A Distributional Approach for Causal Inference Using Propensity Scores

Zhiqiang TAN

Drawing inferences about the effects of treatments and actions is a common challenge in economics, epidemiology, and other fields. We adopt Rubin's potential outcomes framework for causal inference and propose two methods serving complementary purposes. One can be used to estimate average causal effects, assuming no confounding given measured covariates. The other can be used to assess how the estimates might change under various departures from no confounding. Both methods are developed from a nonparametric likelihood perspective. The propensity score plays a central role and is estimated through a parametric model. Under the assumption of no confounding, the joint distribution of covariates and each potential outcome is estimated as a weighted empirical distribution. Expectations from the joint distribution are estimated as weighted averages or, equivalently to first order, regression estimates. The likelihood estimator is at least as efficient and the regression estimator is at least as efficient and robust as existing estimators. Regardless of the no-confounding assumption, the marginal distribution of covariates times the conditional distribution of observed outcome given each treatment assignment and covariates is estimated. For a fixed bound on unmeasured confounding, the marginal distribution of covariates times the conditional distribution of counterfactual outcome given each treatment assignment and covariates is explored to the extreme and then compared with the composite distribution corresponding to observed outcome given the same treatment assignment and covariates. We illustrate the methods by analyzing the data from an observational study on right heart catheterization.

KEY WORDS: Causal inference; Control variate; Nonparametric likelihood; Observational study; Propensity score; Sensitivity analysis.

1. INTRODUCTION

Drawing inferences about the effects of treatments and actions is a common challenge in economics, epidemiology, and other fields. Although randomized experiments remain the gold standard for research, observational studies are often necessary due to ethical or practical considerations. In observational data, systematic differences can exist between treated and untreated groups with respect to various covariates, and direct comparisons of observed outcomes from the two groups are not appropriate. Methods for defining and estimating causal effects from observational data are complicated and even controversial. In this article we adopt Rubin's causal model and contribute two methods for estimation and sensitivity analysis from a nonparametric likelihood perspective.

Rubin's (1974, 1977, 1978) causal model has become widely accepted as a framework for causal inference in both experiments and observational studies; this model is summarized in Section 2. In that framework, a number of basic concepts and assumptions are formalized. Causal effects are defined as comparisons of potential outcomes that would be observed under different treatments. For each subject, only one potential outcome can actually be observed, depending on which treatment is assigned. Therefore, the mechanism of treatment assignment plays a crucial role in causal inference. Randomization is an assignment mechanism that allows causal effects to be estimated straightforwardly. However, in an observational study, the assignment mechanism is unknown, and inferences about causal effects necessarily rely on some assumptions about it.

First, consider the assumption of no confounding, in which treatment assignment and potential outcomes are independent given measured covariates. Under this assumption, the average causal effect can be consistently estimated, and various methods have been proposed for doing so. Some methods focus on the relationship between covariates and potential outcomes, and others work with the relationship between covariates and treatment assignment or the propensity score (Rosenbaum and Rubin 1983a). These approaches both rely on the assumption of no confounding but make different modeling assumptions. The first approach requires a correctly specified outcome regression model, whereas the second requires a correctly specified propensity score model. There are also methods that combine outcome regression with propensity score matching, subclassification, or weighting (see Imbens 2004 for a review). In particular, Robins, Rotnitzky, and others introduced a class of estimators under a propensity score model and derived optimal estimators for when an outcome regression model is also correctly specified. The optimal estimators and several variants are locally efficient; that is, they achieve the semiparametric variance bound under the propensity score model if both the propensity score model and the outcome regression model are correct. Moreover, some of them are doubly robust; that is, remain consistent and asymptotically normal if either the propensity score model or the outcome regression model is correct. (See van der Laan and Robins 2003 for a theory of doubly robust locally efficient estimation.)

