Chapter 10
Positivity

Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang,
Mark J. van der Laan

The identifiability of causal effects requires sufficient variability in treatment or exposure assignment within strata of confounders. The causal inference literature refers to the assumption of adequate exposure variability within confounder strata as the assumption of positivity or experimental treatment assignment. Positivity violations can arise for two reasons. First, it may be theoretically impossible for individuals with certain covariate values to receive a given exposure of interest. For example, certain patient characteristics may constitute an absolute contraindication to receipt of a particular treatment. The threat to causal inference posed by such structural or theoretical violations of positivity does not improve with increasing sample size. Second, violations or near violations of positivity can arise in finite samples due to chance. This is a particular problem in small samples but also occurs frequently in moderate to large samples when the treatment is continuous or can take multiple levels, or when the covariate adjustment set is large or contains continuous or multilevel covariates. Regardless of the cause, causal effects may be poorly or nonidentified when certain subgroups in a finite sample do not receive some of the treatment levels of interest. In this chapter we will use the term "sparsity" to refer to positivity violations and near-violations arising from either of these causes, recognizing that other types of sparsity can also threaten valid inference.

Data sparsity can increase both the bias and variance of a causal effect estimator; the extent to which each is impacted will depend on the estimator. An estimator-specific diagnostic tool is thus needed to quantify the extent to which positivity violations threaten the validity of inference for a given causal effect parameter (for a given model, data-generating distribution, and finite sample). Wang et al. (2006) proposed such a diagnostic based on the parametric bootstrap. Application of a candidate estimator to bootstrapped data sampled from the estimated data-generating distribution provides information about the estimator's behavior under a data-generating distribution that is based on the observed data. The true parameter value in the bootstrap data is known and can be used to assess estimator bias. A large bias estimate can alert the analyst to the presence of a parameter that is poorly

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_10,

identified, an important warning in settings where data sparsity may not be reflected in the variance of the causal effect estimate.

Once bias due to violations in positivity has been diagnosed, the question remains how best to proceed with estimation. We review several approaches. Identifiability can be improved by extrapolating based on subgroups in which sufficient treatment variability does exist; however, such an approach requires additional parametric model assumptions. Alternative approaches to responding to sparsity include the following: restriction of the sample to those subjects for whom the positivity assumption is not violated (known as trimming); redefinition of the causal effect of interest as the effect of only those treatments that do not result in positivity violations (estimation of the effects of "realistic" or "intention to treat" dynamic regimes); restriction of the covariate adjustment set to exclude those covariates responsible for positivity violations; and, when the target parameter is defined using a marginal structural working model, use of a projection function that focuses estimation on areas of the data with greater support.

As we discuss, all of these approaches change the parameter being estimated by trading proximity to the original target of inference for improved identifiability. We advocate incorporation of this tradeoff into the effect estimator itself. This requires defining a family of parameters whose members vary in their proximity to the initial target and in their identifiability. An estimator can then be defined that selects among the members of this family according to some prespecifed criterion.

10.1 Framework for Causal Effect Estimation

We proceed from the basic premise that model assumptions should honestly reflect investigator knowledge. The SCM framework provides a systematic approach for translating background knowledge into a causal model and corresponding statistical model, defining a target causal parameter, and assessing the identifiability of that parameter. We illustrate this approach using a simple point treatment data structure $O = (W, A, Y) \sim P$ and a nonparametric statistical model augmented with possibly additional causal assumptions, with the SCM given by (2.1), which we restate as follows: $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$. Again, let W denote a set of baseline covariates, A denote a treatment or exposure variable, Y denote an outcome, and $U = (U_W, U_A, U_Y) \sim P_U$ denotes the set of background factors that deterministically assign values to (W, A, Y) according to functions (f_W, f_A, f_Y). We minimize notation by focusing on discrete-valued random variables.

Target parameter. A causal effect can be defined in terms of the joint distribution of the observed data under an intervention on one or more of the structural equations in the corresponding SCM or, equivalently, under an intervention on the corresponding causal graph. For example, consider the postintervention distribution of Y under an intervention on the structural model to set $A = a$. Such an intervention corresponds to replacing $A = f_A(W, U_A)$ with $A = a$ in the structural model (2.1) presented in

Chap. 2. The counterfactual outcome that a given subject with background factors u would have had if he or she were to have received treatment level a is denoted $Y_a(u)$. This counterfactual can be derived as the solution to the structural equation f_Y in the modified equation system with f_A set equal to a and with input $U = u$.

Let F_X denote the distribution of $X = (W, (Y_a : a \in \mathcal{A}))$, where $\mathcal{A}$ denotes the possible values that the treatment variable can take (e.g. $\{0, 1\}$ for a binary treatment). F_X describes the joint distribution of the baseline covariates and counterfactual outcomes under a range of interventions on treatment variable A. A causal effect can be defined as some function of F_X. For example, a common target parameter for binary A is the average treatment effect:

$$E_{F_X}(Y_1 - Y_0), \tag{10.1}$$

or the difference in expected counterfactual outcome if every subject in the population had received vs. had not received treatment.

Alternatively, an investigator may be interested in estimating the average treatment effect separately within certain strata of the population or for nonbinary treatments. Specification of a marginal structural model (a model on the conditional expectation of the counterfactual outcome given effect modifiers of interest) provides one option for defining the target causal parameter in such cases (Chap. 9). Marginal structural models take the following form: $E_{F_X}(Y_a \mid V) = m(a, V \mid \beta)$, where $V \subset W$ denotes the strata in which one wishes to estimate a conditional causal effect. For example, one might specify the following model:

$$m(a, V \mid \beta) = \beta_1 + \beta_2 a + \beta_3 V + \beta_4 aV.$$

For a binary treatment $\mathcal{A} \in \{0, 1\}$, such a model implies an average treatment effect within stratum $V = v$ equal to $\beta_2 + \beta_4 v$.

The true functional form of $E_{F_X}(Y_a \mid V)$ will generally not be known. One option is to assume that the parametric model $m(a, V \mid \beta)$ is correctly specified, or in other words that $E_{F_X}(Y_a \mid V) = m(a, V \mid \beta)$ for some value β. Such an approach, however, can place additional restrictions on the allowed distributions of the observed data and thus change the statistical model. In order to respect the premise that the statistical model should faithfully reflect the limits of investigator knowledge and not be altered in order to facilitate definition of the target parameter, we advocate an alternative approach in which the target causal parameter is defined using a marginal structural working model. Under this approach the target parameter β is defined as the projection of the true causal curve $E_{F_X}(Y_a \mid V)$ onto the specified model $m(a, V \mid \beta)$ according to some projection function $h(a, V)$:

$$\beta(F_X, m, h) = \underset{\beta}{\operatorname{argmin}} E_{F_X}\left[\sum_{a \in \mathcal{A}} (Y_a - m(a, V \mid \beta))^2 h(a, V)\right]. \tag{10.2}$$

When $h(a, V) = 1$, the target parameter β corresponds to an unweighted projection of the entire causal curve onto the working model $m(a, V \mid \beta)$; alternative choices

of h correspond to placing greater emphasis on specific parts of the curve [i.e., on certain (a, V) values].

Use of a marginal structural working model such as (10.2) is attractive because it allows the target causal parameter to be defined within the original statistical model. However, this approach by no means absolves the investigator of careful consideration of marginal structural working model specification. A poorly specified working model $m(a, V \mid \beta)$ may result in a target parameter that provides a poor summary of the features of the true causal relationship that are of interest.

In the following sections we discuss the parameter $\beta(F_X, m, 1)$ as the target of inference, corresponding to estimation of the treatment-specific mean for all levels $a \in \mathcal{A}$ within strata of V as projected onto working model m, with projection $h(a, V) = 1$ chosen to reflect a focus on the entire causal curve. To simplify notation we use β to refer to this target parameter unless otherwise noted.

Identifiability. We assess whether the target parameter β of the counterfactual data distribution F_X is identified as a parameter of the observed data distribution P under causal model (2.1). Because the background factors U are assumed to be jointly independent in SCM (2.1), or in other words the model is assumed to be Markov, we have that

$$P_{F_X}(Y_a = y) = \sum_w P(Y = y \mid W = w, A = a)P(W = w), \tag{10.3}$$

identifying the target parameter β according to projection (10.2) (Pearl 2009). This identifiability result is often referred to as the g-computation formula or g-formula (Robins 1986, 1987a,b). The weaker assumption of randomization (10.4), or the assumption that A and Y_a are conditionally independent given W, is also sufficient for identifiability result (10.3) to hold:

$$A \perp\!\!\!\perp Y_a \mid W \text{ for all } a \in \mathcal{A}. \tag{10.4}$$

Whether or not a given structural model implies that assumption (10.4) holds can be assessed from the graph using the backdoor criterion.

