_i \mid A_i, W_i)Q_{W,n}(\epsilon)(W_i). \tag{28.11}$$

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 (28.11), 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 is $(\psi_{2.5}, \psi_{97.5})$, where ψ_k is the kth percentile of this posterior distribution.

Note that simulating observations from this posterior distribution is just one possible way of computing the quantities of interest. Alternatively, one can use the analytic formula of the posterior density of ϵ and the mapping f_n to find the analytical form of the posterior distribution of ψ. Recall that f_n is assumed to be invertible. Note that $\epsilon = f_n^{-1}(\psi)$. We have

$$\begin{aligned} P(\psi \mid O_1, \ldots, O_n) \propto & \left|\frac{d\,f_n^{-1}(\psi)}{d\psi}\right| \pi^*(f_n^{-1}(\psi)) \times \\ & \prod_{i=1}^n Q_{Y,n}(f_n^{-1}(\psi))(Y_i \mid A_i, W_i)Q_{W,n}(f_n^{-1}(\psi))(W_i), \end{aligned} \tag{28.12}$$

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 and Rubin (2006, p. 21), is presented in Appendix 2 of this chapter. 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 TMLE, μ_{ψ_0} is the prior mean, and w_1 and w_2 are weights given in Appendix 2 of this chapter. 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.

28.2 Convergence of 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 1980). This result is analogous to the central limit theorem and is very useful in establishing the 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 are used to make inferences. An analogous 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 28.1. *Let P_n^* be the TMLE of P_0, and let $\{P_n^*(\epsilon) : \epsilon\} \subset \mathcal{M}$ be a parametric fluctuation satisfying $P_n^*(0) = P_n^*$ and $d/d\epsilon \log P_n^*(\epsilon)|_{\epsilon=0} = D^*(P_n^*)$, where $D^*(P)$ is the efficient influence curve of $\Psi(P)$, defined in (28.4). Define the mapping $f_n^* : \epsilon \to \Psi(P_n^*(\epsilon))$ to be invertible. 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*

$$h(\psi_n, P_n^*)(O) \equiv \frac{d^2}{d\psi^2} \log p(f_n^{*-1}(\psi))(O)\Big|_{\psi=\psi_n}$$

and $h(\psi_0, P^*)$ is defined analogously. Assume 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., ψ_n is the TMLE of ψ_0). Note that $S(\psi_n) = 0$, where

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

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 (28.12). The sequence $\sqrt{n}(\tilde{\psi}_n - \psi_n)$ converges in distribution to T, where $T \sim N(0, \sigma^2)$ and

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

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

A proof is provided in Appendix 1 of this chapter. 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 $\bar{Q}_n^*$ (as used by the TMLE P_n^*) is consistent. 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 $D^*(P_0)$ at P_0, providing the analogue of the standard result cited above (Lindley 1980). 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.

28.3 Frequentist Properties of 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. Recall that the TMLE is double robust under certain conditions. Assume that those conditions and the conditions of Theorem 28.1 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. We only wish to rely on the consistency of either g_n or $\bar{Q}_n$, so that the posterior mean is consistent (and asymptotically linear). We now provide a correction factor that can be applied to the credible intervals so that they preserve the desired $(1-\alpha)$ asymptotic coverage when the TMLE is consistent and asymptotically linear.

Under the assumptions for asymptotic linearity, the TMLE satisfies

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

where IC denotes the influence curve of ψ_n. Assume that the conditions of Theorem 28.1 hold, so 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 28.1 and $\sigma^{2*} = var_{P^*}(IC(O))$. Denote the β-percentile of the distribution of $\tilde{\psi}_n$ with q_β. 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

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

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

$$z_{1-\beta}\frac{\sigma}{\sigma^*} = z_{1-\alpha/2}, \tag{28.13}$$

hence

$$\beta = 1 - \Phi^{-1}\left(z_{1-\alpha/2}\frac{\sigma^*}{\sigma}\right),$$

where Φ is the $N(0, 1)$ cumulative distribution function. Since P_0 and P^* are unknown, the values of σ^2 and σ^{2*} cannot be computed explicitly. However, estimates can be obtained by replacing P_0 with P_n and P^* with P_n^0. The variance σ^{2*} can also be estimated by the empirical variance of the estimated influence curve values $IC_n(O_i)$, $i = 1, \ldots, n$.

