Chapter 8
Inference and Non-regularity

Inference plays a key role in almost all statistical problems. In the context of DTRs, one can think of inference for mainly two types of quantities: (i) inference for the parameters indexing the optimal regime; and (ii) inference for the value function (mean outcome) of a regime – either a regime that was pre-specified, or one that was estimated. The literature contains several instances of estimation and inference for the value functions of one or more pre-specified regimes (Lunceford et al. 2002; Wahed and Tsiatis 2004, 2006; Thall et al. 2000, 2002, 2007a). However there has been relatively little work on inference for the value function of an estimated policy, mainly due to the difficulty of the problem.

Constructing confidence intervals (CIs) for the parameters indexing the optimal regime is important for the following reasons. First, if the CIs for some of these parameters contain zero, then perhaps the corresponding components of the patient history need not be collected to make optimal decisions using the estimated DTR. This has the potential to reduce the cost of data collection in a future implementation of the estimated optimal DTR. Thus in the present context, CIs can be viewed as a tool – albeit one that is not very sophisticated – for doing variable selection. Such CIs can be useful in exploratory data analysis when trying to interactively find a suitable model for, say, the Q-functions. Second, note that when linear models are used for the Q-functions, the difference in predicted mean outcomes corresponding to two treatments, e.g. a contrast of Q-functions or a blip function, becomes a linear combination of the parameters indexing the optimal regime. Point-wise CIs for these linear combinations can be constructed over a range of values of the history variables based on the CIs for individual parameters. These CIs can dictate when there is insufficient support in the data to recommend one treatment over another; in such cases treatment decisions can be made based on other considerations, e.g. cost, familiarity, burden, preference etc.

An additional complication in inference for the parameters indexing the optimal regime arises because of a phenomenon called *non-regularity*. It was Robins (2004) who first considered the problem of inference for the parameters of the optimal DTR in the context of G-estimation. As originally discussed by Robins, the treatment effect parameters at any stage prior to the last can be *non-regular* under

B. Chakraborty and E.E.M. Moodie, *Statistical Methods for Dynamic Treatment Regimes*, Statistics for Biology and Health 76, DOI 10.1007/978-1-4614-7428-9_8,

certain longitudinal distributions of the data. By non-regularity, we mean that the asymptotic distribution of the estimator of the treatment effect parameter does not converge uniformly over the parameter space; see below for further details. This technical phenomenon of non-regularity has considerable practical consequences; it often causes bias in estimation, and leads to poor frequentist properties of Wald-type or other standard confidence intervals. Any inference technique that aims to provide good frequentist properties such as appropriate Type I error and nominal coverage of confidence intervals has to address the problem of non-regularity. In this chapter, we consider various approaches to inference in the context of Q-learning and G-estimation.

8.1 Inference for the Parameters Indexing the Optimal Regime Under Regularity

All of the recursive methods of estimation considered in previous chapters, including Q-learning and G-estimation, can be viewed as two-step (substitution) estimators at each stage. At stage j, the first step of estimation requires finding the effect of treatment at all future stages, and then substituting these into the stage j estimating equation in order to find the estimator of ψ_j. For example, in the Q-learning context, for a two-stage example, the pseudo-outcome at the first stage equals $\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2^{opt}(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2)$ which relies on estimators of β_2 and ψ_2. Similarly, in the recursive implementation of G-estimation, the stage-1 estimating function includes $G_{\text{mod},1}(\psi_1) = Y + \left[\gamma_1(h_1, d_1^{opt}; \psi_1) - \gamma_1(h_1, a_1; \psi_1)\right] + \left[\gamma_2(h_2, d_2^{opt}; \hat{\psi}_2) - \gamma_2(h_2, a_2; \hat{\psi}_2)\right]$, which requires estimators of ψ_2 (as well as estimators of propensity score model parameters). At each stage, the decision rule parameters for that particular stage are treated as parameters of interest, and any other parameters including those for treatment models or subsequent treatment stages are considered nuisance parameters.

Newey and McFadden (1994) provide a discussion of the impact of the first-step estimation on the standard errors of the second-step estimates. Van der Laan and Robins (2003) also discuss the issue of second-step estimates' standard errors, arriving at the same standard error as Newey and McFadden found by a different, more measure-theoretic approach. We briefly review the theory of variance derivations for estimating equations, and apply these methods to Q-learning and G-estimation. Throughout this section, we consider only regular estimators, which in the DTR context implies that there is a unique optimal treatment for each possible treatment and covariate history at each stage. We will then consider the more challenging problem of non-regular estimators in Sect. 8.2 and subsequent sections.

8.1.1 A Brief Review of Variances for Estimating Equations

In this section, we provide a concise overview of the theory of estimating equations, since most methods of estimation discussed in this book are *M-estimators*, i.e. estimators which can be obtained as the minima of sums of functions of the data or are roots of an estimating function. In particular, as implemented in previous chapters, Q-learning and G-estimation are both M-estimators. This development will enable us to derive and discuss measures of variability and confidence of the estimators of decision rule parameters more precisely.

A function of the parameter and data, $U_n(\theta) = U_n(\theta, Y) = \mathbb{P}_n U(\theta, Y_i)$, which is of the same dimensionality as the parameter θ for which $E[U_n(\theta)] = 0$ is considered. $U_n(\theta)$ is said to be an *estimating function* (EF), and $\hat{\theta}$ is an *EF estimator* if it is a solution to the *estimating equation* $U_n(\theta) = 0$. That is, $\hat{\theta}$ is an EF estimator if $U_n(\hat{\theta}) = 0$. Note that the EF $U_n(\theta, Y)$ is itself a random variable, since it is a function of the random variable Y. To perform inference, we derive the frequency properties of the EF and can then transfer these properties to the resultant estimator with the help of a Taylor approximation and the delta method. Excellent resources on asymptotic theory of statistics are given by Van der Vaart (1998) and Ferguson (1996); or for a particular focus on semi-parametric methods, see Bickel et al. (1993) and Tsiatis (2006).

The corresponding estimating equation that defines the estimator $\hat{\theta}$ has the form

$$U_n(\hat{\theta}) = U_n(\hat{\theta}, Y) = \mathbb{P}_n U(\hat{\theta}, Y_i) = 0. \tag{8.1}$$

The estimating Eq. (8.1) is said to be unbiased if $E[U_n(\theta)] = 0$, and so

$$\begin{aligned} Var[U_n(\theta)] &= E\left[(U_n(\theta) - E[U_n(\theta)]) (U_n(\theta) - E[U_n(\theta)])^T \right] \\ &= E[U_n(\theta) U_n(\theta)^T], \end{aligned}$$

which converges to some matrix Σ_U. Further, $U_n(\theta)$ is a sum of conditionally independent terms, so under standard regularity conditions

$$U_n(\theta) \to_d \mathcal{N}(0, \Sigma_U). \tag{8.2}$$

Using a first order Taylor expansion, we find

$$0 = U_n(\hat{\theta}_n) = U_n(\theta) + \left(\frac{\partial U_n(\theta)}{\partial \theta} \right) (\hat{\theta} - \theta) + o_p(1).$$

This gives that $(\hat{\theta} - \theta) = -\left(\frac{\partial U_n(\theta)}{\partial \theta} \right)^{-1} U_n(\theta) + o_p(1)$. From this, we can deduce that $\hat{\theta} \to_p \theta$ and

$$\sqrt{n}\,(\hat{\theta} - \theta) \to_d \mathrm{N}_p(0, A^{-1} \Sigma_U (A^T)^{-1}) \tag{8.3}$$

where $A = -E\left[\frac{\partial}{\partial\theta}U_n(\theta)\right]$. That is, $\hat{\theta}$ is a consistent and asymptotically normally distributed estimator. The form of the variance in expression (8.3) has led to it being called the *sandwich estimator*, where A forms the "bread" and Σ_U is the "filling" of the sandwich.

It follows from Eq. (8.3) that $\sqrt{n}\Sigma_{\hat{\theta}}^{-1/2}(\hat{\theta}-\theta) \rightarrow_d \mathrm{N}_p(0,\mathbf{I}_p)$ where $\Sigma_{\hat{\theta}} = A^{-1}\Sigma_U(A^T)^{-1}$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix, implying that confidence intervals can be constructed, and significance tests performed, using a Wald statistic of the form $\sqrt{n}\Sigma_{\hat{\theta}}^{-1/2}\hat{\theta}$. In a more familiar form, this would give, for example, a 95 % confidence interval of $\hat{\theta} \pm 1.96SE(\hat{\theta})$ for a scalar-valued parameter θ (for p-dimensional parameter θ, one can similarly construct component-wise CIs) and a test statistic $W = \sqrt{n}\Sigma_{\hat{\theta}}^{-1/2}\hat{\theta}$.

Confidence intervals for θ can also be constructed directly using the EF and its standard error. From Eq. (8.2), we have $\Sigma_U^{-1/2}U_n(\theta) \rightarrow_d \mathcal{N}(0,\mathbf{I}_p)$. It is therefore the case that we can construct a *score* or *Rao* interval by searching for values θ that satisfy $|\Sigma_U^{-1/2}U_n(\theta)| \leq 1.96$. Unlike the Wald intervals, which rely only on the value of the estimated parameter and its standard error, score-based intervals may be more computationally burdensome as they may require a search over the space of θ. However, score-based intervals may exhibit better finite sample properties even when standard regularity conditions do not hold, since these intervals do not require derivatives of the EF (Robins 2004; Moodie and Richardson 2010).

Now suppose that θ is vector-valued and can be partitioned such that $\theta = (\psi^T, \beta^T)^T$ where ψ is of interest, and β contains nuisance parameters (such as, for example, parameters associated with predictive variables in Q-learning, or parameters from a propensity score model in G-estimation). If interest lies in performing significance tests about ψ leaving β unspecified, i.e. in testing null hypotheses of the form

$$\mathcal{H}_0 : \psi = \psi_0, \beta = \text{'anything'}$$

versus the alternative hypothesis

$$\mathcal{H}_A : (\psi, \beta) \neq (\psi_0, \beta)$$

then we have what is called a *composite* null hypothesis. Suppose further that the EF $U_n(\theta)$ can be decomposed into $U_n(\theta) = \begin{pmatrix} U_n(\psi) \\ U_n(\beta) \end{pmatrix}$.

To derive the correct variance for the composite null hypothesis, consider a Taylor expansion of $U_n(\psi)$ about the limiting value, β, of a consistent estimator, $\hat{\beta}$, of the nuisance parameter β:

$$U_{\text{adj}}(\psi) = U_n(\psi) + \left[\frac{\partial}{\partial\beta}U_n(\psi)\right](\hat{\beta}-\beta) + o_p(1) \tag{8.4}$$

$$= U_n(\psi) - \left[\frac{\partial}{\partial\beta}U_n(\psi)\right]\left[\frac{\partial}{\partial\beta}U_n(\beta)\right]^{-1}U_n(\beta) + o_p(1). \tag{8.5}$$

From Eq. (8.4), it can be seen that $E[U_n(\psi)] = E[U_{\text{adj}}(\psi)]$ so $U_{\text{adj}}(\psi)$ is an unbiased EF; Eq. (8.5) follows from Eq. (8.4) via a substitution from a Taylor expansion of the EF for β about its limiting value. From Eq. (8.5), we can derive the asymptotic distribution of the parameter of interest ψ to be

$$\sqrt{n}\,(\hat{\psi} - \psi) \rightarrow_d \mathrm{N}_p(0, \Sigma_{\hat{\psi}})$$

where $\Sigma_{\hat{\psi}} = A_{\text{adj}}^{-1}\Sigma_{U_{\text{adj}}}(A_{\text{adj}}^T)^{-1}$ is the asymptotic variance of $\hat{\psi}$ with A_{adj} the probability limit of $-E\left[\frac{\partial}{\partial\psi}U_{\text{adj}}(\psi)\right]$ and $\Sigma_{U_{\text{adj}}}$ is the probability limit of $E\left[U_{\text{adj}}(\psi)U_{\text{adj}}(\psi)^T\right]$. Note that $\hat{\psi}$ is the substitution estimator defined by finding the solution to the EF where an estimate of the (vector) nuisance parameter $\hat{\beta}$ has been plugged into the equation in place of the true value, β.

It is interesting to consider the variance of the substitution estimator $\hat{\psi}$ with the estimator, say $\tilde{\psi}$, that would result from plugging in the true value of the nuisance parameter (a feasible estimator only when such true values are known). That is, we may wish to consider $\Sigma_{\hat{\psi}}$ and $\Sigma_{\tilde{\psi}}$. It turns out that no general statement regarding the two estimators' variances can be made, however there are special cases in which relationships can be derived (see Henmi and Eguchi (2004) for a geometric consideration of EF which serves to elucidate the variance relationships). For example, if the EF is the score function for θ in a parametric model, there is a cost (in terms of information loss or variance inflation) that is incurred for having to estimate the nuisance parameters. In contrast, in the semi-parametric setting where the score functions for ψ and β are orthogonal and that the score function is used as the EF for β, it can be shown that $\Sigma_{\tilde{\psi}} - \Sigma_{\hat{\psi}}$ is positive definite. That is, efficiency is *gained* by estimating rather than knowing the nuisance parameter β.

8.1.2 Asymptotic Variance for Q-learning Estimators

We now apply the theory of the previous section to Q-learning for the case where we use linear models parameterized by $\theta_j = (\psi_j, \beta_j)$ of the form $Q_j^{opt}(H_j, A_j; \beta_j, \psi_j) = \beta_j^T H_{j0} + (\psi_j^T H_{j1})A_j$. For simplicity of exposition, we will focus on the two-stage setting, but extensions to the general, K-stage setting follow directly. Following the algorithm for Q-learning outlined in Sect. 3.4.1, we begin with a regression of Y_2 using the model $Q_2^{opt}(H_2, A_2; \beta_2, \psi_2) = \beta_2^T H_{20} + (\psi_2^T H_{21})A_2$. Letting X_2 denote $(H_{20}, H_{21}A_2)$, this gives a linear regression of the familiar form $E[Y_2|X_2] = X_2\theta_2$, with $Var[\hat{\theta}_2] = (X_2^T X_2)^{-1}\sigma^2$ where σ^2 denotes the variance of the residuals $Y_2 - X_2\theta_2$. Confidence intervals can then be formed, and significance tests performed, for the vector parameter θ_2. If composite tests of the form $\mathscr{H}_0 : \psi_2 = 0$ are desired, hypothesizing that the variables contained in H_{21} are not significantly useful tailoring variables without specifying any hypothesized values for the value of β_2, then the Wald statistic should be scaled using $\mathscr{I}_{\psi_2\psi_2.\beta_2}^{1/2} = (\mathscr{I}_{\psi_2\psi_2} - \mathscr{I}_{\psi_2\beta_2}\mathscr{I}_{\beta_2\beta_2}^{-1}\mathscr{I}_{\beta_2\psi_2})^{1/2}$, where

$$\begin{pmatrix} \mathscr{I}_{\psi_2\psi_2} & \mathscr{I}_{\psi_2\beta_2} \\ \mathscr{I}_{\beta_2\psi_2} & \mathscr{I}_{\beta_2\beta_2} \end{pmatrix}$$

is a block-diagonal matrix decomposition of the information of the regression parameters at the second stage, and similarly $\mathscr{I}^{1/2}_{\psi_2\psi_2\cdot\beta_2}$ should be used to determine the limits of a confidence interval.

Now, let us consider the first-stage estimator. First stage estimation proceeds by first forming the pseudo-outcome $Y_1 + \beta_2^T H_{20} + |\psi_2^T H_{21}|$, which we implement in practice using the estimate $\hat{Y}_1 = Y_1 + \hat{\beta}_2^T H_{20} + |\hat{\psi}_2^T H_{21}|$, and regressing this on $(H_{10}, H_{11}A_1)$ using the model $Q_1^{opt}(H_1, A_1; \beta_1, \psi_1) = \beta_1^T H_{10} + (\psi_1^T H_{11})A_1$. This two-stage regression-based estimation can be viewed as an estimating equation based procedure as follows. Define

$$\begin{aligned} U_{2,n}(\theta_2) &= \mathbb{P}_n\Big(Y_2 - Q_2^{opt}(H_2, A_2; \beta_2, \psi_2)\Big)\Big(\frac{\partial}{\partial \theta_2} Q_2^{opt}(H_2, A_2; \beta_2, \psi_2)\Big) \\ &= \mathbb{P}_n\Big(Y_2 - \beta_2^T H_{20} - (\psi_2^T H_{21})A_2\Big)(H_{20}^T, H_{21}^T A_2)^T, \\ U_{1,n}(\theta_1, \theta_2) &= \mathbb{P}_n\Big(Y_1 + \max_{A_2} Q_2^{opt}(H_2, A_2; \beta_2, \psi_2) - \\ &\qquad Q_1^{opt}(H_1, A_1; \beta_1, \psi_1)\Big)\Big(\frac{\partial}{\partial \theta_1} Q_1^{opt}(H_1, A_1; \beta_1, \psi_1)\Big) \\ &= \mathbb{P}_n\Big(Y_1 + \beta_2^T H_{20} + |\psi_2^T H_{21}| - \beta_1^T H_{10} - (\psi_1^T H_{11})A_1\Big)(H_{10}^T, H_{11}^T A_1)^T. \end{aligned}$$

Then the (joint) estimating equation for all the parameters from both stages of Q-learning is given by

$$\begin{pmatrix} U_{2,n}(\theta_2) \\ U_{1,n}(\theta_1, \theta_2) \end{pmatrix} = 0.$$

At the first stage, then, both the main effect parameters β_1 and all second stage parameters can be considered nuisance parameters. Collecting these into a single vector $\beta^\sharp = (\beta_1, \beta_2, \psi_2)$, we use a similar form to above, forming Wald test statistics or CIs for the tailoring variable parameters using $\mathscr{I}^{1/2}_{\psi_1\psi_1\cdot\beta_1^\sharp} = (\mathscr{I}_{\psi_1\psi_1} - \mathscr{I}_{\psi_1\beta_1^\sharp}\mathscr{I}^{-1}_{\beta_1^\sharp\beta_1^\sharp}\mathscr{I}_{\beta_1^\sharp\psi_1})^{1/2}$, where

$$\begin{pmatrix} \mathscr{I}_{\psi_1\psi_1} & \mathscr{I}_{\psi_1\beta_1^\sharp} \\ \mathscr{I}_{\beta_1^\sharp\psi_1} & \mathscr{I}_{\beta_1^\sharp\beta_1^\sharp} \end{pmatrix}$$

is a block-diagonal matrix decomposition of the inverse-variance of all parameters.