The need for experimentation in treatment assignment. The g-formula (10.3) is valid only if the conditional distributions in the formula are well defined. Let $g(a \mid W) = P(A = a \mid W), a \in \mathcal{A}$ denote the conditional distribution of treatment variable A under the observed data distribution P. If one or more treatment levels of interest do not occur within some covariate strata, the conditional probability $P(Y = y \mid A = a, W = w)$ will not be well defined for some value(s) (a, w) and the identifiability result (10.3) will break down.

A simple example provides insight into the threat to parameter identifiability posed by sparsity of this nature. Consider an example in which $W = I(woman)$, A is a binary treatment, and no women are treated ($g(1 \mid W = 1) = 0$). In this data-generating distribution there is no information regarding outcomes among treated women. Thus, as long as there are women in the target population (i.e., $P(W = 1) >$

0), the average treatment effect $E_{F_X}(Y_1 - Y_0)$ will not be identified without additional parametric assumptions.

This simple example illustrates that a given causal parameter under a given model may be identified for some joint distributions of the observed data but not for others. An additional assumption beyond (10.4) is thus needed to ensure identfiability. We begin by presenting the strong version of this assumption, needed for the identification of $P_{F_X}((Y_a = y, W = w) : a, y, w)$ in a nonparametric model.

Strong Positivity Assumption

$$\inf_{a \in \mathcal{A}} g(a \mid W) > 0, \ P\text{-a.e.} \tag{10.5}$$

The strong positivity assumption states that each possible treatment level occurs with some positive probability within each stratum of W.

Parametric model assumptions may allow the positivity assumption to be weakened. In the example above, an assumption that the treatment effect is the same among treated men and women would result in identification of the average treatment effect (10.1) based on extrapolation from the treatment effect among men (assuming that other identifiability assumptions were met). Parametric model assumptions of this nature are dangerous, however, because they extrapolate to regions of the joint distribution of (A, W) that are not supported by the data. Such assumptions should be approached with caution and adopted only when they have a solid foundation in background knowledge.

In addition to being model-specific, the form of the positivity assumption needed for identifiability is parameter-specific. Many target causal parameters require much weaker versions of positivity than (10.5). To take one simple example, if the target parameter is $E(Y_1)$, the identifiability result only requires that $g(1 \mid W) > 0$ hold; it doesn't matter if there are some strata of the population in which no one was treated. Similarly, the identifiability of $\beta(F_X, m, h)$, defined using a marginal structural working model, relies on a weaker positivity assumption.

Positivity Assumption for $\beta(F_X, h, m)$

$$\sup_{a \in \mathcal{A}} \frac{h(a, V)}{g(a \mid W)} < \infty, \ P\text{-a.e.} \tag{10.6}$$

The choice of projection function $h(a, V)$ used to define the target parameter thus has implications for how strong an assumption about positivity is needed for identifiability. In Sect. 10.4 we consider specification of alternative target parameters that allow for weaker positivity assumptions than (10.5), including parameters indexed by alternative choices of $h(a, V)$. For now we focus on the target parameter β in-

dexed by the choice $h(a, V) = 1$ and note that (10.5) and (10.6) are equivalent for this parameter.

10.2 Estimator-Specific Behavior Under Positivity Violations

Let $\Psi(P_0)$ denote the target parameter of the observed data distribution P_0 of O, which under the assumptions of randomization (10.4) and positivity (10.6) equals the target causal parameter $\beta(F_{X,0}, m, h)$. Estimators of this parameter are denoted $\hat{\Psi}(P_n)$, where P_n is the empirical distribution of a sample of n i.i.d. observations from P_0. We use $Q_{W,0}(w) \equiv P_0(W = w)$, $Q_{Y,0}(y \mid A, W) = P_0(Y = y \mid A, W)$, $\bar{Q}_0 = E_0(Y \mid A, W)$, and $Q_0 \equiv (Q_{W,0}, \bar{Q}_0)$. Recall that $g_0(a \mid W) = P_0(A = a \mid W)$. See Chaps. 1–6 for an introductory presentation of the estimators in this section.

We focus our discussion on bias in the point estimate of the target parameter β_0. While estimates of the variance of estimators of β_0 can also be biased when data are sparse, methods exist to improve variance estimation or to provide upper bounds for the true variance. The nonparametric or semiparametric bootstrap provides one straightforward approach to variance estimation in settings where the central limit theorem may not apply as a result of sparsity; alternative approaches to correct for biased variance estimates are also possible (Rosenblum and van der Laan 2009b). These methods will not, however, protect against misleading inference if the point estimate itself is biased.

10.2.1 MLE

MLEs provide a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{MLE}$. The estimator $\hat{\Psi}_{MLE}(P_n)$ is a substitution estimator based on identifiability result (10.3). It is implemented based on an estimator of Q_0 and its consistency relies on the consistency of this estimator. $Q_{W,0}$ can generally be estimated based on the empirical distribution of W. However, even when positivity is not violated, the dimension of (A, W) is frequently too large for $\bar{Q}_0$ to be estimated simply by evaluating the mean of Y within strata of (A, W). Given an estimator $\bar{Q}_n$ of $\bar{Q}_0$, MLE can be implemented by generating a predicted counterfactual outcome for each subject under each possible treatment: $\hat{Y}_{a,i} = \bar{Q}_n(a, W_i)$ for $a \in \mathcal{A}$, $i = 1, ..., n$. The estimate $\hat{\beta}_{MLE}$ is then obtained by regressing $\hat{Y}_a$ on a and V according to the model $m(a, V \mid \beta)$, with weights based on the projection function $h(a, V)$. When all treatment levels of interest are not represented within all covariate strata [i.e., assumption (10.5) is violated], some of the conditional probabilities in the nonparametric g-formula (10.3) will not be defined. A given estimate $\bar{Q}_n$ may allow the MLE to extrapolate based on covariate strata in which sufficient experimentation in treatment level does exist. Importantly, however, this requires extrapolation of the

fit $\bar{Q}_n$ into areas not supported by the data, and the resulting effect estimates will be biased if the extrapolation used to estimate $\bar{Q}_0$ is misspecified.

10.2.2 IPTW Estimator

The IPTW estimator $\hat{\Psi}_{IPTW}(P_n)$ provides a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{IPTW}$ based on an estimator g_n of $g_0(A \mid W)$. The estimator is defined as the solution in β to the following estimating equation:

$$0 = \sum_{i=1}^{n} \frac{h(A_i, V_i)}{g_n(A_i \mid W_i)} \frac{d}{d\beta} m(A_i, V_i \mid \beta)(Y_i - m(A_i, V_i \mid \beta)),$$

where $h(A, V)$ is the projection function used to define the target causal parameter $\beta(F_X, m, h)$ according to (10.2). The IPTW estimator of the true value β_0 can be implemented as the solution to a weighted regression of the outcome Y on treatment A and effect modifiers V according to model $m(A, V \mid \beta)$, with weights equal to $h(A, V)/g_n(A \mid W)$. Consistency of $\hat{\Psi}_{IPTW}(P_n)$ requires that g_0 satisfy positivity and that g_n be a consistent estimator of g_0. Depending on the choice of projection function, implementation may further require estimation of $h(A, V)$; if one defines the desired projection function as the estimand of the estimator h_n then a consistent estimator of $h(A, V)$ is not required to ensure consistency of the IPTW estimator.

The IPTW estimator is particularly sensitive to bias due to data sparsity. Bias can arise due to structural positivity violations (positivity may not hold for g_0) or may occur by chance because certain covariate and treatment combinations are not represented in a given finite sample [$g_n(a \mid W = w)$ may have values of zero or close to zero for some (a, w) even when positivity holds for g_0 and g_n is consistent] (Wang et al. 2006; Neugebauer and van der Laan 2005; Bembom and van der Laan 2007a; Cole and Hernan 2008; Moore et al. 2009). In the latter case, as fewer individuals within a given covariate stratum receive a given treatment, the weights of those rare individuals who do receive the treatment become more extreme. The disproportionate reliance of the causal effect estimate on the experience of a few unusual individuals can result in substantial finite sample bias.