28.4 Simulations

In order to explore additional frequentist properties of the targeted posterior distribution, and, in particular, to compare the posterior mean of the targeted posterior distribution with the TMLE itself, a simulation study was performed. We only considered the case where Y is binary. The data were generated based on the following scheme:

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

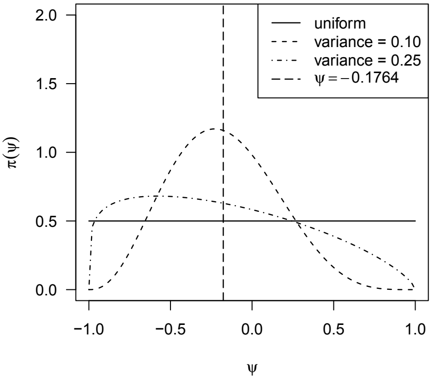


Fig. 28.1 Prior densities of ψ_0

2. Given $W = w$, simulate A from a Bernoulli distribution with probability $\text{expit}\,(-0.2 + 0.1w_1 - 0.2w_2 + .05w_1 \times w_2)$, where expit is the inverse of the logit function.
3. Given $W = w$ and $A = a$, draw Y from a Bernoulli distribution with probability $\text{expit}\,(-0.2 + 0.07a - 0.2w_1 + 0.02w_2 + 0.2a \times w_1 - 0.5a \times w_2 - 0.01w_1 \times w_2 - 0.003a \times w_1 \times w_2)$.

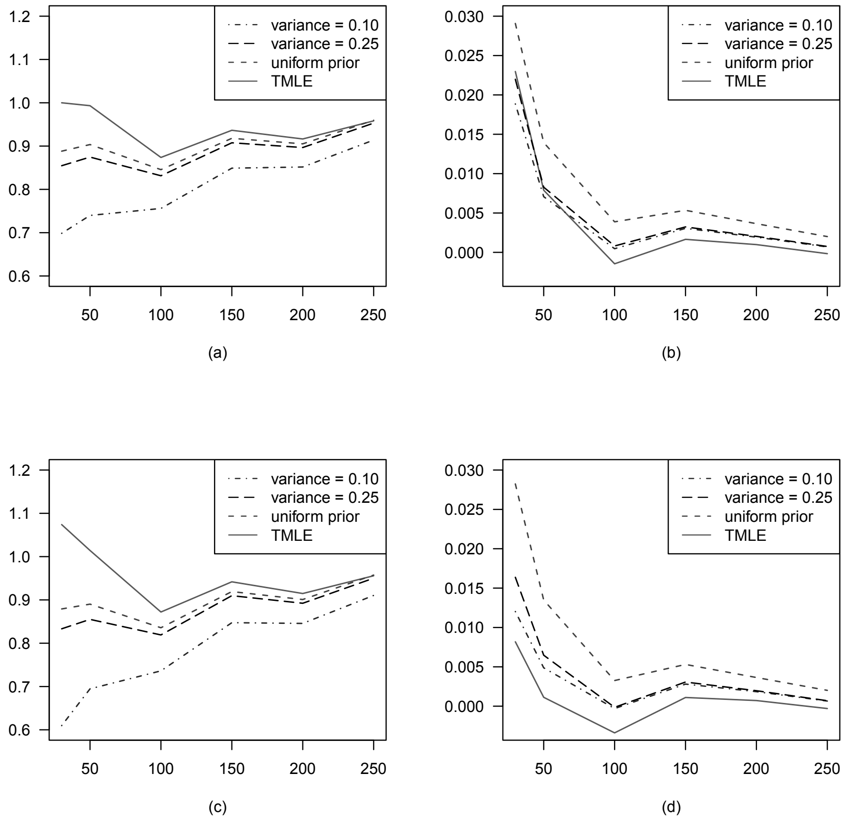
This probability distribution yields a parameter value of $\psi_0 = -.1764$. For each of the sample sizes 30, 50, 100, 150, 200, and 250, 1000 data sets were generated. We consider three different prior distributions on ψ_0, all from the beta family in the interval $(-1, 1)$. The first one boils down to a uniform prior, while the second and third ones have mean ψ_0 and variances 0.1 and 0.25, respectively. 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 Fig. 28.1.

Consider the following model:

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

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 (28.11). A brief description of this algorithm is presented in Appendix 3 of this chapter. 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 (28.13) 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 true parameter value ψ_0. Bias, variance, coverage probability, and average length were also computed for the TMLE and its confidence interval. The results are shown in Figs. 28.2 and 28.3.