8.1.3 Asymptotic Variance for G-estimators

The variance of the optimal decision rule parameters $\hat{\psi}$ must adjust for the plug-in estimates of nuisance parameters in the estimating function of Eq. (4.3), $U(\psi) = \sum_{i=1}^{n}\sum_{j=1}^{K} U_j(\psi_j, \hat{\varsigma}_j(\psi_j), \hat{\alpha}_j)$. In the derivations that follow, we assume the parameters are not shared between stages, however the calculations are similar in the shared-parameter setting. Second derivatives of the estimating functions for all parameters are needed, and thus we require that each subject's optimal regime must be unique at every stage except possibly the first. If for any individual, the optimal treatment is not unique, then it is the case that $\gamma_j(h_j, a_j) = 0$, or equivalently that for a Q-function $\beta_j^T H_{j0} + (\psi_j^T H_{j1})(A_j + 1)/2$, $\psi_j^T H_{j1} = 0$. Provided the rule is unique, then the estimating functions used in each stage of estimation for G-estimation will be differentiable and so the asymptotic variance can be determined.

Robins (2004) derives the variance of $U(\psi, \varsigma(\psi), \alpha)$ by performing a first order Taylor expansion of the function about the limiting values of $\hat{\varsigma}(\psi)$ and $\hat{\alpha}$, ς and α:

$$U_{\text{adj}}(\psi) = U(\psi, \varsigma, \alpha) + E\left[\frac{\partial}{\partial \varsigma} U(\psi, \varsigma, \alpha)\right](\hat{\varsigma}(\psi) - \varsigma) + E\left[\frac{\partial}{\partial \alpha} U(\psi, \varsigma, \alpha)\right](\hat{\alpha} - \alpha) + o_p(1)$$

to obtain the adjusted G-estimating function, $U_{\text{adj}}(\psi)$, which estimates the parameters from all stages $j = 1, \ldots, K$ simultaneously. Of course, with the limiting values of the nuisance parameters unknown, this expression does not provide a practical EF. If $\dot{l}_\alpha$ and $\dot{l}_\varsigma$ denote the (score) EF for the treatment model and expected counterfactual model nuisance parameters, respectively, then we can again apply a Taylor expansion to find

$$\hat{\alpha} - \alpha = -\left(E\left[\frac{\partial}{\partial \alpha}\dot{l}_\alpha(\alpha)\right]\right)^{-1} \dot{l}_\alpha(\alpha) + o_p(1),$$
$$\hat{\varsigma}(\psi) - \varsigma = -\left(E\left[\frac{\partial}{\partial \varsigma}\dot{l}_\varsigma(\varsigma)\right]\right)^{-1} \dot{l}_\varsigma(\varsigma) + o_p(1).$$

This gives

$$U_{\text{adj}}(\psi) = U(\psi, \varsigma, \alpha) - E\left[\frac{\partial}{\partial \varsigma} U(\psi, \varsigma, \alpha)\right]\left(E\left[\frac{\partial}{\partial \varsigma}\dot{l}_\varsigma(\varsigma)\right]\right)^{-1} \dot{l}_\varsigma(\varsigma) - E\left[\frac{\partial}{\partial \alpha} U(\psi, \varsigma, \alpha)\right] E\left[\frac{\partial}{\partial \alpha}\dot{l}_\alpha(\alpha)\right]^{-1} \dot{l}_\alpha(\alpha).$$

Thus the estimating function has variance $E[U_{\text{adj}}(\psi)^{\otimes 2}] = E[U_{\text{adj}}(\psi) U_{\text{adj}}(\psi)^T]$. It follows that the variance of the blip function parameters which index the decision rules, $\hat{\psi} = (\hat{\psi}_1^T, \hat{\psi}_2^T, \ldots, \hat{\psi}_K^T)^T$, is given by

$$\Sigma_{\hat{\psi}} = E\left[\left\{\left(E\left[\frac{\partial}{\partial\psi}U_{\text{adj}}(\psi,\varsigma,\alpha)\right]\right)^{-1}U_{\text{adj}}(\psi,\varsigma,\alpha)\right\}^{\otimes 2}\right].$$

Suppose at each of two stages, p different parameters are estimated. Then $\Sigma_{\hat{\psi}}$ is the $(2p)\times(2p)$ covariance matrix

$$\Sigma_{\hat{\psi}} = \begin{pmatrix} \Sigma_{\hat{\psi}}^{(11)} & \Sigma_{\hat{\psi}}^{(12)} \\ \Sigma_{\hat{\psi}}^{(21)} & \Sigma_{\hat{\psi}}^{(22)} \end{pmatrix}.$$

The $p\times p$ covariance matrix of $\hat{\psi}_2 = (\hat{\psi}_{20},\ldots,\hat{\psi}_{2(p-1)})$ that accounts for using the substitution estimates $\hat{\varsigma}_2$ and $\hat{\alpha}_2$ is $\Sigma_{\hat{\psi}}^{(22)}$, and accounting for substituting $\hat{\psi}_2$ as well as $\hat{\varsigma}_1$ and $\hat{\alpha}_1$ to estimate ψ_1 gives the $p\times p$ covariance matrix $\Sigma_{\hat{\psi}}^{(11)}$ for $\hat{\psi}_1 = (\hat{\psi}_{10},\ldots,\hat{\psi}_{1(p-1)})$.

However, as shown in Sect. 4.3.1, parameters can be estimated separately at each stage using G-estimation recursively at each stage. In such a case, it is possible to estimate the variances $\Sigma_{\hat{\psi}}^{(22)}$ and $\Sigma_{\hat{\psi}}^{(11)}$ of the stage-specific parameters recursively as well (Moodie 2009a). The development for the estimation of the diagonal components, $\Sigma_{\hat{\psi}}^{(jj)}$, of the covariance matrix $\Sigma_{\hat{\psi}}$ will be undertaken in a two-stage setting, but the extension to the K stage case follows directly.

Let $U_{\text{adj},1}(\psi_1,\psi_2)$ and $U_{\text{adj},2}(\psi_2)$ denote, respectively, the first and second components of $U_{\text{adj}}(\psi)$. At the second stage, use $U_{\text{adj},2}$ to calculate $\Sigma_{\hat{\psi}}^{(22)}$. To find the covariance matrix of $\hat{\psi}_1$, use a Taylor expansion of $U_1(\psi_1,\hat{\psi}_2,\hat{\varsigma}_1(\psi_1),\hat{\alpha}_1)$ about the limiting values of the nuisance parameters $(\psi_2,\varsigma_1,\alpha_1)$. After some simplification, this gives:

$$\begin{aligned} U_{\text{adj},1}^{\varepsilon}(\psi_1,\psi_2) = U_{\text{adj},1}(\psi_1,\psi_2) - E\left[\frac{\partial}{\partial\psi_2}U_1(\psi_1,\psi_2,\varsigma_1,\alpha_1)\right]\cdot \\ \left(E\left[\frac{\partial}{\partial\psi_2}U_{\text{adj},2}(\psi_2,\varsigma_2,\alpha_2)\right]\right)^{-1}U_{\text{adj},2}(\psi_2,\varsigma_2,\alpha_2) \\ +o_p(1). \end{aligned}$$

It then follows that $\sqrt{n}(\hat{\psi}_1-\psi_1)$ converges in distribution to

$$N\left(0, E\left[\left\{\left(E\left[\frac{\partial}{\partial\psi_1}U_{\text{adj},1}^{\varepsilon}\right]\right)^{-1}U_{\text{adj},1}^{\varepsilon}\right\}^{\otimes 2}\right]\right).$$

Thus, the diagonal components of $\Sigma_{\hat{\psi}}$ are obtained using a more tractable calculation.

Note that if there are $K>2$ stages, the similar derivations can be used, but require the use of $K-j$ adjustment terms to $U_{\text{adj},j}$ for the estimation and substitution of all future decision rule parameters, $\psi_{j+1},\ldots,\psi_K$. Note that $U_{\text{adj}}^{\varepsilon}$ and U_{adj}

produce numerically the same variance estimate at each stage: that is, the recursive variance calculation simply provides a more convenient and less computationally intensive approach by taking advantage of known independences (i.e. zeros in the matrix of derivatives of $U(\psi)$ with respect to ψ) which arise because decision rules do not share parameters at different stages. The asymptotic variances can lead to coverage below the nominal level in small samples, but perform well for samples of size 1,000 or greater in regular settings where differentiability of the EFs holds (Moodie 2009a).

8.1.4 Projection Confidence Intervals

Berger and Boos (1994) and Berger (1996) proposed a general method for constructing valid hypothesis tests in the presence of a nuisance parameter. One can develop an asymptotically exact confidence interval for the stage 1 parameter ψ_1 by inverting these hypothesis tests, based on the following nuisance parameter formulation. As we have noted above, many DTR parameter estimators are obtained via substitution because the true value of the stage 2 parameter ψ_2 is unknown and must be estimated (see Sect. 8.2 for details). Instead, if the true value of ψ_2 were known a priori, the asymptotic distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ would be *regular* (in fact, normal), and standard procedures could be used to construct an asymptotically valid confidence interval although performance of such asymptotic variance estimators may be poor in small samples. Thus, while ψ_2 is not of primary interest for analyzing stage 1 decisions, it nevertheless plays an essential role in the asymptotic distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$. In this sense, ψ_2 is a nuisance parameter. This idea was used by Robins (2004) to construct a projection confidence interval for ψ_1.

The basic idea is as follows. Let $\mathscr{S}_{n,1-\alpha}(\psi_2)$ denote an asymptotically exact confidence interval for ψ_1 if ψ_2 were known, i.e., $P(\psi_1 \in \mathscr{S}_{n,1-\alpha}(\psi_2)) = 1-\alpha+o_P(1)$. Of course, the exact value of ψ_2 is not known, but since $\sqrt{n}(\hat{\psi}_2 - \psi_2)$ is regular and asymptotically normal, it is straightforward to construct a $(1-\varepsilon)$ asymptotic confidence interval for ψ_2, say $\mathscr{C}_{n,1-\varepsilon}$, for arbitrary $\varepsilon > 0$. Then, it follows that $\bigcup_{\gamma \in \mathscr{C}_{n,1-\varepsilon}} \mathscr{S}_{n,1-\alpha}(\gamma)$ is a $(1-\alpha-\varepsilon)$ confidence interval for ψ_1. To see this, note that

$$P\big(\psi_1 \in \bigcup_{\gamma \in \mathscr{C}_{n,1-\varepsilon}} \mathscr{S}_{n,1-\alpha}(\gamma)\big) \geq 1-\alpha+o_P(1)+P\big(\psi_2 \notin \mathscr{C}_{n,1-\varepsilon}\big) = 1-\alpha-\varepsilon+o_P(1). \tag{8.6}$$

Thus, the projection confidence interval is the union of the confidence intervals $\mathscr{S}_{n,1-\alpha}(\gamma)$ over all values $\gamma \in \mathscr{C}_{n,1-\varepsilon}$, and is an asymptotically valid $(1-\alpha-\varepsilon)$ confidence interval for ψ_1. The main downside of this approach is that it is potentially highly conservative. Also, its implementation can be computationally highly expensive.

8.2 Exceptional Laws and Non-regularity of the Parameters Indexing the Optimal Regime

The cumulative distribution function of the observed longitudinal data is said to be *exceptional* if, at some stage j, the optimal treatment decision depends on at least one component of covariate and treatment history *and* the probability that the optimal rule is not unique is positive (Robins 2004). The combination of three factors makes a law exceptional: (i) the form of the blip or Q-function model, (ii) the true value of the blip model parameters, and (iii) the distribution of treatments and state variables. For a law to be exceptional, then, condition (i) requires the blip or Q-function model to depend on at least one covariate such as prior treatment; conditions (ii) and (iii) require that the model takes the value zero with positive probability, that is, there is some subset of the population in which the optimal treatment is not unique. Exceptional laws may commonly arise in practice: under the hypothesis of no treatment effect, for a blip or Q-function that includes at least one component of treatment and state variable history, every distribution is an exceptional law. More generally, it may be the case that a treatment is ineffective in a sub-group of the population under study. Exceptional laws give rise to non-regular estimators.

The issue of non-regularity can be better understood with a simple but instructive example discussed by Robins (2004). Consider the problem of estimating $|\mu|$ based on n i.i.d. observations $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, 1)$. Note that $|\bar{X}_n|$ is the maximum likelihood estimator of $|\mu|$, where $\bar{X}_n$ is the sample average. It can be shown that the asymptotic distribution of $\sqrt{n}(|\bar{X}_n| - |\mu|)$ for $\mu = 0$ is different from that for $\mu \neq 0$, and more importantly, the change in the distribution at $\mu = 0$ happens abruptly. Thus $|\bar{X}_n|$ is a non-regular estimator of $|\mu|$. Also, for $\mu = 0$,

$$\lim_{n\to\infty} E[\sqrt{n}(|\bar{X}_n| - |\mu|)] = \sqrt{\frac{2}{\pi}}.$$

Robins referred to this quantity as the *asymptotic bias* of the estimator $|\bar{X}_n|$. This asymptotic bias is one symptom of the underlying non-regularity, as discussed by Moodie and Richardson (2010).

We can graphically illustrate the asymptotic bias resulting from non-regularity using a class of generative models in which exceptional laws arise (see Sect. 8.8 for details). Thus there are many combinations of parameters that lead to (near-) non-regularity, and thereby bias in the parameter estimates. Hence it makes sense to study the prevalence and magnitude of bias over regions of the parameter space.

Moodie and Richardson (2010) employed a convenient way to study this bias in the context of G-estimation and the associated hard-threshold estimators using bias maps. We employ the same technique here in the Q-learning context; see Fig. 8.1. Bias maps show the absolute bias in $\hat{\psi}_{10}$ (parameter denoting main effect of treatment at stage 1) as a function of sample size n and one of the stage 2 parameters, ψ_{20}, ψ_{21}, or ψ_{22} (which are equal to the generative parameters γ_5, γ_6 and γ_7, respectively). The plots represent the average bias over 1,000 simulated data sets, computed over a range of 2 units (on a 0.1 unit grid) for each parameter at sample sizes $250, 300, \ldots, 1000$. From the bias maps, it is clear that there exist many regions

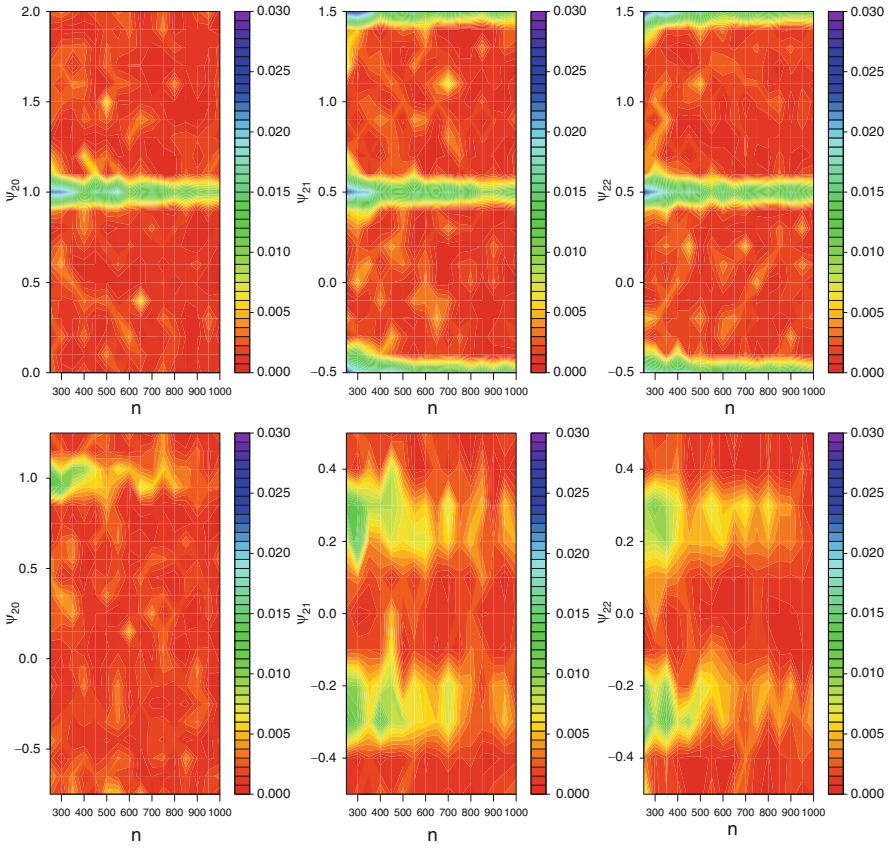

Fig. 8.1 Absolute bias of $\hat{\psi}_{10}$ in hard-max Q-learning in different regions (regular and non-regular) of the underlying parameter space. Different plots correspond to different parameter settings.

of the parameter space that lead to bias in $\hat{\psi}_{10}$, thereby reinforcing the necessity to address the problem through careful estimation and inference techniques.