While values of $g_n(a \mid W)$ remain positive for all $a \in \mathcal{A}$, elevated weights inflate the variance of the effect estimate and can serve as a warning that the data may poorly support the target parameter. However, as the number of individuals within a covariate stratum who receive a given treatment level shifts from few (each of whom receives a large weight and thus elevates the variance) to none, estimator variance can decrease while bias increases rapidly. In other words, when $g_n(a \mid W = w) = 0$ for some (a, w), the weight for a subject with $A = a$ and $W = w$ is infinity; however, as no such individuals exist in the data set, the corresponding threat to valid inference will not be reflected in either the weights or in estimator variance.

Weight truncation. Weights are commonly truncated or bounded in order to improve the performance of the IPTW estimator in the face of data sparsity (Wang et al. 2006; Moore et al. 2009; Cole and Hernan 2008; Kish 1992; Bembom and van der Laan 2008). Weights are truncated at either a fixed or relative level (e.g., at the 1st and 99th percentiles), thereby reducing the variance arising from large weights and limiting the impact of a few possibly nonrepresentative individuals on the effect estimate. This advantage comes at a cost, however, in the form of increased bias due to misspecification of the treatment model g_n, a bias that does not decrease with increasing sample size.

Stabilized weights. The use of projection function $h(a, V) = 1$ implies the use of unstabilized weights. In contrast, stabilized weights, corresponding to a choice of $h(a, V) = g_0(a \mid V)$ [where $g_0(a \mid V) = P_0(A = a \mid V)$] are generally recommended for the implementation of marginal structural-model-based effect estimation. The choice of $h(a, V) = g_0(a \mid V)$ results in a weaker positivity assumption by (10.6). It is important to stress the contrast between assuming a marginal structural model vs. using it as a working model. For example, if A is an ordinal variable with multiple levels, $V = \{\}$, and the target parameter is defined as the true β_0 of a linear marginal structural model $m(a, V \mid \beta) = \beta_{(0)} + \beta_{(1)}a$, it is possible to identify this parameter by using a weight function h that is only nonzero at two values of a chosen such that $g_0(a \mid W) > 0$ for these two values. The corresponding IPTW estimator will extrapolate to levels of A that are sparsely represented in the data by assuming a linear relationship between $E_0 Y_a$ and a for $a \in \mathcal{A}$. However, when the target parameter β is defined using a marginal structural working model according to (10.2) [an approach that acknowledges that the model $m(A, V \mid \beta)$ may be misspecified], the choice of h, including the choice of stabilized vs. unstabilized weights, corresponds to a choice of the target parameter (Neugebauer and van der Laan 2007).

10.2.3 Double Robust Estimators

Double robust approaches to estimation of β include the A-IPTW estimator and the TMLE we focus on in this text. Implementation of double robust estimators requires estimators of both Q_0 and g_0. Double robust estimators remain consistent if either (1) g_n is a consistent estimator of g_0 and g_0 satisfies positivity or (2) Q_n is a consistent estimator of Q_0 and g_n converges to a distribution g^* that satisfies positivity. Thus when positivity holds, these estimators are truly double robust, in the sense that consistent estimation of either g_0 or Q_0 results in a consistent estimator. When positivity fails, however, the consistency of the double robust estimators relies entirely on consistent estimation of Q_0. In the setting of positivity violations, double robust estimators are thus faced with the same vulnerabilities as MLE.

In addition to illustrating how positivity violations increase the vulnerability of double robust estimators to bias resulting from inconsistent estimation of Q_0, these asymptotic results have practical implications for the implementation of the double

robust estimators. Specifically, they suggest that the use of an estimator g_n that yields predicted values in $[0 + \gamma, 1 - \gamma]$ (where γ is some small number) can improve finite sample performance. One way to achieve such bounds is by truncating the predicted probabilities generated by g_n, similar to the process of weight truncation described for the IPTW estimator.

10.3 Diagnosing Bias Due to Positivity Violations

Positivity violations can result in substantial bias, with or without a corresponding increase in variance, regardless of the causal effect estimator used. Practical methods are thus needed to diagnose and quantify estimator-specific positivity bias for a given model, parameter, and sample. Basic descriptive analyses of treatment variability within covariate strata can be helpful; however, this approach quickly becomes unwieldy when the covariate set is moderately large and includes continuous or multilevel variables. Cole and Hernan (2008) suggest a range of informal approaches to diagnose and quantify estimator specific positivity bias when the IPTW estimator is applied. As they note, well-behaved weights are not sufficient to ensure the absence of positivity violations. An alternative formulation is to examine the distribution of the estimated propensity score values given by $g_n(a \mid W)$ for $a \in \mathcal{A}$. However, while useful in diagnosing the presence of positivity violations, examination of the estimated propensity scores does not provide any quantitative estimate of the degree to which such violations result in estimator bias and may pose a threat to inference. The parametric bootstrap can be used to provide an optimistic bias estimate specifically targeted at bias caused by positivity violations and near-violations (Wang et al. 2006).

10.3.1 The Parametric Bootstrap as a Diagnostic Tool

We focus on the bias of estimators that target a parameter of the observed data distribution; this target observed data parameter is equal under the randomization assumption (10.4) to the target causal parameter. [Divergence between the target observed data parameter and target causal parameter when (10.4) fails is a distinct issue not addressed by the proposed diagnostic.] The bias in an estimator is the difference between the true value of the target parameter of the observed data distribution and the expectation of the estimator applied to a finite sample from that distribution:

$$\text{Bias}(\hat{\Psi}, P_0, n) = E_{P_0} \hat{\Psi}(P_n) - \Psi(P_0),$$

where we recall that $\Psi(P_0)$ is the target observed data parameter, $\hat{\Psi}(P_n)$ is an estimator of that parameter (which may be a function of g_n or Q_n or both), and P_n

denotes the empirical distribution of a sample of n i.i.d. observations from the true observed data distribution P_0.

Bias in an estimator can arise due to a range of causes. First, the estimators g_n and Q_n may be inconsistent. Second, g_0 may not satisfy the positivity assumption. Third, consistent estimators g_n and Q_n may still have substantial finite sample bias. This latter type of finite sample bias arises in particular due to the curse of dimensionality in a nonparametric or semiparametric model when g_n or Q_n is a data-adaptive estimator, although it can also be substantial for parametric estimators. Fourth, estimated values of g_n may be equal or close to zero or one, despite use of a consistent estimator g_n and a distribution g_0 that satisfies positivity. The relative contribution of each of these sources of bias will depend on the model, the true data-generating distribution, the estimator, and the finite sample.

The parametric bootstrap provides a tool that allows the analyst to explore the extent to which bias due to any of these causes is affecting a given parameter estimate. The parametric bootstrap-based bias estimate is defined as

$$\widehat{\text{Bias}}_{PB}(\hat{\Psi}, \hat{P}_0, n) = E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0),$$

where $\hat{P}_0$ is an estimate of P_0 and $P_n^{\#}$ is the empirical distribution of a bootstrap sample obtained by sampling from $\hat{P}_0$. In other words, the parametric bootstrap is used to sample from an estimate of the true data-generating distribution, resulting in multiple simulated data sets. The true data-generating distribution and target parameter value in the bootstrapped data are known. A candidate estimator is then applied to each bootstrapped data set and the mean of the resulting estimates compared with the known "truth" (i.e., the true parameter value for the bootstrap data-generating distribution).

We focus on a particular algorithm for parametric bootstrap-based bias estimation, which specifically targets the component of estimator-specific finite sample bias due to violations and near-violations of the positivity assumption. The goal is not to provide an accurate estimate of total bias, but rather to provide a diagnostic tool that can serve as a "red flag" warning that positivity bias may pose a threat to inference. The distinguishing characteristic of the diagnostic algorithm is its use of an estimated data-generating distribution $\hat{P}_0$ that both approximates the true P_0 as closely as possible and is compatible with the estimators $\bar{Q}_n$ and g_n used in $\hat{\Psi}(P_n)$. In other words, $\hat{P}_0$ is chosen such that the estimator $\hat{\Psi}$ applied to bootstrap samples from $\hat{P}_0$ is guaranteed to be consistent unless g_0 fails to satisfy the positivity assumption or g_n is truncated. As a result, the parametric bootstrap provides an optimistic estimate of finite sample bias, in which bias due to model misspecification other than truncation is eliminated.