All existing propensity score methods are based on estimating equations. Robins and Ritov (1997) pointed out that any method based on the usual likelihood should not depend on the propensity score. In Section 3 we propose a likelihood formulation for propensity score weighting by ignoring part of all information about the joint distributions of covariates and potential outcomes. The idea is connected to Kong, McCullagh, Meng, Nicolae, and Tan's (2003) formulation for Monte Carlo integration by ignoring part of all information about the baseline measure. We derive a nonparametric likelihood estimator under a propensity score model and suggest a closed-form regression estimator as a first-order approximation. We also establish that (a) the likelihood estimator is locally efficient and the regression estimator is locally efficient and doubly robust, and (b) the

Zhiqiang Tan is Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ztan@jhsph.edu*). This research is supported by a Faculty Innovation Fund from the school. The author thanks Tom Louis, James Robins, and Dan Scharfstein for valuable discussions, Constantine Frangakis and Chuck Rhode for helpful comments, and the editor and an associate editor for handling the review.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Theory and Methods
DOI 10.1198/016214506000000023

likelihood estimator is at least as efficient and the regression estimator is at least as efficient and robust as existing estimators. In case (b), efficiency is evaluated when the propensity score model is correct but the outcome regression model is misspecified, and robustness is evaluated when the propensity score model is misspecified. We present a simulation study to compare the behaviors of different estimators in medium samples. Our estimators have the overall smallest mean squared errors under various settings in the study.

For data analysis, the propensity score subclassification method was originally illustrated by Rosenbaum and Rubin (1984). The process of fitting the propensity score by logit regression, checking the balance of measured covariates within subclasses using simple statistical methods (such as analysis of variance and bar plot), and refining the fitted propensity score is attractive to a broad audience. Our method follows a similar process but does not rely on subclassification (to create approximately homogeneous subclasses in the propensity score), which still lacks a complete frequentist theory. For a fitted propensity score, the joint distribution of covariates and each potential outcome is estimated. Only if the propensity score model is correctly specified are the two marginal distributions of covariates asymptotically the same and does the difference between the expectations of two potential outcomes indicate the average causal effect. Our method takes into account estimation of the propensity score in calculating standard errors by the frequentist theory that we establish. In Section 5 we illustrate our method by analyzing the data from an observational study on right heart catheterization.

Second, the assumption of no confounding requires that all covariates relevant to both treatment assignment and potential outcomes be measured. Although care can be taken to identify many such covariates, there is always a possibility of overlooking some in an observational study. In this case it is desirable to conduct a sensitivity analysis to assess how the estimates might change under various departures from no confounding. Following Cornfield et al. (1959), a common approach to sensitivity analysis is to investigate the consequences of leaving out a relevant but unmeasured covariate. Some methods postulate how this covariate affects potential outcomes (Lin, Psaty, and Kronmal 1998; Rosenbaum 1986), some postulate how this covariate influences treatment assignment (Rosenbaum 2002a), and others postulate how this covariate is related to both treatment assignment and potential outcomes (Imbens 2003; Rosenbaum and Rubin 1983b). An alternative approach that does not involve any unmeasured covariate is to postulate how potential outcomes are associated with treatment assignment (Brumback, Hernan, Haneuse, and Robins 2004; Robins 1999), or how treatment assignment depends on potential outcomes (Robins, Rotnitzky, and Scharfstein 1999).

Each sensitivity analysis method assumes a model that describes unmeasured confounding and contains no confounding as a special case. However, the model itself is susceptible to misspecification (see Rotnitzky, Scharfstein, Su, and Robins 2001, sec. 5, for a discussion). What information can be learned from observed data regardless of the no-confounding assumption? How can the information be exploited for a robust sensitivity analysis? In Section 4 we show from a nonparametric likelihood perspective that the estimation method in Section 3 extends to general confounding cases. For a fitted propensity score, the marginal distribution of covariates times the conditional distribution of observed outcome given each treatment assignment and covariates is estimated. In contrast, the conditional distribution of counterfactual outcome given each treatment assignment and covariates is not identifiable from observed data. The assumption of no confounding equates the two conditional distributions. Next, unmeasured confounding can be characterized through the density ratio of each potential outcome given different treatment assignments or, equivalently, the odds ratio of receiving the treatment given different values of each potential outcome. We consider a sensitivity analysis model that places bounds on the density ratio or odds ratio and propose a nonparametric method for obtaining conservative bounds on the expectation of counterfactual outcome given each treatment assignment and covariates and then marginalized over covariates. In Section 5 we apply our method to the right heart catheterization study.