As noted by Moodie and Richardson (2010), the bias maps can be used to visually represent the asymptotic results concerning DTR estimators. Consistency may be visualized by looking at a horizontal cross-section of a bias map: as sample size increases, the bias of the first-stage estimator will decrease to be smaller than any fixed, positive number at all non-regular parameter settings, even those that are nearly non-regular. However, as derived by Robins (2004), there exist sequences of data generating processes $\{\psi_{(n)}\}$ for which the second-stage parameters ψ_2 decrease with increasing n in such a way that the asymptotic bias of the first-stage estimator $\hat{\psi}_1$ is strictly positive. Contours of constant bias can be found along the lines on the bias map traced by plotting $g_2(\psi_2) = kn^{-1/2}$ against n, for some constant k. The asymptotic bias is bounded and, in finite samples, the value of the second-stage parameters (i.e. the "nearness" to non-regularity) and the sample size both determine the bias of the first-stage parameter estimator.

In many situations where the asymptotic distribution of an estimator is unavailable, bootstraping is used as an alternative approach to conduct inference. But the success of the bootstrap also hinges on the underlying smoothness of the estimator. When an estimator is non-smooth, the ordinary bootstrap procedure produces an inconsistent bootstrap estimator (Shao 1994). Inconsistency of bootstrap in the above toy example has been discussed by Andrews (2000). Poor performance of usual bootstrap CIs in the Q-learning context has been illustrated by Chakraborty et al. (2010). We first discuss non-regularity in the specific contexts of G-estimation and Q-learning, then, in the following sections, we consider several different approaches to inference that attempt to address the problem of non-regularity.

8.2.1 Non-regularity in Q-learning

With (3.8) as the model for Q-functions, the optimal DTR is given by

$$d_j^{opt}(H_j) = \arg\max_{a_j} (\psi_j^T H_{j1}) a_j = sign(\psi_j^T H_{j1}), \quad j = 1, 2, \tag{8.7}$$

where $sign(x) = 1$ if $x > 0$, and -1 otherwise. Note that the term $\beta_j^T H_{j0}$ on the right hand side of (3.8) does not feature in the optimal DTR. Thus for estimating optimal DTRs, the ψ_js are the parameters of interest, while β_js are nuisance parameters. These ψ_js are the policy parameters for which we want to construct confidence intervals.

Inference for ψ_2, the stage 2 parameters, is straightforward since this falls in the framework of standard linear regression. In contrast, inference for ψ_1, the stage 1 parameters, is complicated by the previously discussed problem of *non-regularity* resulting from the underlying non-smooth maximization operation in the estimation procedure. To further understand the problem, recall that the stage 1 pseudo-outcome in Q-learning for the i-th subject is

$$\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2^{opt}(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2) = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|, \; i = 1, \ldots, n,$$

which is a non-smooth (the absolute value function is non-differentiable at zero) function of $\hat{\psi}_2$. Since $\hat{\psi}_1$ is a function of $\hat{Y}_{1i}$, $i = 1, \ldots, n$, it is in turn a non-smooth function of $\hat{\psi}_2$. As a consequence, the distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ does not converge uniformly over the parameter space of (ψ_1, ψ_2) (Robins 2004). More specifically, the asymptotic distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ is normal if ψ_2 is such that $P[H_2 : \psi_2^T H_{21} = 0] = 0$, but is non-normal if $P[H_2 : \psi_2^T H_{21} = 0] > 0$, and this change in the asymptotic distribution happens abruptly. A precise expression for the asymptotic distribution can be found in Laber et al. (2011). The parameter ψ_1 is called a *non-regular* parameter and the estimator $\hat{\psi}_1$ a *non-regular* estimator; see Bickel et al. (1993) for a precise definition of non-regularity. Because of this non-regularity, given the noise level present in small samples, the estimator $\hat{\psi}_1$ oscillates between

the two asymptotic distributions across samples. Consequently, $\hat{\psi}_1$ becomes a biased estimator of ψ_1, and Wald type CIs for components of ψ_1 show poor coverage rates (Robins 2004; Moodie and Richardson 2010).

8.2.2 Non-regularity in G-estimation

Let us again consider a typical, two-stage scenario with linear optimal blip functions,

$$\gamma_1(h_1,a_1) = (\psi_{10}+\psi_{11}o_1)(a_1+1)/2, \text{ and}$$
$$\gamma_2(h_2,a_2) = (\psi_{20}+\psi_{21}o_2+\psi_{22}(a_1+1)/2+\psi_{23}o_2(a_1+1)/2)(a_2+1)/2.$$

Let $\eta_2 = \psi_{20}+\psi_{21}o_2+\psi_{22}(a_1+1)/2+\psi_{23}o_2(a_1+1)/2$ and similarly define $\hat{\eta}_2 = \hat{\psi}_{20}+\hat{\psi}_{21}o_2+\hat{\psi}_{22}(a_1+1)/2+\hat{\psi}_{23}o_2(a_1+1)/2$. The G-estimating function for ψ_2 is unbiased, so $E[\hat{\eta}_2] = \eta_2$. The sign of η_2 is used to decide optimal treatment at the second stage: $d_2^{opt} = sign(\eta_2) = sign(\psi_{20}+\psi_{21}o_2+\psi_{22}a_1+\psi_{23}o_2a_1)$ and $\hat{d}_2^{opt} = sign(\hat{\eta}_2)$ so that now the G-estimating equation solved for ψ_1 at the first interval contains:

$$\begin{aligned} G_{\text{mod},1}(\psi_1) &= Y-\gamma_1(o_1,a_1;\psi_1)+[\gamma_2(h_2,\hat{d}_2^{opt};\hat{\psi}_2)-\gamma_2(h_2,a_2;\hat{\psi}_2)] \\ &= Y-\gamma_1(o_1,a_1;\psi_1)+[(\hat{d}_2^{opt}-a_2)(\hat{\psi}_{20}+\hat{\psi}_{21}o_2+\hat{\psi}_{22}a_1+\hat{\psi}_{23}o_2a_1)/2] \\ &= Y-\gamma_1(o_1,a_1;\psi_1)+sign(\hat{\eta}_2)\hat{\eta}_2/2-a_2\hat{\eta}_2/2 \\ &\overset{E}{\geq} Y-\gamma_1(o_1,a_1;\psi_1)+sign(\eta_2)\eta_2/2-a_2\eta_2/2 = 0, \end{aligned}$$

where $\overset{E}{\geq}$ is used to denote "greater than or equal to in expectation". The quantity $\left[\gamma_2(h_2,d_2^{opt};\psi_2)-\gamma_2(h_2,a_2;\psi_2)\right]$ in $G_{\text{mod},1}(\psi_1)$ – or more generally, the sum $\sum_{k>j}\left[\gamma_k(h_k,d_k^{opt};\psi_k)-\gamma_k(h_k,a_k;\psi_k)\right]$ in $G_{\text{mod},j}(\psi_j)$ – corresponds conceptually to $|\mu|$ in the toy example with normally-distributed random variables X_i that was introduced at the start of the section. By using a biased estimate of $sign(\eta_2)\eta_2$ in $G_{\text{mod},1}(\psi_1)$, some strictly positive value is added into the G-estimating equation for ψ_1. The estimating function no longer has expectation zero and hence is asymptotically biased.

8.3 Threshold Estimators with the Usual Bootstrap

In this section, we will present two approaches to "regularize" the non-regular estimator (also called the *hard-max* estimator because of the maximum operation used in the definition) by thresholding and/or shrinking the effect of the term involving the maximum, i.e. $|\hat{\psi}_2^T H_{21}|$, towards zero. Usual bootstrap procedures in conjunction with these regularized estimators offer considerable improvement over the original hard-max procedure, as verified in extensive simulations. While these

estimators are quite intuitive in nature, only limited theoretical results are available. We present these in the context of Q-learning, but these can equally be applied in a G-estimation setting.

8.3.1 The Hard-Threshold Estimator

The general form of the hard-threshold pseudo-outcome is

$$\hat{Y}_{1i}^{HT} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \mathbb{I}[|\hat{\psi}_2^T H_{21,i}| > \lambda_i],\ i = 1, \ldots, n, \tag{8.8}$$

where λ_i (>0) is the threshold for the i-th subject in the sample (possibly depending on the variability of the linear combination $\hat{\psi}_2^T H_{21,i}$ for that subject). One way to operationalize this is to perform a preliminary test (for each subject in the sample) of the null hypothesis $\psi_2^T H_{21,i} = 0$ ($H_{21,i}$ is considered fixed in this test), set $\hat{Y}_{1i}^{HT} = \hat{Y}_{1i}$ if the null hypothesis is rejected, and replace $|\hat{\psi}_2^T H_{21,i}|$ with the "better guess" of 0 in the case that the test fails to reject the null hypothesis. Thus the hard-threshold pseudo-outcome can be written as

$$\hat{Y}_{1i}^{HT} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \mathbb{I}\left[\frac{\sqrt{n}|\hat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}}} > z_{\alpha/2}\right] \tag{8.9}$$

for $i = 1, \ldots, n$, where $n^{-1}\hat{\Sigma}_{\hat{\psi}_2}$ is the estimated covariance matrix of $\hat{\psi}_2$. The corresponding estimator of ψ_1, denoted by $\hat{\psi}_1^{HT}$, will be referred to as the hard-threshold estimator. The hard-threshold estimator is common in many areas like variable selection in linear regression and wavelet shrinkage (Donoho and Johnstone 1994). Moodie and Richardson (2010) proposed this estimator for bias correction in the context of G-estimation, and called it the *Zeroing Instead of Plugging In* (ZIPI) estimator. In regular data-generating settings, ZIPI estimators converge to the usual recursive G-estimators and therefore are asymptotically consistent, unbiased and normally distributed. Furthermore, in any non-regular setting where there exist some individuals for whom there is a unique optimal regime, ZIPI estimators have smaller asymptotic bias than the recursive G-estimators provided parameters are not shared across stages (Moodie and Richardson 2010).

Note that $\hat{Y}_1^{HT}$ is still a non-smooth function of $\hat{\psi}_2$ and hence $\hat{\psi}_1^{HT}$ is a non-regular estimator of ψ_1. However, the problematic term $|\hat{\psi}_2^T H_{21}|$ is thresholded, and hence one might expect that the degree of non-regularity is somewhat reduced. An important issue regarding the use of this estimator is the choice of the significance level α of the preliminary test, which is an unknown tuning parameter. As discussed by Moodie and Richardson (2010), this is a difficult problem even in better-understood settings where preliminary test based estimators are used; no widely applicable data-driven method for choosing α in this setting is available. Chakraborty et al. (2010) studied the behavior of the usual bootstrap in conjunction with this estimator empirically.

8.3.2 The Soft-Threshold Estimator

The general form of the soft-threshold pseudo-outcome considered here is

$$\hat{Y}_{1i}^{ST} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{\lambda_i}{|\hat{\psi}_2^T H_{21,i}|^2}\right)^+, \; i = 1, \ldots, n, \quad (8.10)$$

where $x^+ = x\mathbb{I}[x > 0]$ stands for the positive part of a function, and λ_i (>0) is a tuning parameter associated with the i-th subject in the sample (again possibly depending on the variability of the linear combination $\hat{\psi}_2^T H_{21,i}$ for that subject). In the context of regression shrinkage (Breiman 1995) and wavelet shrinkage (Gao 1998), the third term on the right side of (8.10) is generally known as the *non-negative garrote* estimator. As discussed by Zou (2006), the non-negative garrote estimator is a special case of the *adaptive lasso* estimator. Chakraborty et al. (2010) proposed this soft-threshold estimator in the context of Q-learning.

Like the hard-threshold pseudo-outcome, $\hat{Y}_1^{ST}$ is also a non-smooth function of $\hat{\psi}_2$ and hence $\hat{\psi}_1^{ST}$ remains a non-regular estimator of ψ_1. However, the problematic term $|\hat{\psi}_2^T H_{21}|$ is thresholded and shrunk towards zero, which reduces the degree of non-regularity. As in the case of hard-threshold estimators, a crucial issue here is to choose a data-driven tuning parameter λ_i; see below for a choice of λ_i following a Bayesian approach. Figure 8.2 presents the hard-max, the hard-threshold, and the soft-threshold pseudo-outcomes.

Choice of Tuning Parameters

A hierarchical Bayesian formulation of the problem, inspired by the work of Figueiredo and Nowak (2001) in wavelets, was used by Chakraborty et al. (2010) to choose the λ_is in a data-driven way. It turns out that the estimator (8.10) with $\lambda_i = 3H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}/n$, $i = 1, \ldots, n$, where $n^{-1}\hat{\Sigma}_{\hat{\psi}_2}$ is the estimated covariance matrix of $\hat{\psi}_2$, is an approximate empirical Bayes estimator. The following theorem can be used to derive the choice of λ_i.

Theorem 8.1. *Let X be a random variable such that $X|\mu \sim N(\mu, \sigma^2)$ with known variance σ^2. Let the prior distribution on μ be given by $\mu|\phi^2 \sim N(0, \phi^2)$, with Jeffrey's noninformative hyper-prior on ϕ^2, i.e., $p(\phi^2) \propto 1/\phi^2$. Then an empirical Bayes estimator of $|\mu|$ is given by*

$$\begin{aligned} |\hat{\mu}|^{EB} &= X\left(1 - \frac{3\sigma^2}{X^2}\right)^+ \left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1 - \frac{3\sigma^2}{X^2}\right)^+}\right) - 1\right) \\ &\quad + \sqrt{\frac{2}{\pi}}\,\sigma\sqrt{\left(1 - \frac{3\sigma^2}{X^2}\right)^+} \exp\left\{-\frac{X^2}{2\sigma^2}\left(1 - \frac{3\sigma^2}{X^2}\right)^+\right\}, \end{aligned} \quad (8.11)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

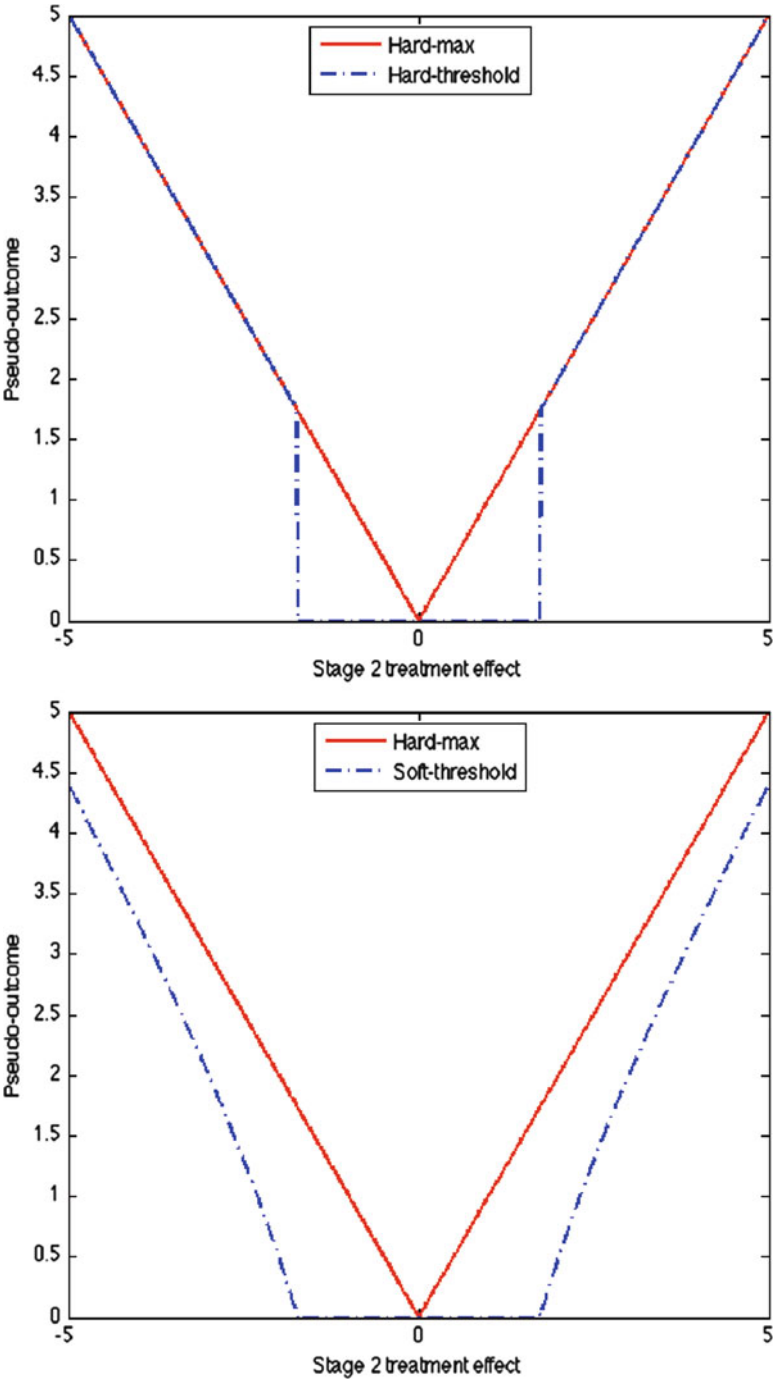

Fig. 8.2 Hard-threshold and soft-threshold pseudo-outcomes compared with the hard-max pseudo-outcome

The proof can be found in Chakraborty et al. (2010).