(a)

(b)

(c)

(d)

Fig. 28.2 Variance and bias of the posterior mean. (a) Variance (multiplied by n) for correctly specified $\bar{Q}$. (b) Bias for correctly specified $\bar{Q}$. (c) Variance (multiplied by n) for misspecified $\bar{Q}$. (d) Bias for misspecified $\bar{Q}$

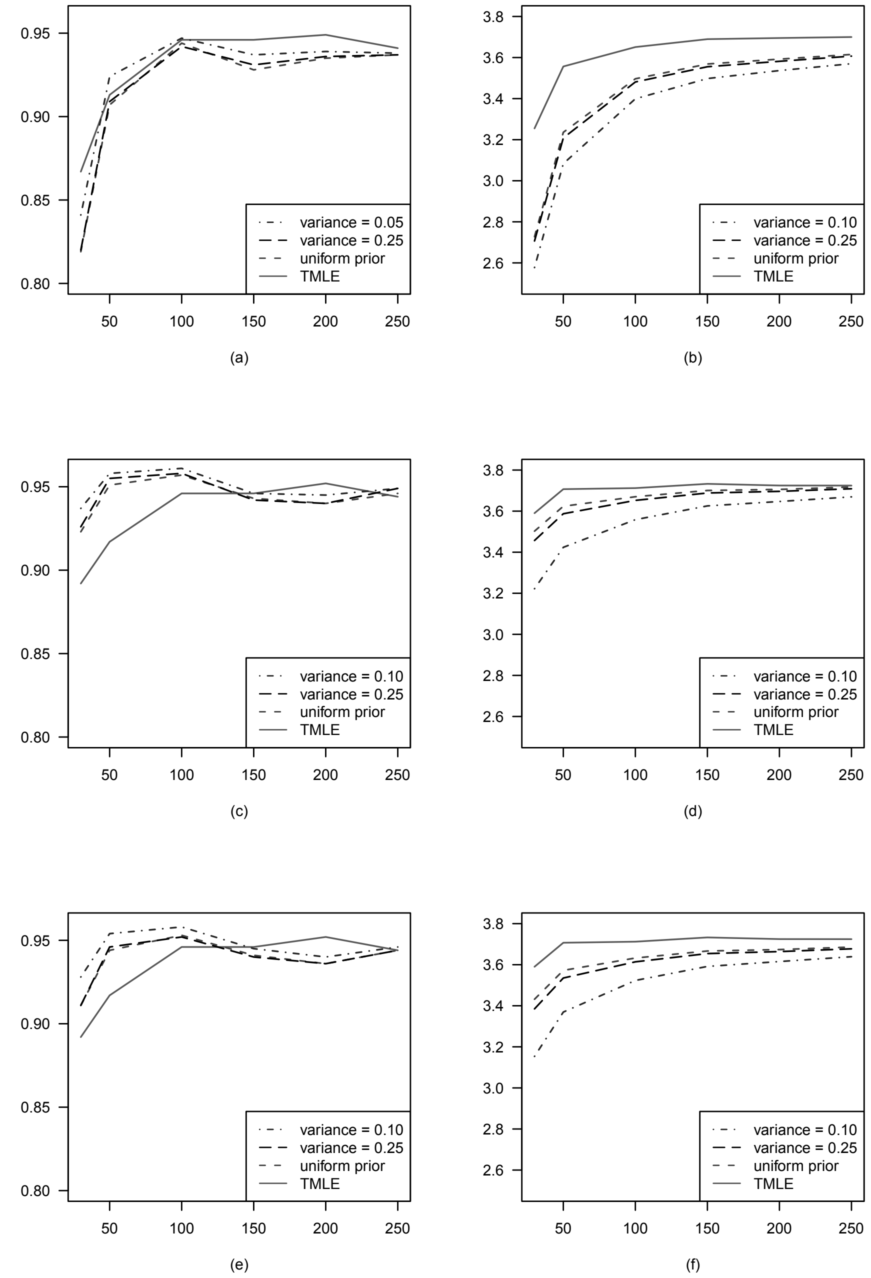


Fig. 28.3 Coverage probability and length of credible intervals. (a) Coverage probability for correctly specified $\bar{Q}_0$. (b) Length (multiplied by n) for correctly specified $\bar{Q}_0$. (c) Coverage probability for misspecified $\bar{Q}_0$. (d) Length (multiplied by n) for misspecified $\bar{Q}_0$. (e) Coverage probability of the corrected intervals. (f) Length (multiplied by n) of the corrected intervals

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. The mean of the targeted posterior distribution appears to be more biased than the TMLE, especially if $\bar{Q}_0$ is misspecified and a uniform distribution is used as prior for ψ_0. However, all the estimators appear to be asymptotically unbiased.