2. CAUSAL MODEL

We adopt Rubin's (1974, 1977, 1978) causal model. Let $\Omega = \{\omega\}$ be a population endowed with a probability measure P. If Ω is finite with N units, then define P to be uniform placing mass N^{-1} at each unit. Let $\{0, 1\}$ be a treatment set, where 0 represents the control treatment ("control") and 1 represents the active treatment ("treatment").

Potential Outcomes and Covariates. Let $Y_0 = Y_0(\omega)$ be the response that would be observed if unit ω received treatment 0 and let $Y_1 = Y_1(\omega)$ be the response that would be observed if unit ω received treatment 1. The two variables are called potential outcomes (Neyman 1923; Rubin 1974). Assume that there is no interference between different units; that is, the potential outcomes of a unit are independent of which treatments other units receive (Cox 1958; Rubin 1980). In addition, let $\mathbf{X} = \mathbf{X}(\omega)$ be a vector of measured covariates whose values are not changed by either treatment.

Defining Causal Effects. Holland and Rubin (1988) distinguished three levels of causal inferences: unit level, subpopulation level, and population level. The definition of causal effect at the unit level is a comparison of $Y_0(\omega)$ and $Y_1(\omega)$, typically the difference $Y_1(\omega) - Y_0(\omega)$. Subpopulations can be classified by the values of covariates. The average causal effect over a subpopulation $\{\omega : \mathbf{X}(\omega) = \mathbf{x}\}$ is $E(Y_1|\mathbf{X} = \mathbf{x}) - E(Y_0|\mathbf{X} = \mathbf{x})$. The average causal effect over the population Ω is $E(Y_1) - E(Y_0)$. The three levels are ordered by decreasing strength in the sense that knowledge of all unit-level causal inferences implies knowledge of all subpopulation-level causal inferences, and knowledge of all subpopulation-level causal inferences for a partition of Ω implies knowledge of population-level causal inferences, but not vice versa.

Assigning Treatments. Imagine that treatment assignment is done before selecting units for a study. Let $T = T(\omega)$ be the binary variable taking value 0 or 1 if unit ω receives treatment 0 or 1. The conditional distribution $P(T|\mathbf{X}, Y_0, Y_1)$ is called assignment mechanism. An important class of assignment mechanisms is the class of unconfounded assignment mechanisms

defined by $P(T|\mathbf{X}, Y_0, Y_1) = P(T|\mathbf{X})$ or, in the notation of conditional independence, $T \perp (Y_0, Y_1)|\mathbf{X}$. That is, treatment assignment T and potential outcomes (Y_0, Y_1) are independent conditionally on covariates $\mathbf{X}$. A probability assignment mechanism is defined by $0 < P(T = 1|\mathbf{X}, Y_0, Y_1) < 1$, so that each unit has a positive probability of receiving either treatment. For technical convenience, the bounds are often assumed to be δ and $1 - \delta$ for a small number $\delta > 0$.

Recording Data. There are at least two sources of missingness. First, either $Y_0(\omega)$ or $Y_1(\omega)$, but not both, can actually be observed, depending on the value of $T(\omega)$. Denote by $Y = Y(\omega)$ the observed outcome $(1 - T(\omega))Y_0(\omega) + T(\omega)Y_1(\omega)$. This inherent fact of observational life is called the fundamental problem of causal inference (Holland 1986). Second, some elements in $\mathbf{X}(\omega)$ and $Y(\omega)$ can be missing. For simplicity, we throughout the article assume that the second source is absent.

Selecting Units. In a study, a sample $\{\omega_1, \ldots, \omega_n\}$ is selected from the population Ω. Basic sampling designs include random sampling from the population Ω, random sampling from the treated group $\{\omega : T(\omega) = 1\}$ and the untreated group $\{\omega : T(\omega) = 0\}$, and random sampling from the case group $\{\omega : Y(\omega) = 1\}$ and the referent group $\{\omega : Y(\omega) = 0\}$ when (Y_0, Y_1) are dichotomous. Here we focus on the first design and assume that $\{\omega_1, \ldots, \omega_n\}$ is an independent and identically distributed (iid) sample.