Clearly, $|\hat{\mu}|^{EB}$ is a thresholding rule, since $|\hat{\mu}|^{EB} = 0$ for $|X| < \sqrt{3}\sigma$. Moreover, when $|X/\sigma|$ is large, the second term of (8.11) goes to zero exponentially fast, and

$$\left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\right)-1\right) \approx (2\,\mathbb{I}[X>0]-1) = sign(X).$$

Consequently, the empirical Bayes estimator is approximated by

$$|\hat{\mu}|^{EB} \approx X\left(1-\frac{3\sigma^2}{X^2}\right)^+ sign(X) = |X|\left(1-\frac{3\sigma^2}{X^2}\right)^+. \tag{8.12}$$

Now for $i = 1,\ldots,n$ separately, put $X = \hat{\psi}_2^T H_{21,i}$, and $\mu = \psi_2^T H_{21,i}$ (for fixed $H_{21,i}$); and plug in $\hat{\sigma}^2 = H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}/n$ for σ^2. This leads to a choice of λ_i in the soft-threshold pseudo-outcome (8.10):

$$\begin{aligned}\hat{Y}_{1i}^{ST} &= Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{3H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right)^+, \\ &= Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{3H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right) \cdot \mathbb{I}\left[\frac{\sqrt{n}|\hat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}}} > \sqrt{3}\right], \\ &\qquad i = 1,\ldots,n. \end{aligned} \tag{8.13}$$

The presence of the indicator function in (8.13) indicates that $\hat{Y}_{1i}^{ST}$ is a thresholding rule for small values of $|\hat{\psi}_2^T H_{21,i}|$, while the term just preceding the indicator function makes $\hat{Y}_{1i}^{ST}$ a shrinkage rule for moderate to large values of $|\hat{\psi}_2^T H_{21,i}|$ (for which the indicator function takes the value one).

Interestingly, the thresholding rule in (8.13) also provides some guidance for choosing the tuning parameter of the hard-threshold estimator. Note that the indicator function in (8.13) corresponds to a pretest that uses a critical value of $\sqrt{3} = 1.7321$; equating this value to $z_{\alpha/2}$ and solving for α, we get $\alpha = 0.0833$. Hence a hard-threshold estimator with tuning parameter $\alpha = 0.0833 \approx 0.08$ corresponds to the soft-threshold estimator without the shrinkage effect. Chakraborty et al. (2010) empirically showed that the hard-threshold estimator with $\alpha = 0.08$ outperformed other choices of this tuning parameter as reported in the original paper by Moodie and Richardson (2010).

8.3.3 *Analysis of Smoking Cessation Data: An Illustration, Continued*

To demonstrate the use of the soft-threshold method in a health application, here we present the analysis of the smoking cessation data described earlier in Sects. 2.4.1

and 3.4.3. The variables considered here are the same as those considered in Sect. 3.4.3. To find the optimal DTR, we applied both the hard-max and the soft-threshold estimators within the Q-learning framework. This involved:

1. Fit stage 2 regression ($n = 281$) of `FF6Quitstatus` using the model:

$$\begin{aligned}\texttt{FF6Quitstatus} = {} & \beta_{20} + \beta_{21} \times \texttt{motivation} + \beta_{22} \times \texttt{source} \\ & + \beta_{23} \times \texttt{selfefficacy} + \beta_{24} \times \texttt{story} \\ & + \beta_{25} \times \texttt{education} + \beta_{26} \times \texttt{PQ6Quitstatus} \\ & + \beta_{27} \times \texttt{source} \times \texttt{selfefficacy} \\ & + \beta_{28} \times \texttt{story} \times \texttt{education} \\ & + \Big(\psi_{20} + \psi_{21} \times \texttt{PQ6Quitstatus}\Big) \times \texttt{FFarm} + \texttt{error}.\end{aligned}$$

2. Construct the hard-max pseudo-outcome ($\hat{Y}_1$) and the soft-threshold pseudo-outcome ($\hat{Y}_1^{ST}$) for the stage 1 regression by plugging in the stage 2 estimates:

$$\begin{aligned}\hat{Y}_1 = {} & \texttt{PQ6Quitstatus} + \hat{\beta}_{20} + \hat{\beta}_{21} \times \texttt{motivation} + \hat{\beta}_{22} \times \texttt{source} \\ & + \hat{\beta}_{23} \times \texttt{selfefficacy} + \hat{\beta}_{24} \times \texttt{story} \\ & + \hat{\beta}_{25} \times \texttt{education} + \hat{\beta}_{26} \times \texttt{PQ6Quitstatus} \\ & + \hat{\beta}_{27} \times \texttt{source} \times \texttt{selfefficacy} + \hat{\beta}_{28} \times \texttt{story} \times \texttt{education} \\ & + \Big|\hat{\psi}_{20} + \hat{\psi}_{21} \times \texttt{PQ6Quitstatus}\Big|;\end{aligned}$$

and

$$\begin{aligned}\hat{Y}_1^{ST} = {} & \texttt{PQ6Quitstatus} + \hat{\beta}_{20} + \hat{\beta}_{21} \times \texttt{motivation} + \hat{\beta}_{22} \times \texttt{source} \\ & + \hat{\beta}_{23} \times \texttt{selfefficacy} + \hat{\beta}_{24} \times \texttt{story} \\ & + \hat{\beta}_{25} \times \texttt{education} + \hat{\beta}_{26} \times \texttt{PQ6Quitstatus} \\ & + \hat{\beta}_{27} \times \texttt{source} \times \texttt{selfefficacy} + \hat{\beta}_{28} \times \texttt{story} \times \texttt{education} \\ & + \Big|\hat{\psi}_{20} + \hat{\psi}_{21} \times \texttt{PQ6Quitstatus}\Big| \\ & \quad \times \left(1 - \frac{3\text{Var}(\hat{\psi}_{20} + \hat{\psi}_{21} \times \texttt{PQ6Quitstatus})}{|\hat{\psi}_{20} + \hat{\psi}_{21} \times \texttt{PQ6Quitstatus}|^2}\right)^+.\end{aligned}$$

   Note that in this case one can construct both versions of the pseudo-outcomes for everyone who participated at stage 1, since there are no variables from post-stage 1 required to do so.
3. Fit stage 1 regression ($n = 1,401$) of the pseudo-outcome using a model of the form:

$$\begin{aligned}\hat{Y}_1 \text{ or } \hat{Y}_1^{ST} = {} & \beta_{10} + \beta_{11} \times \texttt{motivation} \\ & + \beta_{12} \times \texttt{selfefficacy} + \beta_{13} \times \texttt{education}\end{aligned}$$

$$+\left(\psi_{10}^{(1)}+\psi_{11}^{(1)}\times \texttt{selfefficacy}\right)\times \texttt{source}$$
$$+\left(\psi_{10}^{(2)}+\psi_{11}^{(2)}\times \texttt{education}\right)\times \texttt{story}+\texttt{error}.$$

No significant treatment effect was found at the second stage regression, indicating the likely existence of non-regularity. At stage 1, for either estimator, 95 % confidence intervals were constructed by *centered percentile bootstrap* (Efron and Tibshirani 1993) using 1,000 bootstrap replications. The stage 1 analysis summary is presented in Table 8.1. In this case, the hard-max and the soft-threshold estimators produced similar results.

Table 8.1 Regression coefficients and 95 % bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators (significant effects are in bold)

	Hard-max		Soft-threshold	
Variable	Coefficient	95 % CI	Coefficient	95 % CI
`motivation`	0.04	(−0.00, 0.08)	**0.04**	(0.00, 0.08)
`selfefficacy`	**0.03**	(0.00, 0.06)	**0.03**	(0.00, 0.06)
`education`	−0.01	(−0.07, 0.06)	−0.01	(−0.07, 0.06)
`source`	−0.15	(−0.35, 0.06)	−0.15	(−0.35, 0.06)
`source` × `selfefficacy`	**0.03**	(0.00, 0.06)	**0.03**	(0.00, 0.06)
`story`	0.05	(−0.01, 0.11)	0.05	(−0.01, 0.11)
`story` × `education`	**−0.07**	(−0.13, −0.01)	**−0.07**	(−0.13, −0.01)

From the above analysis, it is found that at stage 1 subjects with higher level of `motivation` or `selfefficacy` are more likely to quit. The highly personalized level of `source` is more effective for subjects with a higher `selfefficacy` ($\geq$7), and deeply tailored level of `story` is more effective for subjects with lower `education` ($\leq$ high school); these two conclusions can be drawn from the interaction plots (with confidence intervals) presented in Fig. 3.2 (see Sect. 3.4.3). Thus to maximize each individual's chance of quitting over the two stages, the web-based smoking cessation intervention should be designed in future such that: (1) smokers with high `self-efficacy` ($\geq$7) are assigned to highly personalized level of `source`, and (2) smokers with lower `education` are assigned to deeply tailored level of `story`.

8.4 Penalized Q-learning

In the threshold methods considered earlier, the stage 1 pseudo-outcomes can be viewed as *shrinkage functionals* of the least squares estimators of the stage 2 parameters. However, they are not optimizers of any explicit objective function (except in the special case of only one covariate or an orthonormal design). The *Penalized Q-learning* (hereafter referred to as *PQ-learning*) approach, recently proposed by Song et al. (2011), applies the shrinkage idea with Q-learning by considering an explicit penalized regression at stage 2. The main distinction between the penalization used

here and that used in the context of variable selection is in the "target" of penalization: while penalties are applied to each variable (covariate) in a variable selection context, they are applied on each subject in the case of PQ-learning.

Let $\theta_j = (\beta_j^T, \psi_j^T)^T$ for $j = 1, 2$. PQ-learning starts by considering a penalized least squares optimization at stage 2; it minimizes the objective function

$$W_2(\theta_2) = \sum_{i=1}^{n} \left(Y_{2i} - Q_2^{opt}(H_{2i}, A_{2i}; \beta_2, \psi_2) \right)^2 + \sum_{i=1}^{n} J_{\lambda_n}\left(|\psi_2^T H_{21,i}| \right)$$

with respect to θ_2 to obtain the stage 2 estimates $\hat{\theta}_2$, where $J_{\lambda_n}(\cdot)$ is a pre-specified penalty function and λ_n is a tuning parameter. The penalty function can be taken directly from the variable selection literature; in particular Song et al. (2011) uses the *adaptive lasso* (Zou 2006) penalty, where $J_{\lambda_n}(\theta) = \lambda_n \theta / |\hat{\theta}|^\alpha$ with $\alpha > 0$ and $\hat{\theta}$ being a $\sqrt{n}$-consistent estimator of θ. Furthermore, as in the adaptive lasso procedure, the tuning parameter λ_n is taken to satisfy $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$. The rest of the Q-learning algorithm (hard-max version) is unchanged in PQ-learning.

The above minimization is implemented via *local quadratic approximation* (LQA), following Fan and Li (2001). The procedure starts with an initial value $\hat{\psi}_{2(0)}$ of ψ_2, and then uses LQA for the penalty terms in the objective function:

$$J_{\lambda_n}\left(|\psi_2^T H_{21,i}| \right) \approx J_{\lambda_n}\left(|\hat{\psi}_{2(0)}^T H_{21,i}| \right) + \frac{1}{2} \frac{J'_{\lambda_n}\left(|\hat{\psi}_{2(0)}^T H_{21,i}| \right)}{|\hat{\psi}_{2(0)}^T H_{21,i}|} \left((\psi_2^T H_{21,i})^2 - (\hat{\psi}_{2(0)}^T H_{21,i})^2 \right)$$

for ψ_2 close to $\hat{\psi}_{2(0)}$. Hence the objective function can be locally approximated, up to a constant, by

$$\sum_{i=1}^{n} \left(Y_{2i} - Q_2^{opt}(H_{2i}, A_{2i}; \beta_2, \psi_2) \right)^2 + \frac{1}{2} \sum_{i=1}^{n} \frac{J'_{\lambda_n}\left(|\hat{\psi}_{2(0)}^T H_{21,i}| \right)}{|\hat{\psi}_{2(0)}^T H_{21,i}|} (\psi_2^T H_{21,i})^2.$$

When Q-functions are approximated by linear models as in (3.8), the above minimization problem has a closed form solution:

$$\begin{aligned}
\hat{\psi}_2 &= [\mathbf{X}_{22}(\mathbf{I} - \mathbf{X}_{21}(\mathbf{X}_{21}^T\mathbf{X}_{21})^{-1}\mathbf{X}_{21}^T + \mathbf{D})\mathbf{X}_{22}]^{-1}\mathbf{X}_{22}^T(\mathbf{I} - \mathbf{X}_{21}(\mathbf{X}_{21}^T\mathbf{X}_{21})^{-1}\mathbf{X}_{21}^T)Y_2, \\
\hat{\beta}_2 &= (\mathbf{X}_{21}^T\mathbf{X}_{21})^{-1}\mathbf{X}_{21}^T(\mathbf{Y}_2 - \mathbf{X}_{22}\hat{\psi}_2),
\end{aligned}$$

where $\mathbf{X}_{22}$ is the matrix with i-th row equal to $H_{21,i}^T A_{2i}$, $\mathbf{X}_{21}$ is the matrix with i-th row equal to $H_{20,i}^T$, $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{D}$ is an $n \times n$ diagonal matrix with $D_{ii} = \frac{1}{2} J'_{\lambda_n}\left(|\hat{\psi}_{2(0)}^T H_{21,i}| \right) / |\hat{\psi}_{2(0)}^T H_{21,i}|$, and Y_2 is the vector of Y_{2i} values. The above minimization procedure can be continued for more than one step or until convergence. However, as discussed by Fan and Li (2001), either the one-step or multi-step procedure will be as efficient as the fully iterative procedure as long as the initial estimators are good enough.

Inference for θ_js in the context of PQ-learning is conducted via asymptotic theory. Under a set of regularity conditions, Song et al. (2011) proved that:

1. $\hat{\theta}_2$ is a $\sqrt{n}$-consistent estimator for the true value of θ_2.
2. *Oracle Property*: With probability tending to 1, PQ-learning can identify the individuals for whom the stage 2 treatment effect is zero.
3. Both $\hat{\theta}_2$ and $\hat{\theta}_1$ are asymptotically normal.

Variance Estimation

Song et al. (2011) provided a sandwich type plug-in estimator for the variance of $\hat{\theta}_2$:

$$\widehat{cov}(\hat{\theta}_2) = (\hat{I}_{20} + \hat{\Sigma})^{-1} \hat{I}_{20} (\hat{I}_{20} + \hat{\Sigma})^{-1},$$

where $\hat{I}_{20} \equiv \mathbb{P}_n[\nabla^2_{\theta_2\theta_2}(Y_2 - Q_2^{opt}(H_2, A_2; \theta_2))^2]$ is the empirical Hessian matrix and $\hat{\Sigma} = diag\{\mathbf{0}, \mathbb{P}_n J''_{\lambda_n}(|\hat{\psi}_2^T H_{21}|) H_{21} H_{21}^T\}$. The above variance formula can be further approximated by ignoring $\hat{\Sigma}$, in which case $\widehat{cov}(\hat{\theta}_2) = \hat{I}_{20}^{-1}$. Song et al. (2011) reported having used this reduced formula in their simulation studies and achieved good empirical performance. Likewise, the estimated variance for $\hat{\theta}_1$ is given by:

$$\widehat{cov}(\hat{\theta}_1) = \hat{I}_{10}^{-1} \Big[cov\Big\{ \nabla_{\theta_1} Q_1^{opt}(H_1, A_1; \hat{\theta}_1) \Big(Y_1 + \max_{a_2} Q_2^{opt}(H_2, a_2; \hat{\theta}_2) - Q_1^{opt}(H_1, A_1; \hat{\theta}_1) \Big) + \mathbb{P}_n Z_1 \vec{S}_2{}^T \widehat{cov}(\hat{\theta}_2) \vec{S}_2 Z_1^T \Big\} \Big] \hat{I}_{10}^{-1},$$

where $\hat{I}_{10} \equiv \mathbb{P}_n[\nabla^2_{\theta_1\theta_1}(Y_1 + \max_{a_2} Q_2^{opt}(H_2, a_2; \hat{\theta}_2) - Q_1^{opt}(H_1, A_1; \theta_1))^2]$ is the empirical Hessian matrix.