We refer informally to the resulting bias estimate as Bias_{ETA} because in many settings it will be predominantly composed of bias from the following sources: (1) violation of the positivity assumption by g_0; (2) truncation, if any, of g_n in response to positivity violations; and (3) finite sample bias arising from values of g_n close to zero or one (sometimes referred to as practical violations of the positivity assumption). The term Bias_{ETA} is imprecise because the bias estimated by the proposed

algorithm will also capture some of the bias in $\hat{\Psi}(P_n)$ due to finite sample bias of the estimators g_n and $\bar{Q}_n$ (a form of sparsity only partially related to positivity). Due to the curse of dimensionality, the contribution of this latter source of bias may be substantial when g_n or Q_n is a data-adaptive estimator in a nonparametric or semiparametric model. However, the proposed diagnostic algorithm will only capture a portion of this bias because, unlike P_0, $\hat{P}_0$ is guaranteed to have a functional form that can be well approximated by the data-adaptive algorithms employed by g_n and Q_n. The diagnostic algorithm for Bias_{ETA} is implemented as follows.

Step 1: Estimate P_0. Estimation of P_0 requires estimation of $Q_{W,0}$, g_0, and $Q_{Y,0}$, We define $Q_{\hat{P}_0,W} = Q_{P_n,W}$ (or, in other words, use an estimate based on the empirical distribution of the data), $g_{\hat{P}_0} = g_n$, and $\bar{Q}_{\hat{P}_0} = \bar{Q}_n$. Note that the estimators Q_{P_nW}, g_n, and $\bar{Q}_n$ were all needed for implementation of the IPTW, MLE, and double robust estimators; the same estimators can be used here. Additional steps may be required to estimate the entire conditional distribution of Y given (A, W) (beyond the estimate of its mean given by $\bar{Q}_n$). The true target parameter for the known distribution $\hat{P}_0$ is only a function of $Q_n = (Q_{P_nW}, \bar{Q}_n)$, and $\Psi(\hat{P}_0)$ is the same as the MLE (using Q_n) applied to the observed data:

$$\Psi(\hat{P}_0) = \hat{\Psi}_{MLE}(P_n).$$

Step 2: Generate $P_n^{\#}$ by sampling from $\hat{P}_0$. In the second step, we assume that $\hat{P}_0$ is the true data-generating distribution. Bootstrap samples $P_n^{\#}$, each with n i.i.d. observations, are generated by sampling from $\hat{P}_0$. For example, W can be sampled from the empirical distribution, a binary A might be generated as a Bernoulli with probability $g_n(1 \mid W)$, and a continuous Y can be generated by adding a $N(0, 1)$ error to $\bar{Q}_n(A, W)$ (alternative approaches are also possible).

Step 3: Estimate $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#})$. Finally, the estimator $\hat{\Psi}$ is applied to each bootstrap sample. Depending on the estimator being evaluated, this step involves applying the estimators g_n or Q_n or both to each bootstrap sample. If Q_n or g_n is a data-adaptive estimator, the corresponding data-adaptive algorithm should be rerun in each bootstrap sample; otherwise, the coefficients of the corresponding models should be refit. Bias_{ETA} is calculated by comparing the mean of the estimator $\hat{\Psi}$ across bootstrap samples $[E_{\hat{P}_0}\hat{\Psi}_{IPTW}(P_n^{\#})]$ with the true value of the target parameter under the bootstrap data-generating distribution $[\Psi(\hat{P}_0)]$. Application of the bootstrap to the IPTW estimator offers one particularly sensitive assessment of positivity bias because, unlike the MLE and double robust estimators, the IPTW estimator cannot extrapolate based on $\bar{Q}_n$. However, this approach can be applied to any causal effect estimator, including estimators introduced in Sect. 10.4 that trade off identifiability for proximity to the target parameter. In assessing the threat posed by positivity violations, the bootstrap should ideally be applied to both the IPTW estimator and the estimator of choice.

Remarks on interpretation of the bias estimate. We caution against using the parametric bootstrap for any form of bias correction. The true bias of the estimator is $E_{P_0}\hat{\Psi}(P_n) - \Psi(P_0)$, while the parametric bootstrap estimates $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0)$.

The performance of the diagnostic thus depends on the extent to which $\hat{P}_0$ approximates the true data-generating distribution. This suggests the importance of using flexible data-adaptive algorithms to estimate P_0. Regardless of estimation approach, however, when the target parameter $\Psi(P_0)$ is poorly identified due to positivity violations, $\Psi(\hat{P}_0)$ may be a poor estimate of $\Psi(P_0)$. In such cases one would not expect the parametric bootstrap to provide a good estimate of the true bias. Further, the Bias_{ETA} implementation of the parametric bootstrap provides a deliberately optimistic bias estimate by excluding bias due to model misspecifcation for the estimators g_n and $\bar{Q}_n$.

Rather, the parametric bootstrap is proposed as a diagnostic tool. Even when the data-generating distribution is not estimated consistently, the bias estimate provided by the parametric bootstrap remains interpretable in a world where the estimated data-generating mechanism represents the truth. If the estimated bias is large, an analyst who disregards the implied caution is relying on an unsubstantiated hope that first, he or she has inconsistently estimated the data-generating distribution but still done a reasonable job estimating the causal effect of interest; and second, the true data-generating distribution is less affected by positivity (and other finite sample) bias than is the analyst's best estimate of it.

The threshold level of Bias_{ETA} that is considered problematic will vary depending on the scientific question and the point and variance estimates of the causal effect. With that caveat, we suggest the following two general situations in which Bias_{ETA} can be considered a "red flag" warning: (1) when Bias_{ETA} is of the same magnitude as (or larger than) the estimated standard error of the estimator and (2) when the interpretation of a bias-corrected confidence interval would differ meaningfully from initial conclusions.

10.3.2 Simulations

Data were simulated using a data-generating distribution published by Freedman and Berk (2008). Two baseline covariates, $W = (W_1, W_2)$, were generated bivariate normal, $N(\mu, \Sigma)$, with $\mu_1 = 0.5$, $\mu_2 = 1$, and

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

Y was generated as $1 + A + W_1 + 2W_2 + N(0, 1)$, and $g_0(1 \mid W)$ was given by $\Phi(0.5 + 0.25W_1 + 0.75W_2)$, where Φ is the CDF of the standard normal distribution. With this treatment mechanism $g_0 \in [0.001, 1]$. The target parameter was $E_0(Y_1 - Y_0)$ [corresponding to $\beta_{(1)}$ marginal structural model $m(a \mid \beta) = \beta_{(0)} + \beta_{(1)}a$)]. The true value of the target parameter $\Psi(P_0) = 1$.

The bias, variance, and mean squared error of the MLE, IPTW, A-IPTW, and TMLE estimators were estimated by applying each estimator to 250 samples of size 1000 drawn from this data-generating distribution. The four estimators were imple-

mented with each of the following three approaches: (1) correctly specified model to estimate both $\bar{Q}_0$ and g_0 (CC), (2) correctly specified model to estimate $\bar{Q}_0$ and a misspecified model to estimate g_0 obtained by omitting W_2 from g_n (CM), and (3) correctly specified model to estimate g_0 and a misspecified model to estimate $\bar{Q}_0$ obtained by omitting W_2 from $\bar{Q}_n$ (MC). The double robust and IPTW estimators were further implemented using the following sets of bounds for the values of g_n: $[0, 1]$ (no bounding), $[0.025, 0.975]$, $[0.050, 0.950]$, and $[0.100, 0.900]$. For the IPTW estimator, the latter three bounds correspond to truncation of the unstabilized weights at $[1.03, 40]$, $[1.05, 20]$, and $[1.11, 10]$.

The parametric bootstrap was then applied using the Bias_{ETA} algorithm to 10 of the 250 samples. For each sample and for each model specification, Q_n and g_n were used to draw 1000 parametric bootstrap samples. Specifically, W was drawn from the empirical distribution for that sample, A was generated given the bootstrapped values of W as a series of Bernoulli trials with probability $g_n(1 \mid W)$, and Y was generated given the bootstrapped values of A, W by adding an $N(0, 1)$ error to $\bar{Q}_n(A, W)$. Each candidate estimator was then applied to each bootstrap sample. In this step, the parametric models g_n and $\bar{Q}_n$ were held fixed and their coefficients refit. Bias_{ETA} was calculated for each of the 10 samples as the difference between the mean of the bootstrapped estimator and the initial MLE estimate $\Psi(\hat{P}_0) = \hat{\Psi}_{MLE}(P_n)$ in that sample.