Figure 28.3 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 correct coverage, credible intervals with misspecified $\bar{Q}_0$ are somewhat conservative for some small sample sizes, having wider lengths and a coverage probability that is barely greater than the prespecified level 95%. This means that the variance of the posterior distribution is larger if $\bar{Q}_0$ is misspecified, thereby reflecting some kind of "inefficiency" of the posterior distribution due to misspecification of $\bar{Q}_0$. The correction to the credible intervals proposed in (28.13) operates, causing 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. The inclusion of an unbiased prior with small variance results in a significant reduction in the length of the credible intervals, especially for small sample sizes.

28.5 Discussion

A methodology to carry out targeted inference for an 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 an 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 TMLE. 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 correctly specified. The frequentist can use the posterior mean as an estimator of the target parameter, thereby allowing the incorporation of a prior distribution, while proceeding with frequentist statistical inference.

The asymptotic variance of the targeted posterior distribution has been proven to be equal to the variance under the true distribution P_0 of the efficient influence curve at P_0 when the initial estimator of P_0 is consistent. This implies, amongst other characteristics, that credible intervals will also be frequentist confidence intervals for the target parameter 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 simple adjustment has been provided that provides the desired coverage under the assumption that the TMLE is asymptotically linear.

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 (notably for binary outcome Y), the uncorrected credible intervals based on misspecified $\bar{Q}_0$ also had the right asymptotic coverage probability, suggesting the possibility that for some cases, even if P_n^0 is not consistent, $(1-\alpha)$ credible intervals still have $(1-\alpha)$ coverage probability. Identification of those cases is an interesting issue for future work. The simulation also showed that the credible intervals for a misspecified $\bar{Q}_0$ 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 distribution for any pathwise differentiable parameter defined on any semiparametric model. Our formal asymptotic results for the targeted posterior distribution can be straightforwardly generalized. The targeted posterior distribution is defined by the TMLE, specifically, the loss function and parametric submodel that defined the TMLE, and a prior distribution on the target parameter. 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 formal theoretical studies that will provide a comprehensive understanding of the frequentist properties of the targeted posterior distribution.

Appendix 1

Proof (Theorem 28.1). Let $u_n^*(\psi)(O_i) \equiv P_n^*(f_n^{*-1}(\psi))(O_i)$, where $f_n^* : \epsilon \to \Psi(P_n^*(\epsilon))$. Let $u^*(\psi)(O_i) \equiv P^*(f_n^{*-1}(\psi))(O_i)$, for $f^* : \epsilon \to \Psi(P^*(\epsilon))$. Let $\tilde{\psi}_n$ be a random variable with distribution given by the targeted posterior distribution of ψ_0:

$$\pi(\psi) \prod_{i=1}^{n} u_n^*(\psi)(O_i),$$

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

$$p_{T_n}(t) \propto \pi\left(\psi_n + \frac{t}{\sqrt{n}}\right) \prod_{i=1}^{n} u_n^*\left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i).$$

We have that, for some positive constant c_n independent of ψ_n and t,

$$\log p_{T_n}(t) = \log c_n + \log \pi\left(\psi_n + \frac{t}{\sqrt{n}}\right) + \sum_{i=1}^{n} \log u_n^*\left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i).$$

The first two terms behave for n large as a constant function in t. We will now study the last term as a function in t for n large. A Taylor series expansion in t around zero yields the following asymptotic approximation for the last term:

$$\sum_{i=1}^{n} \log u_n^* \left(\psi_n + \frac{t}{\sqrt{n}}\right)(O_i) = \\ R_n^1 + R_n^2 + \frac{t^2}{2n} \sum_{i=1}^{n} \frac{d^2}{d\psi^2} \log u_n^*(\psi)(O_i)\Big|_{\psi=\psi_n} + R_n^3, \quad (28.14)$$

where

- $R_n^1 = \sum_{i=1}^n \log u_n^*(\psi_n)(O_i) = \sum_{i=1}^n \log P_n^*(O_i)$ does not depend on t nor on ψ_n and satisfies $\log c_n + R_n^1 = \log c'_n$ for some constant c'_n independent of t and ψ_n;
- $R_n^2 = S(\psi_n) = 0$ because