There are a few characteristics of the PQ-learning approach that demand some discussion. First, this approach offers a data-analyst the ability to calculate standard errors using explicit formulae, which should be less time-consuming than a bootstrap procedure. However in the present era of fast computers, the difference in computing time between analytic and bootstrap approaches is gradually diminishing. Second, the asymptotic theory of PQ-learning assumes a finite support for the H_{21} values, which is achieved when only discrete covariates are used in the analysis. Thus, if there are important continuous covariates in a study, one must first discretize the continuous covariates before being able to use PQ-learning. Third, the success of PQ-learning in addressing non-regularity crucially depends on the "oracle property" described above; this property dictates that after the penalized regression in stage 2, all subsequent inference will be the same as if the analyst knew which subjects had no treatment effect. However this property does not say anything about very small effects that are not exactly zero but are indistinguishable from zero in finite samples due to noise in the data (e.g. in "near non-regular" cases; see Sect. 8.8). It has been widely argued (see e.g. Leeb and Pötscher 2005; Pötscher 2007; Pötscher and Schneider 2008; Laber and Murphy 2011) that characterizing non-regular settings by a condition like $P[H_2 : \psi_2^T H_{21} = 0] > 0$ is really a working assumption to reflect

the uncertainty about the optimal treatment for patients with 'small' – rather than zero – treatment effects. Such situations may be better handled by a local asymptotic framework. From this perspective, the PQ-learning method is still non-regular as it is not consistent under local alternatives; see Laber et al. (2011) for further details on this issue.

8.5 Double Bootstrap Confidence Intervals

The *double bootstrap* (see, e.g. Davison and Hinkley 1997; Nankervis 2005) is a computationally intensive method for constructing CIs. Chakraborty et al. (2010) implemented this method for inference in the context of Q-learning. Empirically it was found to offer valid CIs for the policy parameters in the face of non-regularity. Below we present a brief description.

Let $\hat{\theta}$ be an estimator of a parameter θ and $\hat{\theta}^*$ be its bootstrap version. As is well-known, the $100(1-\alpha)\,\%$ percentile bootstrap CI is given by $\left(\hat{\theta}^*_{(\frac{\alpha}{2})}, \hat{\theta}^*_{(1-\frac{\alpha}{2})}\right)$, where $\hat{\theta}^*_\gamma$ is the 100γ-th percentile of the bootstrap distribution. Then the double (percentile) bootstrap CI is calculated as follows:

1. Draw B_1 first-step bootstrap samples from the original data. For each first-step bootstrap sample, calculate the bootstrap version of the estimator $\hat{\theta}^{*b}$, $b = 1, \ldots, B_1$.
2. Conditional on each first-step bootstrap sample, draw B_2 second-step (nested) bootstrap samples and calculate the double bootstrap versions of the estimator, e.g., $\hat{\theta}^{**bm}$, $b = 1, \ldots, B_1$, $m = 1, \ldots, B_2$.
3. For $b = 1, \ldots, B_1$, calculate $u^{*b} = \frac{1}{B_2} \sum_{m=1}^{B_2} \mathbb{I}[\hat{\theta}^{**bm} \leq \hat{\theta}]$, where $\hat{\theta}$ is the estimator based on the original data.
4. The double bootstrap CI is given by $\left(\hat{\theta}^*_{\hat{q}(\frac{\alpha}{2})}, \hat{\theta}^*_{\hat{q}(1-\frac{\alpha}{2})}\right)$, where $\hat{q}(\gamma) = u^*_{(\gamma)}$, the 100γ-th percentile of the distribution of u^{*b}, $b = 1, \ldots, B_1$.

Next we attempt to provide some intuition[1] about the double bootstrap using the *bagged* hard-max estimator. *Bagging* (Breiman 1996), a nickname for *bootstrap aggregating*, is a well-known ensemble method used to smooth "unstable" estimators, e.g. decision trees in classification. Bagging was originally motivated by Breiman as a variance-reduction technique; however Bühlmann and Yu (2002) showed that it is a smoothing operation that also reduces the mean squared error of the estimator in the case of decision trees, where a "hard decision" based on an indicator function is taken. Note that in the context of Q-learning, the hard-max pseudo-outcome can be re-written as

$$\hat{Y}_{1i} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|$$

[1] This is unpublished work, but the first author was pointed to this direction by Dr. Susan Murphy (personal communication).

$$= Y_{1i} + \hat{\beta}_2^T H_{20,i} + (\hat{\psi}_2^T H_{21,i}) \cdot \Big(2 \cdot \mathbb{I}[\hat{\psi}_2^T H_{21,i} > 0] - 1\Big). \tag{8.14}$$

The second term in (8.14) contains an indicator function (as in a decision tree). Hence one can expect that the bagged version of the hard-max estimator will effectively "smooth out" the effect of this indicator function (e.g. replace the hard decision by a soft decision) and hence should reduce the degree of non-regularity. More precisely, bagging would effectively replace the indicator $\mathbb{I}[\hat{\psi}_2^T H_{21,i} > 0]$ by $\Phi\left(\frac{\sqrt{n}\hat{\psi}_2^T H_{21,i}}{\sqrt{H_{21,i}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21,i}}}\right)$; see Bühlmann and Yu (2002) for details. The bagged hard-max estimator of ψ_1 can be calculated as follows:

1. Construct a bootstrap sample of size n from the original data.
2. Compute the bootstrap version $\hat{\psi}_1^*$ of the usual hard-max estimator $\hat{\psi}_1$.
3. Repeat steps 1 and 2 above B_2 times yielding $\hat{\psi}_1^{*1}, \ldots, \hat{\psi}_1^{*B_2}$. Then the bagged hard-max estimator is given by $\hat{\psi}_1^{Bag} = \frac{1}{B_2} \sum_{b=1}^{B_2} \hat{\psi}_1^{*b}$.

When it comes to constructing CIs, the effect of considering a usual bootstrap CI using B_1 replications along with the bagged hard-max estimator (already using B_2 bootstrap replications) is, in a way, equivalent to considering a double bootstrap CI in conjunction with the original (un-bagged) hard-max estimator.

8.6 Adaptive Bootstrap Confidence Intervals

Laber et al. (2011) recently developed a novel *adaptive bootstrap* procedure to construct confidence intervals for linear combinations $c^T \theta_1$ of the stage 1 coefficients in Q-learning, where $\theta_1^T = (\beta_1^T, \psi_1^T)$ and $c \in \mathbb{R}^{\dim(\theta_1)}$ is a known vector. This method is asymptotically valid and gives good empirical performance in finite samples. In this procedure, Laber et al. (2011) considered the asymptotic expansion of $c^T \sqrt{n}(\hat{\theta}_1 - \theta_1)$ and decomposed it as:

$$c^T \sqrt{n}(\hat{\theta}_1 - \theta_1) = \mathbb{W}_n + \mathbb{U}_n,$$

where the first term is smooth and the second term is non-smooth. While $\mathbb{W}_n$ is asymptotically normally distributed, the distribution of $\mathbb{U}_n$ depends on the underlying data-generating process "non-smoothly". To illustrate the effect of this non-smoothness, fix $H_{21} = h_{21}$. If $h_{21}^T \psi_2 > 0$, then $\mathbb{U}_n$ is asymptotically normal with mean zero. On the other hand, $\mathbb{U}_n$ has a non-normal asymptotic distribution if $h_{21}^T \psi_2 = 0$. Thus, the asymptotic distribution of $c^T \sqrt{n}(\hat{\theta}_1 - \theta_1)$ depends abruptly on both the true parameter ψ_2 and the distribution of patient features H_{21}. In particular, the asymptotic distribution of $c^T \sqrt{n}(\hat{\theta}_1 - \theta_1)$ depends on the frequency of patient features $H_{21} = h_{21}$ for which there is no treatment effect (i.e. features for which $h_{21}^T \psi_2 = 0$). As discussed earlier in this chapter, this non-regularity complicates the construction of CIs for $c^T \theta_1$.

The *adaptive bootstrap* confidence intervals are formed by constructing smooth data-dependent upper and lower bounds on $\mathbb{U}_n$, and thereby on $c^T\sqrt{n}(\hat{\theta}_1 - \theta_1)$, by means of a preliminary hypothesis test that partitions the data into two sets: (i) patients for whom there appears to be a treatment effect, and (ii) patients in whom it appears there is no treatment effect, and then drawing bootstrap samples from these upper and lower bounds. The actual bounds are rather complex and difficult to present without going into the details, so the explicit forms will not be presented here. Instead, we focus on communicating the key ideas.

The bounds are formed by finding limits for the error of the overall approximation due to misclassification of patients in the partitioning step. The idea of conducting a preliminary hypothesis test prior to forming estimators or confidence intervals is known as *pretesting* (Olshen 1973). In fact, the hard-threshold estimator (Moodie and Richardson 2010) discussed earlier uses the same notion of pretest. As in the case of hard-thresholding, Laber et al. (2011) conducted a pretest for each *individual* in the data set as follows. Each pretest is based on

$$T_n(h_{21}) \triangleq \frac{n(h_{21}^T\hat{\psi}_2)^2}{h_{21}^T\hat{\Sigma}_{\hat{\psi}_2}h_{21}},$$

where $\hat{\Sigma}_{\hat{\psi}_2}/n$ is the estimated covariance matrix of $\hat{\psi}_2$. Note that $T_n(h_{21})$ corresponds to the usual test statistic when testing the null hypothesis: $h_{21}^T\psi_2 = 0$. The pretests are performed using a cutoff λ_n, which is a tuning parameter of the procedure and can be varied; to optimize performance, Laber et al. (2011) used $\lambda_n = \log\log n$ in their simulation study and data analysis.

Let the upper and lower bounds on $c^T\sqrt{n}(\hat{\theta}_1 - \theta_1)$ discussed above be given by $\mathscr{U}(c)$ and $\mathscr{L}(c)$ respectively; both of these quantities are functions of λ_n. Laber et al. (2011) showed that the limiting distributions of $c^T\sqrt{n}(\hat{\theta}_1 - \theta_1)$ and $\mathscr{U}(c)$ are equal in the case $H_{21}^T\psi_2 \neq 0$ with probability one. Similarly, the limiting distributions of $c^T\sqrt{n}(\hat{\theta}_1 - \theta_1)$ and $\mathscr{L}(c)$ are equal in the case $H_{21}^T\psi_2 \neq 0$ with probability one. That is, when there is a large treatment effect for almost all patients then the upper (or lower) bound is tight. However, when there is a non-null subset of patients for which there is no treatment effect, then the limiting distribution of the upper bound is stochastically larger than the limiting distribution of $c^T\sqrt{n}(\hat{\theta}_1 - \theta_1)$. This adaptivity between non-regular and regular settings is a key feature of this procedure.

Next we discuss how to actually construct the CIs by this procedure. By construction of $\mathscr{U}(c)$ and $\mathscr{L}(c)$, it follows that

$$c^T\hat{\theta}_1 - \frac{\mathscr{U}(c)}{\sqrt{n}} \leq c^T\theta_1 \leq c^T\hat{\theta}_1 - \frac{\mathscr{L}(c)}{\sqrt{n}}.$$

The distributions of $\mathscr{U}(c)$ and $\mathscr{L}(c)$ are approximated using the bootstrap. Let $\hat{u}$ be the $1-\alpha/2$ quantile of the bootstrap distribution of $\mathscr{U}(c)$, and let $\hat{l}$ be the $\alpha/2$ quantile of the bootstrap distribution of $\mathscr{L}(c)$. Then $[c^T\hat{\theta}_1 - \hat{u}/\sqrt{n}, c^T\hat{\theta}_1 - \hat{l}/\sqrt{n}]$ is the adaptive bootstrap CI for $c^T\theta_1$.

Through a series of theorems, Laber et al. (2011) proved the consistency of the bootstrap in this context, and in particular that

$$P\left(c^T\hat{\theta}_1 - \hat{u}/\sqrt{n} \le c^T\theta_1 \le c^T\hat{\theta}_1 - \hat{l}/\sqrt{n}\right) \ge 1-\alpha+o_P(1).$$

The above probability statement is with respect to the bootstrap distribution. Furthermore, if $P(H_{21}^T\psi_2 = 0) = 0$, then the above inequality can be strengthened to equality. This result shows that the adaptive bootstrap method can be used to construct valid (though potentially conservative) confidence intervals regardless of the underlying parameters of the generative model. Moreover, in settings where there is a treatment effect for almost every patient (e.g. regular settings), the adaptive procedure delivers asymptotically exact coverage.

The theory behind adaptive bootstrap CIs uses a *local asymptotic* framework. This framework provides a medium through which a glimpse of finite-sample behavior can be assessed, while retaining the mathematical convenience of large samples. A thorough technical discussion of this framework is beyond the scope of this book; hence here we presented only the key results without making exact statements of the assumptions and theorems. The procedure discussed here can be extended to more than two stages and more than two treatments per stage; see Laber et al. (2011) for details. The main downside to this procedure lies in its complexity – not just in the theory but also in its implementation. Constructing the smooth upper and lower bounds involves solving very difficult nonconvex optimization problems, making it a computationally expensive procedure. This conceptual and computational complexity may be a potential barrier for its wide-spread dissemination.

8.7 *m*-out-of-*n* Bootstrap Confidence Intervals

The *m*-out-of-*n* bootstrap is a well-known tool for producing valid confidence sets for non-smooth functionals (Shao 1994; Bickel et al. 1997). This method is the same as the usual nonparametric bootstrap (Efron 1979) except that the resample size, historically denoted by m, is of a smaller order of magnitude than the original sample size n. More precisely, m depends on n, tends to infinity with n, and satisfies $m = o(n)$. Intuitively, the *m*-out-of-*n* bootstrap works asymptotically by letting the empirical distribution tend to the true generative distribution at a faster rate than the analogous convergence of the bootstrap empirical distribution to the empirical distribution. In essence, this allows the empirical distribution to reach its limit 'first' so that bootstrap resamples behave as if they were drawn from the true generative distribution. Unfortunately, the choice of the resample size m has long been a difficult obstacle since the condition $m = o(n)$ is purely asymptotic and thus provides no guidance for finite samples. Data-driven approaches for choosing m in various contexts were given by Hall et al. (1995), Lee (1999), Cheung et al. (2005), and Bickel and Sakov (2008). However, these choices were not directly

connected with data-driven measures of non-regularity. Chakraborty et al. (2013) recently proposed a method for choosing the resample size m in the context of Q-learning that is directly connected to an estimated degree of non-regularity. This method of choosing m is adaptive in that it leads to the usual n-out-of-n bootstrap in a regular setting and the m-out-of-n bootstrap otherwise. This methodology, developed for producing asymptotically valid confidence intervals for parameters indexing estimated optimal DTRs, is conceptually and computationally simple, making it more appealing to data analysts. This should be contrasted with methods of Robins (2004) and Laber et al. (2011), both of which involve solving difficult non-convex optimization problems (see Laber et al. 2011, for a discussion).

Intuitively, the choice of the resample size m should reflect the degree of non-smoothness in the underlying generative model. The non-smoothness in Q-learning arises when there is an amassing of points on or near the boundary $\{h_{21} : h_{21}^T \psi_2 = 0\}$. Define $p \triangleq P(H_{21}^T \psi_2 = 0)$, and consider the situation where non-regularity does not exist, i.e. $p = 0$. Then $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ is asymptotically normal and the n-out-of-n bootstrap is consistent. However, if $p > 0$, given that $\hat{\psi}_2$ is not exactly equal to the true value, the quantity $\hat{\theta}_1$, as a function of $\hat{\psi}_2$, oscillates with a rate $n^{-1/2}$ around a point where abrupt changes of the asymptotic distribution occur. This is also true for its bootstrap analogue $\hat{\theta}_{1,m}^{(b)}$ while the oscillating rate is $m^{-1/2}$. With a large p, indicating a high degree of non-regularity, it is hoped that this bootstrap analogue oscillates with a rate much slower than $n^{-1/2}$. Therefore, a reasonable class of resample sizes is given by

$$m \triangleq n^{f(p)},$$

where $f(p)$ is a function of p satisfying the following conditions:

(i) $f(p)$ is monotone decreasing in p, takes values in $(0, 1]$ and satisfies $f(0) = 1$; and
(ii) $f(p)$ is continuous and has bounded first derivative.

One still needs to estimate $f(p)$ from data since p is unknown. Define the plug-in estimator for p, $\hat{p} = \mathbb{P}_n \mathbb{I}[n(H_{21}^T \hat{\psi}_2)^2 \leq \tau_n(H_{21})]$ for cutoff $\tau_n(H_{21})$ (see below), where $\mathbb{P}_n$ denotes the empirical average. Thus, naturally, one can use the resample size

$$\hat{m} \triangleq n^{f(\hat{p})}. \tag{8.15}$$

Chakraborty et al. (2013) showed that $\hat{m}/n^{f(p)} \to 1$ almost surely, and thus $\hat{m} \xrightarrow{p} \infty$ and $\hat{m}/n \xrightarrow{p} 0$. For implementation, they proposed a simple form of $f(p)$ satisfying conditions (i) and (ii),

$$\hat{m} \triangleq n^{\frac{1+\alpha(1-\hat{p})}{1+\alpha}}, \tag{8.16}$$

where $\alpha > 0$ is a tuning parameter that can be either fixed at a constant or chosen adaptively using the double bootstrap (see below for the algorithm). Note that for fixed n, $\hat{m}$ is a monotone decreasing function of $\hat{p}$, taking values in the interval $[n^{\frac{1}{1+\alpha}}, n]$. Thus, α governs the smallest acceptable resample size.