Table 10.1 displays the effect of positivity violations and near-violations on estimator behavior across 250 samples. MSE remained minimally biased when the estimator $\bar{Q}_n$ was consistent; use of inconsistent $\bar{Q}_n$ resulted in bias. Given consistent estimators $\bar{Q}_n$ and g_n, the IPTW estimator was more biased than the other three estimators, as expected given the practical positivity violations present in the simulation. The finite sample performance of the A-IPTW and TMLE estimators was also affected by the presence of practical positivity violations. The double robust estimators achieved the lowest MSE when (1) $\bar{Q}_n$ was consistent and (2) g_n was inconsistent but satisfied positivity (as a result either of truncation or of omission of W_2, a major source of positivity bias). Interestingly, in this simulation TMLE still did quite well when $\bar{Q}_n$ was inconsistent and the model used for g_n was correctly specified but its values bounded at $[0.025, 0.925]$.

The choice of bound imposed on g_n affected both the bias and variance of the IPTW estimator, A-IPTW estimator, and TMLE. As expected, truncation of the IPTW weights improved the variance of the estimator but increased bias. Without additional diagnostic information, an analyst who observed the dramatic decline in the variance of the IPTW estimator that occurred with weight truncation might have concluded that truncation improved estimator performance; however, in this simulation weight truncation increased MSE. In contrast, and as predicted by theory, use of bounded values of g_n decreased MSE of the double robust estimators despite the inconsistency introduced into g_n.

Table 10.2 shows the mean of Bias_{ETA} across 10 of the 250 samples; the variance of Bias_{ETA} across the samples was small [results available in Petersen et al. (2010)]. Based on the results shown in Table 10.1, a red flag was needed for the IPTW estimator with and without bounded g_n and for the TMLE without bounded g_n. (The

Table 10.1 Performance of estimators in 250 simulated data sets of size 1000; rows for each estimator indicate the bound on g_n. *CC* is correctly specified $\bar{Q}_n$ and g_n, *CM* is correctly specified $\bar{Q}_n$ and misspecified g_n, and *MC* is misspecified $\bar{Q}_n$ and correctly specified g_n

	CC			*CM*			*MC*		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
MLE									
	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
IPTW									
[0.000, 1.000]	0.544	0.693	0.989	1.547	0.267	2.660	0.544	0.693	0.989
[0.025, 0.975]	1.080	0.090	1.257	1.807	0.077	3.340	1.080	0.090	1.257
[0.050, 0.950]	1.437	0.059	2.123	2.062	0.054	4.306	1.437	0.059	2.123
[0.100, 0.900]	1.935	0.043	3.787	2.456	0.043	6.076	1.935	0.043	3.787
A-IPTW									
[0.000, 1.000]	0.080	0.966	0.972	−0.003	0.032	0.032	−0.096	16.978	16.987
[0.025, 0.975]	0.012	0.017	0.017	0.006	0.017	0.017	0.430	0.035	0.219
[0.050, 0.950]	0.011	0.014	0.014	0.009	0.014	0.014	0.556	0.025	0.334
[0.100, 0.900]	0.009	0.011	0.011	0.008	0.011	0.011	0.706	0.020	0.519
TMLE									
[0.000, 1.000]	0.251	0.478	0.540	0.026	0.059	0.060	−0.675	0.367	0.824
[0.025, 0.975]	0.016	0.028	0.028	0.005	0.021	0.021	−0.004	0.049	0.049
[0.050, 0.950]	0.013	0.019	0.020	0.010	0.016	0.017	0.163	0.027	0.054
[0.100, 0.900]	0.010	0.014	0.014	0.009	0.013	0.013	0.384	0.018	0.166

Table 10.2 Finite sample bias and mean of Bias_{ETA} across ten simulated data sets of size 1000

		Bound on g_n			
		[0.000, 1.000]	[0.025, 0.975]	[0.050, 0.950]	[0.100, 0.900]
MLE					
Finite sample bias	*CC*	7.01e−03	–	–	–
Mean(Bias_{ETA})	*CC*	−8.51e−04	–	–	–
Mean(Bias_{ETA})	*CM*	2.39e−04	–	–	–
Mean(Bias_{ETA})	*MC*	5.12e−04	–	–	–
IPTW					
Finite sample bias	*CC*	5.44e−01	1.08e+00	1.44e+00	1.93e+00
Mean(Bias_{ETA})	*CC*	4.22e−01	1.04e+00	1.40e+00	1.90e+00
Mean(Bias_{ETA})	*CM*	1.34e−01	4.83e−01	7.84e−01	1.23e+00
Mean(Bias_{ETA})	*MC*	2.98e−01	7.39e−01	9.95e−01	1.35e+00
A-IPTW					
Finite sample bias	*CC*	7.99e−02	1.25e−02	1.07e−02	8.78e−03
Mean(Bias_{ETA})	*CC*	1.86e−03	2.80e−03	5.89e−05	1.65e−03
Mean(Bias_{ETA})	*CM*	−3.68e−04	−6.36e−04	2.56e−05	5.72e−04
Mean(Bias_{ETA})	*MC*	−3.59e−04	1.21e−04	−1.18e−04	−1.09e−03
TMLE					
Finite sample bias	*CC*	2.51e−01	1.60e−02	1.31e−02	9.98e−03
Mean(Bias_{ETA})	*CC*	1.74e−01	4.28e−03	2.65e−04	1.84e−03
Mean(Bias_{ETA})	*CM*	2.70e−02	−3.07e−04	2.15e−04	7.74e−04
Mean(Bias_{ETA})	*MC*	1.11e−01	9.82e−04	−2.17e−04	−1.47e−03

A-IPTW estimator without bounded g_n exhibited a small to moderate amount of bias; however, the variance would likely have alerted an analyst to the presence of sparsity.) The parametric bootstrap correctly identified the presence of substantial finite sample bias in the IPTW estimator for all truncation levels and in the TMLE with unbounded g_n. Bias_{ETA} was minimal for the remaining estimators.

For correctly specified $\bar{Q}_n$ and g_n (g_n unbounded), the mean of Bias_{ETA} across the ten samples was 78% and 69% of the true finite sample bias of the IPTW estimator and TMLE, respectively. The fact that the true bias was underestimated in both cases illustrates a limitation of the parametric bootstrap: its performance, even as an intentionally optimistic bias estimate, suffers when the target estimator is not asymptotically normally distributed. Bounding g_n improved the ability of the bootstrap to accurately diagnose bias by improving estimator behavior (in addition to adding a new source of bias due to truncation of g_n). This finding suggests that practical application of the bootstrap to a given estimator should at minimum generate Bias_{ETA} estimates for a single low level of truncation of g_n in addition to any unbounded estimate. When g_n was bounded, the mean of Bias_{ETA} for the IPTW estimator across the 10 samples was 96 to 98% of the true finite sample bias; the finite sample bias for the TMLE with bounded g_n was accurately estimated to be minimal. Misspecification of g_n or $\bar{Q}_n$ by excluding a key covariate led to an estimated data-generating distribution with less sparsity than the true P_0, and as a result the parametric bootstrap underestimated bias to a greater extent for these model specifications.

While use of an unbounded g_n resulted in an underestimate of the true degree of finite sample bias for the IPTW and TMLE, in this simulation the parametric bootstrap would still have functioned well as a diagnostic in each of the ten samples considered. Table 10.3 reports the output that would have been available to an analyst applying the parametric bootstrap to the unbounded IPTW and TMLE for each of the ten samples. In all samples Bias_{ETA} was of roughly the same magnitude as the estimated standard error of the estimator, and in most was of significant magnitude relative to the point estimate of the causal effect.

The simulation demonstrates how the parametric bootstrap can be used to investigate the tradeoffs between bias due to weight truncation/bounding of g_n and positivity bias. The parametric bootstrap accurately diagnosed both an increase in the bias of the IPTW estimator with increasing truncation and a reduction in the bias of the TMLE with truncation. When viewed in light of the standard error estimates under different levels of truncation, the diagnostic would have accurately suggested that truncation of g_n for the TMLE was beneficial, while truncation of the weights for the IPTW estimator was of questionable benefit. The parametric bootstrap can also be used to provide a more refined approach to choosing an optimal truncation constant based on estimated MSE (Bembom and van der Laan 2008).