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

  and $\epsilon_n = 0$ is the maximum likelihood estimator of ϵ in the model $\{P_n^*(\epsilon) : \epsilon\}$;
- The remainder R_n^3 can be written as

$$\frac{t^3}{6} \frac{1}{n^{3/2}} \sum_{i=1}^{n} \frac{d^3}{d\psi^3} \log u_n^*(\psi)(O_i)\Big|_{\psi=\psi_1},$$

  for some ψ_1 between zero and ψ_n, and is thus of order $n^{-\frac{1}{2}}$. This shows that R_n^3 is negligible compared with the other term in (28.14) which is of order 1.

Recall the definition of $h(\psi, P)$ in the theorem, and note that

$$\frac{t^2}{2n} \sum_{i=1}^{n} \frac{d^2}{d\psi^2} \log u_n^*(\psi)(O_i)\Big|_{\psi=\psi_n} = \frac{t^2}{2} P_n h(\psi_n, P_n^*).$$

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 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

$$P_n \frac{d^2}{d\psi^2} \log u_n^*(\psi)\Big|_{\psi=\psi_n} \longrightarrow P_0 \frac{d^2}{d\psi^2} \log u^*(\psi)\Big|_{\psi=\psi_0},$$

which in turn proves that $p_{T_n}(t)$ converges, up to a constant, to

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

where

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

This asymptotic variance satisfies:

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

where $\frac{d}{d\epsilon} \log P^*(\epsilon)\big|_{\epsilon=0} = 0$ because $P_n D^*(P_n^*) = 0$ (a property of the frequentist TMLE), and because of the convergence of P_n and P_n^* to P_0 and P^*, respectively. 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 $\sigma^2(P^*)(A, W) = Var_{P^*}(Y \mid A, W)$ 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 at P^* and $s(P^*)$ is the score of $P^*(\epsilon)$ at $\epsilon = 0$, which is precisely $D^*(P^*)$. Therefore:

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

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 $\bar{Q}$ is fluctuated. If the outcome is continuous, we can consider a linear fluctuation model:

$$\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon H_1^*, \tag{28.15}$$

where H_1^* is defined in (28.6). In this case, the mapping $\Psi(P_n^0(\epsilon))$ can be written as

$$\begin{aligned}\Psi(P_n^0(\epsilon)) =& P_n\left(\bar{Q}_{n,1} - \bar{Q}_{n,0}\right) + \epsilon P_n\left(H_{1,1}^* - H_{1,0}^*\right) \\ =& \Psi\left(P_n^0\right) + \epsilon P_n\left(H_{1,1}^* - H_{1,0}^*\right),\end{aligned} \tag{28.16}$$

where $Q_{n,A}(W) \equiv \bar{Q}_n^0(A, W)$ and $H_{1,A}^*(W) \equiv H_1^*(A, W)$. 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 on ψ_0, the prior distribution on ϵ is characterized by

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

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

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

Let us consider $Q_{Y,n}(\epsilon)(Y \mid A, W)$ 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 let $\sigma_{\bar{Q}_n^0}^2 \equiv \sigma^2(\bar{Q}_n^0)$. The part of the likelihood corresponding to $Q_{Y,n}(\epsilon)(Y \mid A, W)$ can be written as follows:

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

Thus, the posterior density of ϵ is

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

Now let

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

Then,

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

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

Note that the maximum likelihood estimator of ϵ in the model (28.15), 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 ϵ_n and the prior mean μ_ϵ 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

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

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

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

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

where

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

Note the posterior mean of ψ_0 is just a weighted average of the TMLE 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.

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, we refer readers 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 <- numeric(n)
  e[1] <- e0
  z <- rnorm(n-1)
  for(i in 2:n){
      cand <- z[i-1]*sd0 + e[i-1]
      p <- (posterior(cand) * dnorm(e[i-1], mean = cand,
              sd = sd0))/(posterior(e[i-1]) * dnorm(cand,
              mean = e[i-1], sd = sd0))
      pr <- min(p, 1)
      e[i] <- sample(c(cand, e[i-1]), 1, prob = c(pr, 1-pr))
  }
  return(e[-1])}
```