Another potentially important tuning parameter is $\tau_n(H_{21})$. For a given patient history h_{21}, the indicator $\mathbb{I}[n(h_{21}^T \hat{\psi}_2)^2 \leq \tau_n(h_{21})]$ can be viewed as the acceptance

region of the null hypothesis $h_{21}^T \psi_2 = 0$. Thus, a natural choice for $\tau_n(h_{21})$ is $\left(h_{21}^T \hat{\Sigma}_{21} h_{21}\right) \cdot \chi^2_{1,1-\nu}$, where $n^{-1}\hat{\Sigma}_{21}$ is the plug-in estimator of the asymptotic covariance matrix of $\hat{\psi}_2$ and $\chi^2_{1,1-\nu}$ is the $(1-\nu)\times 100$ percentile of a χ^2 distribution with 1 degree of freedom. Chakraborty et al. (2013) used $\nu = 0.001$ in their simulations, and also showed robustness of results to this choice of ν via a thorough sensitivity analysis.

As before, let $c \in \mathbb{R}^{\dim(\theta_1)}$ be a known vector. To form a $(1-\eta)\times 100\,\%$ confidence interval for $c^T\theta_1$, first find $\hat{l}$ and $\hat{u}$, the $(\eta/2)\times 100$ and $(1-\eta/2)\times 100$ percentiles of $c^T\sqrt{m}(\hat{\theta}_1^{(b)} - \hat{\theta}_1)$ respectively, where $\hat{\theta}_1^{(b)}$ is the m-out-of-n bootstrap analog of $\hat{\theta}_1$ (the dependence of $\hat{\theta}_1^{(b)}$ on m is implicit in the notation). The confidence interval is then given by $(c^T\hat{\theta}_1 - \hat{u}/\sqrt{m}, c^T\hat{\theta}_1 - \hat{l}/\sqrt{m})$.

Next we describe the double bootstrap procedure for choosing the tuning parameter α employed to define m. Suppose $c^T\theta_1$ is the parameter of interest, and its estimate from the original data is $c^T\hat{\theta}_1$. Consider a grid of possible values of α; Chakraborty et al. (2013) used $\{0.025, 0.05, 0.075, \ldots, 1\}$ in their simulation study and data analysis. The exact algorithm follows.

1. Draw B_1 usual n-out-of-n first-stage bootstrap samples from the data and calculate the corresponding bootstrap estimates $c^T\hat{\theta}_1^{(b_1)}$, $b_1 = 1, \ldots, B_1$. Fix α at the smallest value in the grid.
2. Compute the corresponding values of $\hat{m}^{(b_1)}$ using Eq. (8.16), $b_1 = 1, \ldots, B_1$.
3. Conditional on each first-stage bootstrap sample, draw B_2 $\hat{m}^{(b_1)}$-out-of-n second-stage (nested) bootstrap samples and calculate the double bootstrap versions of the estimate $c^T\hat{\theta}_1^{(b_1 b_2)}$, $b_1 = 1, \ldots, B_1$, $b_2 = 1, \ldots, B_2$.
4. For $b_1 = 1, \ldots, B_1$, compute the $(\eta/2)\times 100$ and $(1-\eta/2)\times 100$ percentiles of $\left\{c^T\sqrt{\hat{m}^{(b_1)}}\left(\hat{\theta}_1^{(b_1 b_2)} - \hat{\theta}_1^{(b_1)}\right),\ b_2 = 1, \ldots, B_2\right\}$, say $\hat{l}_{\mathrm{DB}}^{(b_1)}$ and $\hat{u}_{\mathrm{DB}}^{(b_1)}$ respectively. Construct the *double centered percentile bootstrap* CI from the b_1-th first-stage bootstrap data as $\left(c^T\hat{\theta}_1^{(b_1)} - \hat{u}_{\mathrm{DB}}^{(b_1)}/\sqrt{\hat{m}^{(b_1)}}, c^T\hat{\theta}_1^{(b_1)} - \hat{l}_{\mathrm{DB}}^{(b_1)}/\sqrt{\hat{m}^{(b_1)}}\right)$, $b_1 = 1, \ldots, B_1$.
5. Estimate the coverage rate of the double bootstrap CI from all the first-stage bootstrap data sets as

$$\frac{1}{B_1}\sum_{b_1=1}^{B_1} \mathbb{I}\left[c^T\hat{\theta}_1^{(b_1)} - \hat{u}_{\mathrm{DB}}^{(b_1)}/\sqrt{\hat{m}^{(b_1)}} \le c^T\hat{\theta}_1 \le c^T\hat{\theta}_1^{(b_1)} - \hat{l}_{\mathrm{DB}}^{(b_1)}/\sqrt{\hat{m}^{(b_1)}}\right].$$

6. If the above coverage rate is at or above the nominal rate, up to Monte Carlo error, then pick the current value of α as the final value. Otherwise, update α to its next higher value in the grid.
7. Repeat steps 2–6, until the coverage rate of the double bootstrap CI, up to Monte Carlo error, attains the nominal coverage rate, or the grid is exhausted.[2]

[2] If this unlikely event does occur, one should examine the observed values of $\hat{p}$. If the values of $\hat{p}$ are concentrated close to zero, ν may be increased; if not, the maximal value in the grid should be increased.

Chakraborty et al. (2013) proved the consistency of the m-out-of-n bootstrap in this context, and in particular that

$$P\left(c^T\hat{\theta}_1 - \hat{u}/\sqrt{\hat{m}} \le c^T\theta_1 \le c^T\hat{\theta}_1 - \hat{l}/\sqrt{\hat{m}}\right) \ge 1-\eta+o_P(1).$$

The above probability statement is with respect to the bootstrap distribution. Furthermore, if $P(H_{21}^T\psi_2 = 0) = 0$, then the above inequality can be strengthened to equality. This result shows that the m-out-of-n bootstrap method can be used to construct valid (though potentially conservative) confidence intervals regardless of the underlying parameters or generative model. Moreover, in settings where there is a treatment effect for every patient (regular setting), the adaptive procedure delivers asymptotically exact coverage. Unlike the theoretical setting of adaptive CIs of Laber et al. (2011), the theory of m-out-of-n bootstrap does not involve a local asymptotic framework (in fact it is not consistent under local alternatives).

The m-out-of-n bootstrap procedure for two stages in the context of Q-learning with linear models has been implemented in the R package `qLearn` that is freely available from the Comprehensive R Archive Network (CRAN):

`http://cran.r-project.org/web/packages/qLearn/index.html`.

8.8 Simulation Study

In this section, we consider a simulation study to provide an empirical evaluation of the available inference methods discussed in this chapter. Nine generative models are used in these evaluations, each of them having two stages of treatment and two treatments at each stage. Generically, these models can be described as follows:

- $O_i \in \{-1,1\}, A_i \in \{-1,1\}$ for $i = 1,2$;
- $P(A_1 = 1) = P(A_1 = -1) = 0.5$, $P(A_2 = 1) = P(A_2 = -1) = 0.5$;
- $O_1 \sim \text{Bernoulli}(0.5)$, $O_2|O_1, A_1 \sim \text{Bernoulli}(\text{expit}(\delta_1 O_1 + \delta_2 A_1))$;
- $Y_1 \equiv 0$,
  $Y_2 = \gamma_1 + \gamma_2 O_1 + \gamma_3 A_1 + \gamma_4 O_1 A_1 + \gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2 + \varepsilon$,

where $\varepsilon \sim \mathcal{N}(0,1)$ and $\text{expit}(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$. This class is parameterized by nine quantities $\gamma_1, \gamma_2, \ldots, \gamma_7, \delta_1, \delta_2$.

The form of the above class of generative models, developed by Chakraborty et al. (2010), is useful as it allows one to influence the degree of non-regularity present in the example problems through the choice of γs and δs, and in turn evaluate performance in these different scenarios. Recall that in Q-learning, non-regularity occurs when more than one stage 2 treatment produces exactly or nearly the same optimal expected outcome for a set of patient histories that occur with positive probability. In the model class above, this occurs if the model generates histories for which $\gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2 \approx 0$, i.e., if it generates histories for which Q_2 depends weakly or not at all on A_2. By manipulating the values of γs and δs, we can control: (i) the probability of generating a patient

history such that $\gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2 = 0$, and (ii) the standardized effect size $E(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)/\sqrt{\text{Var}(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)}$. These two quantities, denoted by p and ϕ, respectively, can be thought of as *measures of non-regularity*. Note that for fixed parameter values, the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$ that governs the non-regularity in an example generative model can take only four possible values corresponding to the four possible (O_2, A_1) cells. The cell probabilities can be easily calculated; the formulae are provided in Table 8.2. Using the quantities presented in Table 8.2, one can write

$$E[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1] = q_1 f_1 + q_2 f_2 + q_3 f_3 + q_4 f_4,$$
$$E[(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)^2] = q_1 f_1^2 + q_2 f_2^2 + q_3 f_3^2 + q_4 f_4^2.$$

From these two, one can calculate $Var[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]$, and subsequently the effect size ϕ.

Table 8.2 Distribution of the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$

(O_2, A_1) cell	Cell probability (averaged over O_1)	Value of the linear combination
$(1,1)$	$q_1 \equiv \frac{1}{4}\Big(expit(\delta_1+\delta_2) + expit(-\delta_1+\delta_2)\Big)$	$f_1 \equiv \gamma_5 + \gamma_6 + \gamma_7$
$(1,-1)$	$q_2 \equiv \frac{1}{4}\Big(expit(\delta_1-\delta_2) + expit(-\delta_1-\delta_2)\Big)$	$f_2 \equiv \gamma_5 + \gamma_6 - \gamma_7$
$(-1,1)$	$q_3 \equiv \frac{1}{4}\Big(expit(\delta_1-\delta_2) + expit(-\delta_1-\delta_2)\Big)$	$f_3 \equiv \gamma_5 - \gamma_6 + \gamma_7$
$(-1,-1)$	$q_4 \equiv \frac{1}{4}\Big(expit(\delta_1+\delta_2) + expit(-\delta_1+\delta_2)\Big)$	$f_4 \equiv \gamma_5 - \gamma_6 - \gamma_7$

Table 8.3 provides the parameter settings; the first six of these settings were constructed by Chakraborty et al. (2010), and were described therein as "non-regular," "near-non-regular," and "regular." Example 1 is a setting where there is no treatment effect for any subject (any possible history) in either stage. Example 2 is similar to example 1, where there is a very weak stage 2 treatment effect for every subject, but it is hard to detect the very weak effect given the noise level in the data. Example 3 is a setting where there is no stage 2 treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects. In example 4, there is a very weak stage 2 treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects (the parameters are close to those in example 3). Example 5 is a setting where there is no stage 2 treatment effect for one-fourth of the subjects in the population, but others have a reasonably large effect. Example 6 is a completely regular setting where there is a reasonably large stage 2 treatment effect for every subject in the population. Song et al. (2011) also used these six examples for empirical evaluation of their PQ-learning method.

To these six, Laber et al. (2011) added three further examples labeled A, B, and C. Example A is an example of a strongly regular setting. Example B is an example of a non-regular setting where the non-regularity is strongly dependent on the stage 1 treatment. In example B, for histories with $A_1 = 1$, there is a moderate effect of

A_2 at the second stage. However, for histories with $A_1 = -1$, there is no effect of A_2 at the second stage, i.e., both actions at the second stage are equally optimal. In example C, for histories with $A_1 = 1$, there is a moderate effect of A_2, and for histories with $A_1 = -1$, there is a small effect of A_2. Thus example C is a "near-non-regular" setting that behaves similarly to example B.

Table 8.3 Parameters indexing the example models

Example	γ^T	δ^T	Type	Regularity Measures	
1	$(0,0,0,0,0,0,0)$	$(0.5,0.5)$	Non-regular	$p=1$	$\phi=0/0$
2	$(0,0,0,0,0.01,0,0)$	$(0.5,0.5)$	Near-non-regular	$p=0$	$\phi=\infty$
3	$(0,0,-0.5,0,0.5,0,0.5)$	$(0.5,0.5)$	Non-regular	$p=1/2$	$\phi=1.0$
4	$(0,0,-0.5,0,0.5,0,0.49)$	$(0.5,0.5)$	Near-non-regular	$p=0$	$\phi=1.02$
5	$(0,0,-0.5,0,1.0,0.5,0.5)$	$(1.0,0.0)$	Non-regular	$p=1/4$	$\phi=1.41$
6	$(0,0,-0.5,0,0.25,0.5,0.5)$	$(0.1,0.1)$	Regular	$p=0$	$\phi=0.35$
A	$(0,0,-0.25,0,0.75,0.5,0.5)$	$(0.1,0.1)$	Regular	$p=0$	$\phi=1.035$
B	$(0,0,0,0,0.25,0,0.25)$	$(0,0)$	Non-regular	$p=1/2$	$\phi=1.00$
C	$(0,0,0,0,0.25,0,0.24)$	$(0,0)$	Near-non-regular	$p=0$	$\phi=1.03$

The Q-learning analysis models used in the simulation study are given by

$$Q_2^{opt}(H_2,A_2;\beta_2,\psi_2) = H_{20}^T\beta_2 + H_{21}^T\psi_2 A_2,$$
$$Q_1^{opt}(H_1,A_1;\beta_1) = H_{10}^T\beta_1 + H_{11}^T\psi_1 A_1,$$

where the following patient history vectors are used:

$$H_{20} = (1,O_1,A_1,O_1A_1)^T,$$
$$H_{21} = (1,O_2,A_1)^T,$$
$$H_{10} = (1,O_1)^T,$$
$$H_{11} = (1,O_1)^T.$$

So the models for the Q-functions are correctly specified. For the purpose of inference, the focus is on ψ_{10} and ψ_{11}, the parameters associated with stage 1 treatment A_1 in the analysis model. They can be expressed in terms of γs and δs, the parameters of the generative model, as follows:

$$\psi_{10} = \gamma_3 + q_1|f_1| - q_2|f_2| + q_3|f_3| - q_4|f_4|,$$
$$\text{and } \psi_{11} = \gamma_4 + q_1'|f_1| - q_2'|f_2| - q_3'|f_3| + q_4'|f_4|,$$

where $q_1' = q_3' = \frac{1}{4}(expit(\delta_1+\delta_2) - expit(-\delta_1+\delta_2))$, and $q_2' = q_4' = \frac{1}{4}(expit(\delta_1 - \delta_2) - expit(-\delta_1-\delta_2))$.

Below we will present simulation results to compare the performances of ten competing methods of constructing CIs for the stage 1 parameters of Q-learning. We will be reporting the results for *centered percentile* bootstrap (CPB) (Efron and Tibshirani 1993) method. Let $\hat{\theta}$ be an estimator of θ and $\hat{\theta}^{(b)}$ be its

bootstrap version. Then the $100(1-\alpha)\,\%$ CPB confidence interval is given by $\left(2\hat{\theta}-\hat{\theta}^{(b)}_{(1-\frac{\alpha}{2})}, 2\hat{\theta}-\hat{\theta}^{(b)}_{(\frac{\alpha}{2})}\right)$, where $\hat{\theta}^{(b)}_{\gamma}$ is the 100γ-th percentile of the bootstrap distribution. The competing methods are listed below:

(i) CPB interval in conjunction with the (original) hard-max estimator (CPB-HM);
(ii) CPB interval in conjunction with the hard-threshold estimator with $\alpha = 0.08$ (CPB-HT$_{0.08}$);
(iii) CPB interval in conjunction with the soft-threshold estimator (CPB-ST);
(iv) Double bootstrap interval in conjunction with the hard-max estimator (DB-HM);
(v) Asymptotic confidence interval in conjunction with the PQ-learning estimator (PQ);
(vi) Adaptive bootstrap confidence interval (ACI);
(vii) m-out-of-n CPB interval with fixed $\alpha = 0.1$, in conjunction with the hard-max estimator ($\hat{m}_{0.1}$-CPB-HM);
(viii) m-out-of-n CPB interval with data-driven α chosen by double bootstrap, in conjunction with the hard-max estimator ($\hat{m}_{\hat{\alpha}}$-CPB-HM);
(ix) m-out-of-n CPB interval with fixed $\alpha = 0.1$, in conjunction with the soft-threshold estimator ($\hat{m}_{0.1}$-CPB-ST);
(x) m-out-of-n CPB interval with data-driven α chosen by double bootstrap, in conjunction with the soft-threshold estimator ($\hat{m}_{\hat{\alpha}}$-CPB-ST)

The comparisons are conducted on a variety of settings represented by examples 1–6, A–C, using $N = 1{,}000$ simulated data sets, $B = 1{,}000$ bootstrap replications, and the sample size $n = 300$. However, the double bootstrap CIs are based on $B_1 = 500$ first-stage and $B_2 = 100$ second-stage bootstrap iterations, due to the increased computational burden. Note that here we simply compile the results from the original papers instead of implementing and running them afresh. As a consequence, the results for all the methods across all examples are not available.