These results further illustrate the benefit of applying the parametric bootstrap to the IPTW estimator in addition to the analyst's estimator of choice. Diagnosis of substantial bias in the IPTW estimator due to positivity violations would have alerted an analyst that MLE was relying heavily on extrapolation and that the double

Table 10.3 Estimated causal treatment effect, standard error, and Bias_{ETA} in ten simulated datasets of size 1000; g_n and Q_n correctly specified, g_n unbounded

	IPTW			TMLE		
Sample	$\hat{\beta}_{IPTW}$	$\widehat{\text{SE}}$	Bias_{ETA}	$\hat{\beta}_{TMLE}$	$\widehat{\text{SE}}$	Bias_{ETA}
1	0.207	0.203	0.473	0.827	0.197	0.172
2	1.722	0.197	0.425	0.734	0.114	0.153
3	1.957	0.184	0.306	1.379	0.105	0.087
4	1.926	0.206	0.510	0.237	0.089	0.252
5	2.201	0.192	0.565	2.548	0.182	0.245
6	0.035	0.236	0.520	0.533	0.228	0.234
7	1.799	0.180	0.346	1.781	0.184	0.150
8	0.471	0.215	0.420	1.066	0.114	0.188
9	2.749	0.184	0.391	1.974	0.114	0.161
10	0.095	0.228	0.263	0.628	0.173	0.099

robust estimators were sensitive to bias arising from misspecification of the model used to estimate $\bar{Q}_0$.

10.3.3 HIV Data Application

We analyzed an observational cohort of HIV-infected patients in order to estimate the effect of mutations in the HIV protease enzyme on viral response to the antiretroviral drug lopinavir. The question, data, and analysis have been described previously (Bembom et al. 2009). Here, a simplified version of prior analyses was performed and the parametric bootstrap was applied to investigate the potential impact of positivity violations on results.

Baseline covariates, mutation profiles prior to treatment change, and viral response to therapy were collected for 401 treatment change episodes (TCEs) in which protease-inhibitor-experienced subjects initiated a new antiretroviral regimen containing the drug lopinavir. We focused on 2 target mutations in the protease enzyme: p82AFST and p82MLC (present in 25% and 1% of TCEs, respectively). The data for each target mutation consisted of $O = (W, A, Y)$, where A was a binary indicator that the target mutation was present prior to treatment change, W was a set of 35 baseline characteristics including summaries of past treatment history, mutations in the reverse transcriptase enzyme, and a genotypic susceptibility score for the background regimen (based on the Stanford scoring system). The outcome Y was the change in $\log_{10}$(viral load) following initiation of the new antiretroviral regimen. The target observed data parameter was $E_0[E_0(Y \mid A = 1, W) - E(Y \mid A = 0, W)]$, equal under (10.4) to the average treatment effect $E_0(Y_1 - Y_0)$.

Effect estimates were obtained for each mutation using the IPTW estimator and TMLE with a logistic fluctuation (Chap. 7). $\bar{Q}_0$ and g_0 were estimated with stepwise forward selection of main terms based on the AIC criterion. Estimators were

implemented using both unbounded values for $g_n(A \mid W)$ and values truncated at $[0.025, 0.975]$. Standard errors were estimated using the influence curve treating the values of g_n as fixed. The parametric bootstrap was used to estimate bias for each estimator using 1000 samples and the Bias_{ETA} algorithm.

Results for both mutations are presented in Table 10.4. p82AFST is known to be a major mutation for lopinavir resistance (Johnson et al. 2009). The current results support this finding; the IPTW and TMLE point estimates were similar and both suggested a significantly more positive change in viral load (corresponding to a less effective drug response) among subjects with the mutation as compared to those without it. The parametric-bootstrap-based bias estimate was minimal, raising no red flag that these findings might be attributable to positivity bias.

The role of mutation p82CLM is less clear based on existing knowledge; depending on the scoring system used it is either not considered a lopinavir resistance mutation, or given an intermediate lopinavir resistance score (http://hivdb.stanford.edu, Johnson et al. 2009). Initial inspection of the point estimates and standard errors in the current analysis would have suggested that p82CLM had a large and highly significant effect on lopinavir resistance. Application of the parametric-bootstrap-based diagnostic, however, would have suggested that these results should be interpreted with caution. In particular, the bias estimate for the unbounded TMLE was larger than the estimated standard error, while the bias estimate for the unbounded IPTW estimator was of roughly the same magnitude. While neither bias estimate was of sufficient magnitude relative to the point estimate to change inference, their size relative to the corresponding standard errors would have suggested that further investigation was warranted.

In response, the nonparametric bootstrap (based on 1000 bootstrap samples) was applied to provide an alternative estimate of the standard error. Using this alternative approach, the standard errors for the unbounded TMLE and IPTW estimator of the effect of p82MLC were estimated to be 2.77 and 1.17, respectively. Nonparametric-bootstrap-based standard error estimates for the bounded TMLE and IPTW estimator were lower (0.84 and 1.12, respectively), but still substantially higher than the initial naive standard error estimates. These revised standard error estimates dramatically changed interpretation of results, suggesting that the current analysis was unable to provide essentially any information on the presence, magnitude, or direction of the p82CLM effect. (Nonparametric-bootstrap-based standard error estimates for p82AFST were also somewhat larger than initial estimates but did not change inference).

In this example, Bias_{ETA} is expected to include some nonpositivity bias due to the curse of dimensionality. However, the resulting bias estimate should be interpreted as highly optimistic (i.e., as an underestimate of the true finite sample bias). The parametric bootstrap sampled from estimates of g_0 and $\bar{Q}_0$ that had been fit using the forward stepwise algorithm. This ensured that g_n and $\bar{Q}_n$ (which applied the same stepwise algorithm) would do a good job approximating $g_{\hat{P}_0}$ and $\bar{Q}_{\hat{P}_0}$ in each bootstrap sample. Clearly, no such guarantee exists for the true P_0. This simple example further illustrates the utility of the nonparametric bootstrap for standard error estimation in the setting of sparse data and positivity violations. In this particular

Table 10.4 Point estimate, standard error, and parametric-bootstrap-based bias estimates for the effect of two HIV resistance mutations on viral response

	TMLE			IPTW		
	$\hat{\beta}_{TMLE}$	$\widehat{\text{SE}}$	Bias_{ETA}	$\hat{\beta}_{IPTW}$	$\widehat{\text{SE}}$	Bias_{ETA}
p82AFST						
[0.000, 1.000]	0.65	0.13	−0.01	0.66	0.15	−0.01
[0.025, 0.975]	0.62	0.13	0.00	0.66	0.15	−0.01
p82MLC						
[0.000, 1.000]	2.85	0.14	−0.37	1.29	0.14	0.09
[0.025, 0.975]	0.86	0.10	−0.01	0.80	0.23	0.08

example, the improved variance estimate provided by the nonparametric bootstrap was sufficient to prevent positivity violations from leading to incorrect inference. As demonstrated in the simulations, however, in other settings even accurate variance estimates may fail to alert the analyst to threats posed by positivity violations.

10.4 Practical Approaches to Positivity Violations

Approach #1: Change the projection function $h(A, V)$. Throughout this chapter we have focused on the target causal parameter $\beta(F_X, m, h)$ defined according to (10.2) as the projection of the $E_{F_X}(Y_a \mid V)$ on the working marginal structural model $m(a, V \mid \beta)$. Choice of function $h(a, V)$ both defines the target parameter by specifying which values of (A, V) should be given greater weight when estimating β_0 and, by assumption (10.6), defines the positivity assumption needed for β_0 to be identifiable.

We have focused on parameters indexed by $h(a, V) = 1$, a choice that gives equal weight to estimating the counterfactual outcome for all values (a, v) (Neugebauer and van der Laan 2007). Alternative choices of $h(a, V)$ can significantly weaken the needed positivity assumption. For example, if the target of inference only involves counterfactual outcomes among some restricted range $[c, d]$ of possible values $\mathcal{A}$, defining $h(a, V) = I(a \in [c, d])$ weakens the positivity assumption by requiring sufficient variability only in the assignment of treatment levels within the target range. In some settings, the causal parameter defined by such a projection over a limited range of $\mathcal{A}$ might be of substantial a priori interest. For example, one may wish to focus estimation of a drug dose response curve only on the range of doses considered reasonable for routine clinical use, rather than on the full range of doses theoretically possible or observed in a given data set.