We focus on the coverage rate and width of CIs for the parameter ψ_{10} that denotes the main effect of treatment; see Table 8.4 for coverage and Table 8.5 for width of CIs. Different authors also reported results for the stage 1 interaction parameter ψ_{11}; however the effect of non-regularity is less pronounced on this parameter, and hence less interesting for the purpose of illustration of non-regularity and comparison of competing methods.

First, let us focus on Table 8.4. As expected from the inconsistency of the usual n-out-of-n bootstrap in the present non-regular problem, the CPB-HM method shows the problem of under-coverage in most of the examples. While CPB-HT$_{0.08}$, by virtue of bias correction via thresholding (see Moodie and Richardson 2010), performs well in Ex. 1–4, it fares poorly in Ex. 5–6 (and was never implemented in Ex. A–C). Similarly CPB-ST performs well, again by virtue of bias correction via thresholding (see Chakraborty et al. 2010), except in Ex. 6, A, and B. The computationally expensive double bootstrap method (DB-HM) performs well across the first six examples (but was never tried on Ex. A–C). The PQ method (see Song et al. 2011) performs well across the first six examples (but was never tried on

Ex. A–C). PQ-learning is probably the cheapest method computationally, because CIs are constructed by asymptotic formulae rather than any kind of bootstrapping. The ACI, as known from the work of Laber et al. (2011), is a consistent bootstrap procedure that is conservative in some of the highly non-regular settings but delivers coverage rates closer to nominal as the settings become more and more regular (as the degree of non-regularity as measured by p decreases). The behavior of the m-out-of-n bootstrap method with fixed $\alpha = 0.1$ ($\hat{m}_{0.1}$-CPB-HM) is quite similar to that of ACI in that these CIs are conservative in highly non-regular settings, but become close-to-nominal as the settings become more regular. Both ACI and $\hat{m}_{0.1}$-CPB-HM deliver nominal coverage in the two strictly regular settings (Ex. 6, Ex. A) and the one mildly non-regular ($p = \frac{1}{4}$) setting (Ex. 5) considered. However, $\hat{m}_{0.1}$-CPB-HM is computationally much less expensive (about 180 times) than ACI which involves solving a very difficult optimization problem. Interestingly, the m-out-of-n bootstrap with data-driven α via double bootstrap ($\hat{m}_{\hat{\alpha}}$-CPB-HM) offers an extra layer of adaptiveness; fine-tuning α via double bootstrapping reduces the conservatism present in the case of ACI and $\hat{m}_{0.1}$-CPB-HM, and provides nominal coverage in all the examples. However, it is computationally expensive (comparable to ACI). The $\hat{m}_{0.1}$-CPB-ST method performs similarly to the other versions of m-out-of-n bootstrap methods, except perhaps a bit more conservately in non-regular examples. However, this conservatism is reduced in the $\hat{m}_{\hat{\alpha}}$-CPB-ST method. The performances of the last two methods of inference show that the use of m-out-of-n bootstrap is not limited to the original hard-max estimator, but can also be successfully used in conjunction with other non-smooth estimators like the soft-threshold estimator. See Chakraborty et al. (2013) for further discussion on the m-out-of-n bootstrap methods in this context.

Table 8.4 Monte Carlo estimates of coverage probabilities of confidence intervals for the main effect of treatment (ψ_{10}) at the 95 % nominal level. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Examples are designated *NR* non-regular, *NNR* near-non-regular, *R* regular

$n = 300$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB-HM	0.936	0.932*	0.928*	0.921*	0.933*	0.931*	0.944	0.925*	0.922*
CPB-HT$_{0.08}$	0.950	0.953	0.943	0.941	0.932*	0.885*	–	–	–
CPB-ST	0.962	0.961	0.947	0.946	0.942	0.918*	0.918*	0.931*	0.938
DB-HM	0.936	0.936	0.948	0.944	0.942	0.950	–	–	–
PQ	0.951	0.940	0.952	0.955	0.953	0.953	–	–	–
ACI	0.994	0.994	0.975	0.976	0.962	0.957	0.950	0.977	0.976
$\hat{m}_{0.1}$-CPB-HM	0.984	0.982	0.956	0.955	0.943	0.949	0.953	0.971	0.970
$\hat{m}_{\hat{\alpha}}$-CPB-HM	0.964	0.964	0.953	0.950	0.939	0.947	0.944	0.955	0.960
$\hat{m}_{0.1}$-CPB-ST	0.993	0.993	0.979	0.976	0.954	0.943	0.939	0.972	0.977
$\hat{m}_{\hat{\alpha}}$-CPB-ST	0.971	0.976	0.961	0.956	0.949	0.935	0.926*	0.971	0.967

Table 8.5 presents the Monte Carlo estimates of the mean width of CIs. Mean widths corresponding to CPB-HT$_{0.08}$, DB-HM and PQ were not reported in the original papers in which they appeared. Among the rest of the methods, as expected,

Table 8.5 Monte Carlo estimates of the mean width of confidence intervals for the main effect of treatment (ψ_{10}) at the 95 % nominal level. Widths with corresponding coverage significantly below nominal are marked with $*$. Examples are designated *NR* non-regular, *NNR* near-non-regular, *R* regular

$n = 300$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB-HM	0.269	0.269*	0.300*	0.300*	0.320*	0.309*	0.314	0.299*	0.299*
CPB-HT$_{0.08}$	–	–	–	–	–	–	–	–	–
CPB-ST	0.250	0.250	0.293	0.293	0.319	0.319*	0.323*	0.303*	0.304
DB-HM	–	–	–	–	–	–	–	–	–
PQ	–	–	–	–	–	–	–	–	–
ACI	0.354	0.354	0.342	0.342	0.341	0.327	0.327	0.342	0.342
$\hat{m}_{0.1}$-CPB-HM	0.346	0.347	0.341	0.341	0.340	0.341	0.332	0.342	0.343
$\hat{m}_{\hat{\alpha}}$-CPB-HM	0.331	0.331	0.321	0.323	0.330	0.336	0.322	0.328	0.328
$\hat{m}_{0.1}$-CPB-ST	0.324	0.324	0.336	0.336	0.343	0.352	0.343	0.353	0.353
$\hat{m}_{\hat{\alpha}}$-CPB-ST	0.273	0.275	0.306	0.306	0.328	0.349	0.331*	0.330	0.332

CIs constructed via the usual n-out-of-n method (CPB-HM and CPB-ST) have the least width; however these are often associated with under-coverage. The widths of the CIs from the last five methods are quite comparable, with $\hat{m}_{\hat{\alpha}}$-CPB-HM and $\hat{m}_{\hat{\alpha}}$-CPB-ST offering narrower CIs more often.

Given the above findings, it is very hard to declare an overall winner. From a purely theoretical standpoint, the ACI method (Laber et al. 2011) is arguably the strongest since it uses a local asymptotic framework. However it is conceptually complicated, computationally expensive, and often conservative in finite samples. In terms of finite sample performance, both versions of the m-out-of-n bootstrap method (Chakraborty et al. 2013) are at least as good as (and often better than) the ACI method; moreover, they are conceptually very simple and hence may be more attractive to practitioners. The version with fixed α ($\hat{m}_{0.1}$-CPB-HM), while similar to ACI in conservatism, is computationally much cheaper. On the other hand, the version with data-driven choice of α ($\hat{m}_{\hat{\alpha}}$-CPB), while computationally as demanding as the ACI, overcomes the conservatism and provides nominal coverage in all the examples. Nonetheless, m-out-of-n bootstrap methods are valid only under fixed alternatives, not under local alternatives. The PQ-learning method (Song et al. 2011) is also valid only under fixed alternatives but not under local alternatives. This method is non-conservative in Ex. 1–6, and is computationally the cheapest. However its coverage performance in Ex. A–C and the mean widths of CIs resulting from this method in all the examples are unknown to us at this point.

Note that the bias maps of Fig. 8.1 in Sect. 8.2 were created in a scenario where $\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1 = 0$ with positive probability. As noted previously, the generative parameters γ_5, γ_6 and γ_7 correspond to the policy parameters ψ_{20}, ψ_{21}, and ψ_{22} of the analysis model, respectively. For all bias maps in the figure, $\gamma_1 = \gamma_2 = \gamma_4 = 0$ and $\gamma_3 = -0.5$; the first three plots (upper panel) explored the extent of bias in regions around the parameter setting given in Ex. 5 of Table 8.3, while the last three plots (lower panel) explore the extent of bias in regions around the parameter setting in

Ex. 6 of Table 8.3. More precisely, in the first three plots, $\delta_1 = 1$, $\delta_2 = 0$; and only one of $\psi_{20}\,(=\gamma_5)$, $\psi_{21}\,(=\gamma_6)$, or $\psi_{22}\,(=\gamma_7)$ was varied while the remaining were fixed (e.g. $(\psi_{21}, \psi_{22}) = (0.5, 0.5)$ fixed in the first plot, $(\psi_{20}, \psi_{22}) = (1.0, 0.5)$ fixed in the second plot, and $(\psi_{20}, \psi_{21}) = (1.0, 0.5)$ fixed in the third plot). Similarly, in the last three plots, $\delta_1 = \delta_2 = 0.1$; and only one of ψ_{20}, ψ_{21}, or ψ_{22} was varied while the remaining were fixed, e.g. $(\psi_{21}, \psi_{22}) = (0.5, 0.5)$ fixed in the first plot of the lower panel, $(\psi_{20}, \psi_{22}) = (0.25, 0.5)$ fixed in the second plot of the lower panel, and $(\psi_{20}, \psi_{21}) = (0.25, 0.5)$ fixed in the third plot of the lower panel.

8.9 Analysis of STAR*D Data: An Illustration

8.9.1 Background and Study Details

Selective serotonin reuptake inhibitors (SSRIs) are the most commonly prescribed class of antidepressants with simple dosing regimens and a preferable adverse effect profile in comparison to other types of antidepressants (Nelson 1997; Mason et al. 2000). Serotonin is a neurotransmitter in the human brain that regulates a variety of functions including mood. SSRIs affect the serotonin based brain circuits. Other classes of antidepressants may act on serotonin in concert with other neurotransmitter systems, or on entirely different neurotransmitter. While a meta-analysis of all efficacy trials submitted to the US Food and Drug Administration of four antidepressants for which full data sets were available found that pharmacological treatment of depression was no more effective than placebo for mild to moderate depression, other studies support the effectiveness of SSRIs and other antidepressants in primary care settings (Arroll et al. 2005, 2009). Few studies have examined treatment patterns, and in particular, few have studied best prescribing practices following treatment failure.

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was a multisite, multi-level randomized controlled trial designed to assess the comparative effectiveness of different treatment regimes for patients with major depressive disorder, and was introduced earlier in Chap. 2. See Sect. 2.4.2 for a detailed description of the study design along with a schematic of the treatment assignment algorithm. Here we will focus on levels 2, 2A, and 3 of the study only. For the purpose of the current analysis, we will classify the treatments into two categories: (i) treatment with an SSRI (alone or in combination): sertraline (SER), CIT + bupropion (BUP), CIT + buspirone (BUS), or CIT + cognitive psychotherapy (CT) or (ii) treatment with one or more non-SSRIs: venlafaxine (VEN), BUP, or CT alone. Only the patients assigned to CIT + CT or CT alone in level 2 were eligible, in the case of a non-satisfactory response, to move to a supplementary level of treatment (level 2A), to receive either VEN or BUP. Patients not responding satisfactorily at level 2 (and level 2A, if applicable) would continue to level 3. Treatment options at level 3 can

again be classified into two categories, i.e. treatment with (i) SSRI: an augmentation of any SSRI-containing level 2 treatment with either lithium (Li) or thyroid hormone (THY), or (ii) non-SSRI: mirtazapine (MIRT) or nortriptyline (NTP), or an augmentation of any non-SSRI level 2 treatment with either Li or THY.

8.9.2 Analysis

Here we present the analysis originally conducted by Chakraborty et al. (2013). In this analysis, level 2A was considered a part of level 2. This implies that a patient who received an SSRI at level 2 but a non-SSRI at level 2A was considered a recipient of SSRI in the combined level $2+2$A for the present analysis. Also, levels 2 (including 2A, if applicable) and 3 were treated as stages 1 and 2 respectively of the Q-learning framework (level 4 data were not considered in this analysis). As a feature of the trial design, the outcome data at stage 2 were available only for the non-remitters from stage 1; so Chakraborty et al. (2013) defined the overall primary outcome (Y) as the average $-$QIDS score over the stage(s) a patient was present in the study, i.e.

$$Y = R_1 \cdot Y_1 + (1 - R_1) \cdot \Big(\frac{Y_1 + Y_2}{2}\Big),$$

where Y_1 and Y_2 denote the $-$QIDS scores measured at the end of stages 1 and 2 respectively (the negative of QIDS score was taken to make higher values correspond to better outcomes), and $R_1 = 1$ if the subject achieved remission (QIDS ≤ 5) at the end of stage 1, and 0 otherwise.

Following Pineau et al. (2007), three covariates (tailoring variables) were included in this analysis: (i) QIDS-score measured at the start of the level (QIDS.start), (ii) the slope of the QIDS-score over the previous level (QIDS.slope), and (iii) preference. While QIDS.start and QIDS.slope are continuous variables, preference is a binary variable, coded 1 for preference to switch previous treatment and -1 for preference to augment previous treatment or no preference. Following the notation used earlier, let O_{1j} denote the QIDS.start at the jth stage, and O_{2j} denote the QIDS.slope at the jth stage, O_{3j} denote the preference at the jth stage, and A_j denote the treatment at the jth stage, for $j = 1, 2$. Treatment at each stage was coded 1 for SSRI and -1 for non-SSRI. The following models for the Q-functions were employed:

$$Q_2^{opt} = \beta_{02} + \beta_{12}O_{12} + \beta_{22}O_{22} + \beta_{32}O_{32} + \beta_{42}A_1 + \Big(\psi_{02} + \psi_{12}O_{12} + \psi_{22}O_{22}\Big)A_2,$$

$$Q_1^{opt} = \beta_{01} + \beta_{11}O_{11} + \beta_{21}O_{21} + \beta_{31}O_{31} + \Big(\psi_{01} + \psi_{11}O_{11} + \psi_{21}O_{21} + \psi_{31}O_{31}\Big)A_1.$$

To avoid singularity, a preference-by-treatment interaction was not included in the model for Q_2^{opt}; similarly no A_1A_2 interaction was included. According to the above models, the optimal DTR is given by the following two decision rules:

$$d_2^{opt}(H_2) = sign(\psi_{02} + \psi_{12}O_{12} + \psi_{22}O_{22}),$$

$$d_1^{opt}(H_1) = sign(\psi_{01} + \psi_{11}O_{11} + \psi_{21}O_{21} + \psi_{31}O_{31}).$$

One thousand two hundred and sixty patients were used at stage 1 (level 2); a small number (19) of patients were omitted altogether due to gross item missingness in the covariates. Of the 1,260 patients at stage 1, there were 792 who were non-remitters (QIDS > 5) who should have moved to stage 2 (level 3); however, only 324 patients were present at stage 2 while the rest dropped out. To adjust for this dropout, the model for Q_2^{opt} was fitted using inverse probability weighting where the probability of being present at stage 2 was estimated by logistic regression using O_{11}, O_{21}, O_{31}, A_1, $-Y_1$, O_{22}, $O_{11}A_1$, $O_{21}A_1$, and $O_{31}A_1$ as predictors.

Another complexity came up in the computation of the pseudo-outcome, $\max_{a_2} Q_2^{opt}$. Note that for $(792 - 324) = 468$ non-remitters who were absent from stage 2, covariates O_{12} (QIDS.start at stage 2) and O_{32} (preference at stage 2) were missing, rendering the computation of the pseudo-outcome impossible for them. For these patients, the value of O_{12} was imputed by the last observed QIDS score in the previous stage – a sensible strategy for a continuous, slowly changing variable like the QIDS score. On the other hand, the missing values of the binary variable O_{32} (preference at stage 2) were imputed using k nearest neighbor (k-NN) classification, where k was chosen via leave-one-out cross-validation. Following these imputations, Q-learning was implemented for this data; the estimates of the parameters of the Q-functions, along with their 95 % bootstrap CIs were computed. While only the usual bootstrap was used at stage 2, both the usual bootstrap and the adaptive m-out-of-n bootstrap procedure (with α chosen via double bootstrap) were employed at stage 1, to facilitate ready comparison.

8.9.3 Results

Results of the above analysis are presented in Table 8.6. In this analysis, m was chosen to be 1,059 in a data-driven way (using double bootstrap). At both stages, the coefficient of QIDS.start (β_{12} and β_{11}) and the coefficient of preference (β_{32} and β_{31}) were statistically significant. Additionally ψ_{31}, the coefficient of preference-by-treatment interaction at stage 1 was significantly different from 0; this fact is particularly interesting because it suggests that the decision rule at stage 1 should be individually tailored based on preference.