An alternative approach, commonly employed in the context of IPTW estimation and introduced in Sect. 10.2.2, is to choose $h(a, V) = g(a \mid V)$, where $g(a \mid V) = P(A = a \mid V)$ is the conditional probability of treatment given the covariates included in the marginal structural model. In the setting of IPTW estima-

tion this choice corresponds to the use of stabilizing weights, a common approach to reducing both the variance of the IPTW estimator in the face of sparsity (Robins et al. 2000a). When the target causal parameter is defined using a marginal structural working model, use of $h(a, V) = g(a, V)$ corresponds to a decision to define a target parameter that gives greater weight to those regions of the joint distribution of (A, V) that are well supported and that relies on smoothing or extrapolation to a greater degree in areas that are not (Neugebauer and van der Laan 2007).

Use of a marginal structural working model makes clear that the utility of choosing $h(a, V) = g(a \mid V)$ as a method to approach data sparsity is not limited to the IPTW estimator. Recall that MLE can be implemented by regressing predicted values for Y_a on (a, V) according to model $m(a, V \mid \beta)$ with weights provided by $h(a, V)$. When the projection function is chosen to be $g_0(a \mid V)$, this corresponds to a weighted regression in which weights are proportional to the degree of support in the data.

Even when one is ideally interested in the entire causal curve [implying a target parameter defined by choice $h(a, V) = 1$], specification of alternative choices for h offers a means of improving identifiability, at a cost of redefining the target parameter. For example, one can define a family of target parameters indexed by $h_\delta(a, V) = I(a \in [c(\delta), d(\delta)])$, where an increase in δ corresponds to progressive restriction on the range of treatment levels targeted by estimation. Fluctuation of δ thus corresponds to trading a focus on more limited areas of the causal curve for improved parameter identifiability. Selection of the final target from among this family can be based on an estimate of bias provided by the parametric bootstrap. For example, the bootstrap can be used to select the parameter with the smallest δ below some prespecified threshold for allowable Bias_{ETA}.

Approach #2: Restrict the adjustment set. Exclusion of problematic Ws (i.e., those covariates resulting in positivity violations or near-violations) from the adjustment set provides a means to trade confounding bias for a reduction in positivity violations (Bembom et al. 2008). In some cases, exclusion of covariates from the adjustment set may come at little or no cost to bias in the estimate of the target parameter. In particular, a subset of W that excludes covariates responsible for positivity violations may still be sufficient to control for confounding. In other words, a subset $W' \subset W$ may exist for which both identifying assumptions (10.4) and (10.5) hold [i.e., $Y_a \coprod A \mid W'$ and $g_0(a \mid W') > 0, a \in \mathcal{A}$], while positivity fails for the full set of covariates. In practice, this approach can be implemented by first determining candidate subsets of W under which the positivity assumption holds, and then using causal graphs to assess whether any of these candidates is sufficient to control for confounding. Even when no such candidate set can be identified, background knowledge (or sensitivity analysis) may suggest that problematic Ws represent a minimal source of confounding bias (Moore et al. 2009). Often, however, those covariates that are most problematic from a positivity perspective are also strong confounders.

As suggested with respect to the choice of projection function $h(a, V)$ in the previous section, the causal effect estimator can be fine-tuned to select the degree of restriction on the adjustment set W according to some prespecified rule for elim-

inating covariates from the adjustment set, and the parametric bootstrap used to select the minimal degree of restriction that maintains Bias_{ETA} below an acceptable threshold (Bembom et al. 2008). In the case of substantial positivity violations, such an approach can result in small covariate adjustment sets. While such limited covariate adjustment accurately reflects a target parameter that is poorly supported by the available data, the resulting estimate can be difficult to interpret and will no longer carry a causal intepretation.

Approach #3: Restrict the sample. An alternative, sometimes referred to as "trimming," discards classes of subjects for whom there exists no or limited variability in observed treatment assignment. A causal effect is then estimated in the remaining subsample. This approach is popular in econometrics and social science (Crump et al. 2006; LaLonde 1986; Heckman et al. 1997; Dehejia and Wahba 1999).

When the subset of covariates responsible for positivity violations is low- or one-dimensional, such an approach can be implemented simply by discarding subjects with covariate values not represented in all treatment groups. For example, say that one aims to estimate the average effect of a binary treatment and, in order to control for confounding, one needs to adjust for W, a covariate with possible levels $\{1, 2, 3, 4\}$. However, inspection of the data reveals that no one in the sample with $W = 4$ received treatment [i.e., $g_n(1 \mid W = 4) = 0$]. The sample can be trimmed by excluding those subjects for whom $W = 4$ prior to applying a given causal effect estimator for the average treatment effect. As a result, the target parameter is shifted from $E_0(Y_1 - Y_0)$ to $E_0(Y_1 - Y_0 \mid W < 4)$, and positivity assumption (10.5) now holds (as $W = 4$ occurs with zero probability).

Often W is too high-dimensional to make this straightforward implementation feasible; in such a case matching on the propensity score provides a means to trim the sample. There is an extensive literature on propensity score-based effect estimators; however, such estimators are beyond the scope of the current review. Several potential problems arise with the use of trimming methods to address positivity violations. First, discarding subjects responsible for positivity violations shrinks sample size and thus runs the risk of increasing the variance of the effect estimate. Further, sample size and the extent to which positivity violations arise by chance are closely related. Depending on how trimming is implemented, new positivity violations can be introduced as sample size shrinks. Second, restriction of the sample may result in a causal effect for a population of limited interest. In other words, as can occur with alternative approaches to improving identifiability by shifting the target of inference, the parameter actually estimated may be far from the initial target. Further, when the criterion used to restrict the sample involves a summary of high-dimensional covariates, such as is provided the propensity score, it can be difficult to interpret the parameter estimated. Finally, when treatment is longitudinal, the covariates responsible for positivity violations may themselves be affected by past treatment. Trimming to remove positivity violations in this setting amounts to conditioning on posttreatment covariates and can thus introduce new bias.

Crump proposes an approach to trimming that falls within the general strategy of redefining the target parameter in order to explicitly capture the tradeoff between

parameter identifiability and proximity to the initial target (Crump et al. 2006). In addition to focusing on the treatment effect in an a priori specified target population, he defines an alternative target parameter corresponding to the average treatment effect in that subsample of the population for which the most precise estimate can be achieved. Crump further suggests the potential for extending this approach to achieve an optimal (according to some user-specified criterion) tradeoff between the representativeness of the subsample in which the effect is estimated and the variance of the estimate.

Approach #4: Change the intervention of interest. A final alternative for improving the identifiability of a causal parameter in the presence of positivity violations is to redefine the intervention of interest. Realistic rules rely on an estimate of the propensity score $g_0(a \mid W)$ to define interventions that explicitly avoid positivity violations. This ensures that the causal parameter estimated is sufficiently supported by existing data.

Realistic interventions avoid positivity violations by first identifying subjects for whom a given treatment assignment is not realistic (i.e., subjects whose propensity score for a given treatment is small or zero) and then assigning an alternative treatment with better data support to those individuals. Such an approach is made possible by focusing on the causal effects of dynamic treatment regimes (van der Laan and Petersen 2007a; Robins et al. 2008). The causal parameters described thus far are summaries of the counterfactual outcome distribution under a fixed treatment applied uniformly across the target population. In contrast, a dynamic regime assigns treatment in response to patient covariate values. This characteristic makes it possible to define interventions under which a subject is only assigned treatments that are possible (or "realistic") given a subject's covariate values.

To continue the previous example in which no subjects with $W = 4$ were treated, a realistic treatment rule might take the form "treat only those subjects with W less than 4." More formally, let $d(W)$ refer to a treatment rule that deterministically assigns a treatment $a \in \mathcal{A}$ based on a subject's covariates W and consider the rule $d(W) = I(W < 4)$. Let Y_d denote the counterfactual outcome under the treatment rule $d(W)$, which corresponds to treating a subject if and only if his or her covariate W is below 4. In this example $E_0(Y_0)$ is identified as $\sum_w E_0(Y \mid W = w, A = 0)P_0(W = w)$; however, since $E_0(Y \mid W = w, A = 1)$ is undefined for $W = 4$, $E_0(Y_1)$ is not identified (unless we are willing to extrapolate based on $W < 4$). In contrast, $E_0(Y_d)$ is identified by the nonparametric g-formula: $\sum_w E_0(Y = y \mid W = w, A = d(W))P_0(W = w)$. Thus the average treatment effect $E_0(Y_d - Y_0)$, but not $E_0(Y_1 - Y_0)$, is identified. The redefined causal parameter can be interpreted as the difference in expected counterfactual outcome if only those subjects with $W < 4$ were treated as compared to the outcome if no one were treated.