The estimated optimal DTR can be explicitly described in terms of the $\hat{\psi}$s: $\hat{d}_2^{opt}(H_2) = sign(-0.18 - 0.01O_{12} - 0.25O_{22})$, and $\hat{d}_1^{opt}(H_1) = sign(-0.73 + 0.01O_{11} + 0.01O_{21} - 0.67O_{31})$. That is, the estimated optimal DTR suggests treating a patient at stage 2 with an SSRI if $(-0.18 - 0.01 \times \text{QIDS.start}_2 - 0.25 \times \text{QIDS.slope}_2) > 0$, and with a non-SSRI otherwise. Similarly, it suggests treating a patient at stage 1 with an SSRI if $(-0.73 + 0.01 \times \text{QIDS.start}_1 + 0.01 \times \text{QIDS.slope}_1 - 0.67 \times \text{preference}_1) > 0$, and with a non-SSRI otherwise.

Table 8.6 Regression coefficients and their 95 % centered percentile bootstrap CIs (both the usual n-out-of-n and the novel m-out-of-n) in the analysis of STAR*D data (significant coefficients are in bold)

Parameter	Variable	Estimate	95 % CI (n-out-of-n)	95 % CI (m-out-of-n)
		Stage 2 ($n = 324$)		
β_{02}	Intercept$_2$	-1.66	$(-3.70, 0.43)$	–
β_{12}	QIDS.start$_2$	$\mathbf{-0.72}$	$(-0.87, -0.56)$	–
β_{22}	QIDS.slope$_2$	0.79	$(-0.32, 1.99)$	–
β_{32}	Preference$_2$	**0.74**	$(0.05, 1.50)$	–
β_{42}	Treatment$_1$	0.26	$(-0.38, 0.89)$	–
ψ_{02}	Treatment$_2$	-0.18	$(-2.15, 2.00)$	–
ψ_{12}	Treatment$_2$ × QIDS.start$_2$	-0.01	$(-0.18, 0.13)$	–
ψ_{22}	Treatment$_2$ × QIDS.slope$_2$	-0.25	$(-1.33, 0.94)$	–
		Stage 1 ($n = 1,260$; $\hat{m} = 1,059$)		
β_{01}	Intercept$_1$	-0.47	$(-1.64, 0.71)$	$(-1.82, 0.97)$
β_{11}	QIDS.start$_1$	$\mathbf{-0.55}$	$(-0.63, -0.48)$	$(-0.65, -0.46)$
β_{21}	QIDS.slope$_1$	0.12	$(-0.36, 0.52)$	$(-0.41, 0.57)$
β_{31}	Preference$_1$	**0.88**	$(0.40, 1.40)$	$(0.35, 1.46)$
ψ_{01}	Treatment$_1$	-0.73	$(-1.84, 0.43)$	$(-1.91, 0.48)$
ψ_{11}	Treatment$_1$ × QIDS.start$_1$	0.01	$(-0.06, 0.09)$	$(-0.07, 0.09)$
ψ_{21}	Treatment$_1$ × QIDS.slope$_1$	0.01	$(-0.44, 0.46)$	$(-0.47, 0.49)$
ψ_{31}	Treatment$_1$ × Preference$_1$	$\mathbf{-0.67}$	$(-1.17, -0.18)$	$(-1.29, -0.16)$

However, these are just the "point estimates" of the optimal decision rules. A measure of confidence for these estimated decision rules can be formulated as follows. Note that the estimated difference in mean outcome at stage 2 corresponding to the two treatment options is given by

$$\begin{aligned} &Q_2^{opt}(H_2, 1; \hat{\beta}_2, \hat{\psi}_2) - Q_2^{opt}(H_2, -1; \hat{\beta}_2, \hat{\psi}_2) \\ &\quad = 2\big(-0.18 - 0.01 \times \text{QIDS.start}_2 - 0.25 \times \text{QIDS.slope}_2\big). \end{aligned}$$

Likewise, the estimated difference in mean pseudo-outcome at stage 1 corresponding to the two treatment options is given by

$$\begin{aligned} &Q_1^{opt}(H_1, 1; \hat{\beta}_1, \hat{\psi}_1) - Q_1^{opt}(H_1, -1; \hat{\beta}_1, \hat{\psi}_1) \\ &\quad = 2\big(-0.73 + 0.01 \times \text{QIDS.start}_1 + 0.01 \times \text{QIDS.slope}_1 - 0.67 \times \text{preference}_1\big). \end{aligned}$$

For any fixed values of QIDS.start, QIDS.slope, and preference, one can construct point-wise CIs for the above difference in mean outcome (or, pseudo-outcome) based on the CIs for the individual ψs, thus leading to a confidence band around the entire function. The mean difference function and its 95 % confidence band over the observed range of QIDS.start and QIDS.slope are plotted for stage 1 (separately for preference = "switch" and preference = "augment or no preference") and for stage 2 (patients with all preferences combined), and are presented in Fig. 8.3. Since the confidence bands in all three panels contain zero, there is insufficient evidence in the data to recommend a unique best treatment.

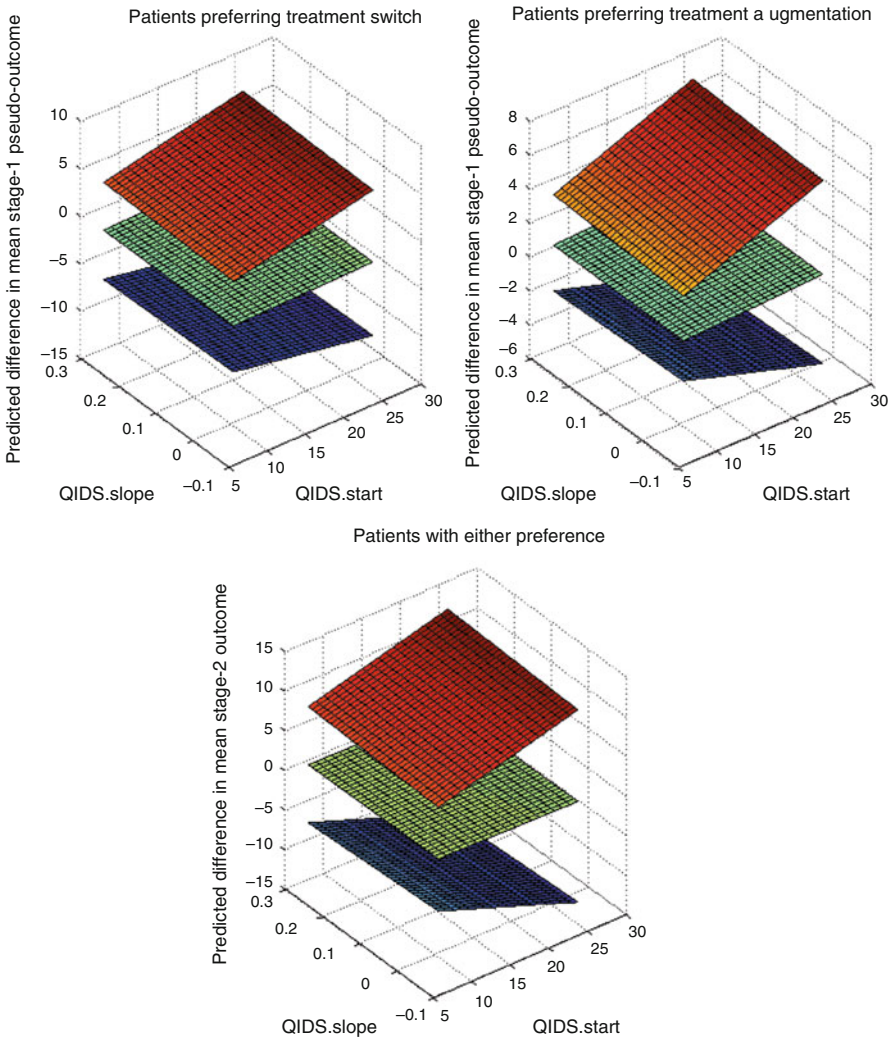

Fig. 8.3 Predicted difference in mean outcome and its 95 % confidence band for: (**a**) patients preferring treatment switch at stage 1; (**b**) patients either preferring treatment augmentation or without preference at stage 1; and (**c**) all patients at stage 2

8.10 Inference About the Value of an Estimated DTR

In Sect. 5.1, we discussed estimation of the *value* of an arbitrary DTR. Once a DTR $\hat{d}$ is estimated from the data (say, via Q-learning, G-estimation, etc.), a key quantity to assess its merit is its true value, $V^{\hat{d}}$. A point estimate of this quantity, say $\hat{V}^{\hat{d}}$, can be obtained, for example, by the IPTW formula (see Sect. 5.1). However it may be more interesting to construct a confidence interval for $V^{\hat{d}}$ and see if the confi-

dence interval contains the optimal value V^{opt} (implying that the estimated DTR is not significantly different from the optimal DTR), or the value of some other prespecified (not necessarily optimal) DTR. It turns out that the estimation of the value of an estimated DTR, or constructing a confidence interval for it, is a very difficult problem.

From Sect. 5.1, we can express the value of $\hat{d}$ by

$$V^{\hat{d}} = \int \Big(\prod_{j=1}^{K} \frac{\mathbb{I}[A_j = \hat{d}_j(H_j)]}{\pi_j(A_j|H_j)} \Big) Y dP_\pi, \tag{8.17}$$

where π is an embedded DTR in the study from which the data arose (e.g. the randomization probabilities in the study); see Sect. 5.1 for further details. Note that (8.17) can be alternatively expressed as

$$\begin{aligned} V^{\hat{d}} &= \int \left\{ \prod_{j=1}^{K} \frac{1}{\pi_j(A_j|H_j)} Y \right\} \Big(\prod_{j=1}^{K} \mathbb{I}[A_j = \hat{d}_j(H_j)] \Big) dP_\pi \\ &= \int c(O_1, A_1, \ldots, O_{K+1}; \pi) \Big(\prod_{j=1}^{K} \mathbb{I}[A_j = \hat{d}_j(H_j)] \Big) dP_\pi \end{aligned} \tag{8.18}$$

where

$$c(O_1, A_1, \ldots, O_{K+1}; \pi) = \left\{ \prod_{j=1}^{K} \frac{1}{\pi_j(A_j|H_j)} Y \right\}$$

is a function of the entire data trajectory and the embedded DTR π. Note that the form of the value function, as expressed in (8.18), is analogous to the test error (misclassification rate) of a classifier in a weighted (or, cost-sensitive) classification problem, where $c(O_1, A_1, \ldots, O_{K+1}; \pi)$ serves as the weight (or, cost) function. Zhao et al. (2012) vividly discussed this analogy in a single-stage decision problem; see also Sect. 5.3.

From this analogy, one can argue that the confidence intervals for the value function could be constructed in ways similar to those for confidence intervals for the test error of a learned classifier. Unfortunately, constructing valid confidence intervals for the test error in classification is an extremely difficult problem due to the inherent non-regularity (note the presence of non-smooth indicator functions in the definition of the value function); see Laber and Murphy (2011) for further details. Standard methods like normal approximation or the usual bootstrap fail in this problem. Laber and Murphy (2011) developed a method for constructing such confidence intervals by use of smooth data-dependent upper and lower bounds on the test error; this method is similar to the method described in Sect. 8.6 in the context of inference for Q-learning parameters. They proved that for linear classifiers, their proposed confidence interval automatically adapts to the non-smoothness of the test error, and is consistent under local alternatives. The method provided nominal coverage on a suite of test problems using a range of classification algorithms and sample

sizes. While intuitively one can expect that this method could be successfully used for constructing confidence intervals for the value function, more research is needed to extend and fine-tune the procedure to the current setting.

8.11 Bayesian Estimation in Non-regular Settings

Robins (2004) considered the behavior of Bayesian estimators under exceptional laws, i.e. the situations where the data-generating distributions lead to non-regularity in frequentist approaches. He considered a prior distribution, $\pi(\psi)$, for the decision rule parameters that is absolutely continuous with respect to a Lebesgue measure and assigns positive mass over the area that includes the true (unknowable) parameter values. Robins (2004) showed that the posterior distribution of the decision rule parameters is non-normal, but that credible intervals based on the posterior distribution are well-defined under all data-generating distributions with probability 1. Furthermore, in many cases the frequentist confidence interval and the Bayesian credible interval based on the highest posterior density will coincide, even at exceptional laws, in very large samples. Robins noted:

> Nonetheless, in practice, if frequentist [confidence interval for ψ] includes exceptional laws (or laws very close to exceptional laws) and thus the set where the likelihood is relatively large contains an exceptional law, it is best not to use a normal approximation, but rather to use either Markov chain Monte Carlo or rejection sampling techniques to generate a sample $\psi^{(v)}, v = 1, \ldots, V$ [...] to construct highest posterior credible intervals, even if one had a prior mass of zero on the exceptional laws.

Following the estimation of the posterior density via direct calculation or, more likely, Markov Chain Monte Carlo, the Bayesian analyst must then formulate optimal decision rules. This can be done in a variety of manners, such as recommending treatment if the posterior median of $H_{j1}^T \psi_j$ is greater than some threshold or if the probability that the posterior mean of $H_{j1}^T \psi_j$ exceeds a threshold is greater than a half. Decisions based on either of these rules will coincide when the posterior is normally distributed, but may not in general (i.e. when laws are exceptional). Alternatively, both Arjas and Saarela (2010) and Zajonc (2012) considered a G-computation like approach, and choose as optimal the rule that maximizes the posterior predictive mean of the outcome.

8.12 Discussion

In this chapter, we have illustrated the problem of non-regularity that arises in the context of inference about the optimal "current" (stage j) treatment rule, when the optimal treatments at subsequent stages are non-unique for at least some non-null proportion of subjects in the population. We have discussed and illustrated the phenomenon using Q-learning as well as G-estimation.

As discussed by Chakraborty et al. (2010), the underlying non-regularity affects the analysis of optimal DTRs in at least two different ways: in some data-generating models it induces bias in the point estimates of the parameters indexing the optimal DTRs, and in other settings it causes lightness of tail of the asymptotic distribution but no bias. The coexistence of these two not-so-well-related issues makes this inference problem unique and challenging.

Non-regularity is an issue in the estimation of the optimal DTRs because it arises when there is no (or a very small) treatment effect at subsequent stages. This is exactly the setting that we are likely to face in many SMARTs in a variety of application areas, due to clinical equipoise (Freedman 1987). Thus we want estimators and inference tools to perform well particularly in non-regular settings. In the case of the hard-max estimator, unfortunately the point of non-differentiability coincides with the parameter value such that $\psi_2^T H_{21} = 0$ (non-unique optimal treatment at the subsequent stage), which causes non-regularity. The threshold estimators (both soft and hard), in some sense, redistribute the non-regularity from this "null point" to two different points symmetrically placed on either side of the null point (see Fig. 8.2). This is one reason why threshold estimators tend to work well in non-regular settings.

However, threshold estimators are still non-smooth, and hence cannot perform uniformly well throughout the parameter space (particularly in regular settings). Furthermore, due to their non-smoothness, the usual bootstrap procedure is still a theoretically invalid inference procedure. Song et al. (2011) extended the idea of thresholding into penalized regression in the Q-learning steps which led to the PQ-learning estimators. Asymptotic CIs for PQ-learning estimators are constructed via analytical formulae, making the procedure computationally cheap. While threshold methods focused primarily on bias correction, PQ-learning was perhaps a more comprehensive attack on the root of the problem.

A different class of methods emerged from the works of Laber et al. (2011) and Chakraborty et al. (2013). These methods do not disturb the original Q-learning (hard-max) estimators, but employ more sophisticated versions of the ordinary bootstrap to mimic the non-regular asymptotic distributions of the estimators. The adaptive method of Laber et al. (2011) is computationally and conceptually complex, while the m-out-of-n bootstrap method is simpler and thus may be more attractive to practitioners. Another computationally expensive method is the double bootstrap, which performs well in conjunction with the original estimator. Yet another method to construct CIs in non-regular settings is the *score method* due to Robins (2004); except for the work of Moodie and Richardson (2010), this approach has not been thoroughly investigated in simulations, likely due to its computational burden.

As discussed by Chakraborty et al. (2013), their adaptive m-out-of-n resampling scheme is conceptually very similar to the *subsampling* method without replacement. In particular, a subsample size of $\tilde{m} = \hat{m}/2$ would enjoy similar asymptotic theory to the adaptive m-out-of-n bootstrap and hence provide consistent confidence sets (see, for example, Politis et al. 1999). One possible advantage of the adaptive m-out-of-n scheme over an adaptive subsampling scheme is that in a regular setting, the m-out-of-n procedure reduces to the familiar n-out-of-n bootstrap which may be

more familiar to applied quantitative researchers. Many of the inference tools discussed in this chapter can be extended to involve more stages and more treatment options at each stage; see, for example, Laber et al. (2011) and Song et al. (2011). Aside from notational complications, extending the adaptive m-out-of-n procedure should also be straightforward.

Finally, we touched on the problems of inference for the value of an estimated DTR, discussing the work of Laber and Murphy (2011), and Bayesian estimation. These are very interesting yet very difficult problems, and little has yet appeared in the literature. More targeted research is warranted.