More generally, realistic rules indexed by a given static treatment a assign a only to those individuals for whom the probability of receiving a is greater than some user-specified probability α (such as $\alpha > 0.05$). Let $d(a, W)$ denote the rule indexed by static treatment a. If A is binary, then $d(1, W) = 1$ if $g(1 \mid W) > \alpha$, otherwise $d(1, W) = 0$. Similarly, $d(0, W) = 0$ if $g(0 \mid W) > \alpha$; otherwise $d(0, W) = 1$. Real-

istic causal parameters are defined as some parameter of the distribution of $Y_{d(a,W)}$ (possibly conditional on some subset of baseline covariates $V \subset W$). Estimation of the causal effects of dynamic rules $d(W)$ allows the positivity assumption to be relaxed to $g(d(W) \mid W) > 0$, a.e (i.e., only those treatments that would be assigned based on rule d to patients with covariates W need to occur with positive probability within strata of W). Realistic rules $d(a, W)$ are designed to satisfy this assumption by definition. When a given treatment level a is unrealistic [i.e., when $g(a \mid W) < \alpha$], realistic rules assign an alternative from among viable (well-supported) choices. The choice of an alternative is straightforward when treatment is binary. When treatment has more than two levels, however, a rule for selecting the alternative treatment level is needed. One option is to assign a treatment level that is as close as possible to the original assignment while still remaining realistic. For example, if high doses of drugs occur with low probability in a certain subset of the population, a realistic rule might assign the maximum dose that occurs with probability $> \alpha$ in that subset. An alternative class of dynamic regimes, referred to as "intent-to-treat" rules, instead assigns a subject to his or her observed treatment value if an initial assignment is deemed unrealistic. Moore et al. (2009) and Bembom and van der Laan (2007a) provide illustrations of these types of realistic rules using simulated and real data.

The causal effects of realistic rules clearly differ from their static counterparts. The extent to which the new target parameter diverges from the initial parameter of interest depends on both the extent to which positivity violations occur in the finite sample (i.e., the extent of support available in the data for the initial target parameter) and on a user-supplied threshold α. The parametric bootstrap approach presented in Sect. 10.3 can be employed to data-adaptively select α based on the level of Bias_{ETA} deemed acceptable (Bembom and van der Laan 2007a).

Selection among a family of parameters. Each of the methods described for estimating causal effects in the presence of data sparsity corresponds to a particular strategy for altering the target parameter in exchange for improved identifiability. In each case, we have outlined how this tradeoff could be made systematically based on some user-specified criterion such as the bias estimate provided by the parametric bootstrap. We now summarize this general approach in terms of a formal method for estimation in the face of positivity violations.

1. Define a family of parameters. The family should include the initial target of inference together with a set of related parameters, indexed by γ in index set I, where γ represents the extent to which a given family member trades improved identifiability for decreased proximity to the initial target. In the examples given in the previous section, γ could be used to index a set of projection functions $h(a, V)$ based on an increasingly restrictive range of the possible values $\mathcal{A}$, degree to which the adjustment covariate set or sample is restricted, or choice of a threshold for defining a realistic rule.
2. Apply the parametric bootstrap to generate an estimate Bias_{ETA} for each $\gamma \in I$. In particular, this involves estimating the data-generating distribution, simulating new data from this estimate, and then applying an estimator to each target indexed by γ.

3. Select the target parameter from among the set that falls below a prespecified threshold for acceptable Bias_{ETA}. In particular, select the parameter from within the set that is indexed by the value γ that corresponds to the greatest proximity to the initial target.

This approach allows an estimator to be defined in terms of an algorithm that identifies and estimates the parameter within a candidate family that is as close to the initial target of inference as possible while remaining within some user-supplied limit on the extent of tolerable positivity violations.

10.5 Discussion

The identifiability of causal effects relies on sufficient variation in treatment assignment within covariate strata. The strong version of positivity requires that each possible treatment occur with positive probability in each covariate stratum; depending on the model and target parameter, this assumption can be relaxed to some extent. In addition to assessing identifiability based on measurement of and control for sufficient confounders, data analyses should directly assess threats to identifiability based on positivity violations. The parametric bootstrap is a practical tool for assessing such threats, and provides a quantitative estimator-specific estimate of bias arising due to positivity violations.

This chapter has focused on the positivity assumption for the causal effect of a treatment assigned at a single time point. Extension to a longitudinal setting in which the goal is to estimate the effect of multiple treatments assigned sequentially over time introduces considerable additional complexity. First, practical violations of the positivity assumption can arise more readily in this setting. Under the longitudinal version of the positivity assumption the conditional probability of each possible treatment history should remain positive regardless of covariate history. However, this probability is the product of time-point-specific treatment probabilities given the past. When the product is taken over multiple time points, it is easy for treatment histories with very small conditional probabilities to arise. Second, longitudinal data make it harder to diagnose the bias arising due to positivity violations. Implementation of the parametric bootstrap in longitudinal settings requires Monte Carlo simulation both to implement the MLE and to generate each bootstrap sample. In particular, this requires estimating and sampling from the time-point-specific conditional distributions of all covariates and treatment given the past. Additional research on assessing the impact of positivity bias on longitudinal causal parameters is needed, including investigation of the parametric bootstrap in this setting.

When positivity violations occur for structural reasons rather than due to chance, a causal parameter that avoids these positivity violations will often be of substantial interest. For example, when certain treatment levels are contraindicated for certain types of individuals, the average treatment effect in the population may be of less interest than the effect of treatment among that subset of the population without contraindications, or, alternatively, than the effect of an intervention that assigns

treatment only to those subjects without contraindications. Similarly, the effect of a multilevel treatment may be of greatest interest for only a subset of treatment levels.

In other cases researchers may be happy to settle for a better estimate of a less interesting parameter. Sample restriction, estimation of realistic parameters, and change in projection function $h(a, V)$ all change the causal effect being estimated; in contrast, restriction of the covariate adjustment set often results in estimation of a noncausal parameter. However, all of these approaches can be understood as means to shift from a poorly identified initial target toward a parameter that is less ambitious but more fully supported by the available data. The new estimand is not determined a priori by the question of interst, but rather is driven by the observed data distribution in the finite sample at hand. There is thus an explicit tradeoff between identifiability and proximity to the initial target of inference. Ideally, this tradeoff will be made in a systematic way rather than on an ad hoc basis at the discretion of the investigator. Definition of an estimator that selects among a family of parameters according to some prespecified criteria is a means to formalize this tradeoff. An estimate of bias based on the parametric bootstrap can be used to implement the tradeoff in practice.

In summary, we offer the following advice for applied analyses. First, define the causal effect of interest based on careful consideration of structural positivity violations. Second, consider estimator behavior in the context of positivity violations when selecting an estimator. Third, apply the parametric bootstrap to quantify the extent of estimator bias under data simulated to approximate the true data-generating distribution. Fourth, when positivity violations are a concern, choose an estimator that selects systematically from among a family of parameters based on the tradeoff between data support and proximity to the initial target of inference.

10.6 Notes and Further Reading

While perhaps less well-recognized than confounding bias, violations and near-violations of the positivity assumption can increase both the variance and bias of causal effect estimates, and if undiagnosed can seriously threaten the validity of causal inference. The dangers of causal effect estimation in the absence of adequate data support have long been understood (Cochran 1957). More recent causal inference literature refers to the need for adequate exposure variability within confounder strata as the assumption of positivity or experimental treatment assignment (Robins 1986, 1987a, 2000). A summary of estimator behavior in the face of positivity violations is also discussed in previous work (Neugebauer and van der Laan 2005, 2007; Bembom and van der Laan 2007a; Moore et al. 2009; Cole and Hernan 2008). Additional simulations are discussed in Petersen et al. (2010), the article from which this chapter was adapted.