In the original work of Rubin (1978), the population Ω is finite, and selecting units appears to precede assigning treatments and recording data. Our view presented earlier is slightly different and tailored toward an infinite population.

3. NO–CONFOUNDING ESTIMATION

Let $\{\omega_1, \ldots, \omega_n\}$ be an iid sample. The data $(\mathbf{X}_i, Y_i, T_i) = (\mathbf{X}(\omega_i), Y(\omega_i), T(\omega_i))$ are iid from the joint distribution of $(\mathbf{X}, Y, T)$. Label the sample such that $T_i = 1$ for $i = 1, \ldots, n_1$ and $=0$ for $i = n_1 + 1, \ldots, n$. Our task is to estimate $\mu_0 = E(Y_0)$ and $\mu_1 = E(Y_1)$, whose difference gives the average causal effect.

Assume that the assignment mechanism is unconfounded: $T \perp (Y_0, Y_1)|\mathbf{X}$. The likelihood of $\{(\mathbf{X}_i, Y_i, T_i)\}$ is

$$L_1 \times L_2 = \prod_{i=1}^{n} \left[(1 - \pi(\mathbf{X}_i))^{1-T_i} \pi(\mathbf{X}_i)^{T_i}\right] \times \prod_{i=1}^{n} \left[G_0(\{\mathbf{X}_i, Y_i\})^{1-T_i} G_1(\{\mathbf{X}_i, Y_i\})^{T_i}\right],$$

where $\pi(\mathbf{X})$ is the propensity score $P(T = 1|\mathbf{X})$, G_0 is the joint distribution of $(\mathbf{X}, Y_0)$, and G_1 is the joint distribution of $(\mathbf{X}, Y_1)$. The distribution G_0 or G_1 can be further factorized as the marginal distribution $P(\mathbf{X})$ times the conditional distribution $P(Y_0|\mathbf{X})$ or $P(Y_1|\mathbf{X})$, but such a factorization is avoided here. The likelihood is a product of two factors, one factor, L_1, involving π only and the other factor, L_2, involving (G_0, G_1) only. By definition, G_0 and G_1 induce the same marginal distribution on the covariate space $\mathcal{X}$. Equivalently, G_0 and G_1 satisfy

$$\int h(\mathbf{x})\, \mathrm{d}G_0(\mathbf{x}, y_0) = \int h(\mathbf{x})\, \mathrm{d}G_1(\mathbf{x}, y_1) \tag{1}$$

for each bounded function h on $\mathcal{X}$.

At this stage, the model is saturated or nonparametric; there is no additional restriction on either π or (G_0, G_1). Parametric submodels can be specified for the regression functions $E(Y_t|\mathbf{X})$, the propensity score $P(T = 1|\mathbf{X})$, or both. Consider the outcome regression model (model R)

$$E(Y_t|\mathbf{X}) = \Psi(\boldsymbol{\alpha}_t^\top \mathbf{g}(\mathbf{X})),$$

where Ψ is a link function, $\mathbf{g} = (1, g_1, \ldots, g_k)^\top$ is a vector of known functions including the constant, and $\boldsymbol{\alpha}_t = (\alpha_{t0}, \alpha_{t1}, \ldots, \alpha_{tk})^\top$ is a vector of parameters $(t = 0, 1)$. The model can be fit by maximum quasi-likelihood, and $E(Y_t) = E[E(Y_t|\mathbf{X})]$ can be estimated by

$$\hat{\mu}_{\mathrm{OR}} = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y_t|\mathbf{X}_i).$$

Alternatively, consider the propensity score model (model S)

$$P(T = 1|\mathbf{X}) = \Pi(\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{X})),$$

where Π is a link function, $\mathbf{f} = (f_1, \ldots, f_l)^\top$ is a vector of known functions, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_l)^\top$ is a vector of parameters. The model can be fit by maximum likelihood, and $E(Y_t)$ can be estimated by the inverse probability weighted (IPW) estimator

$$\hat{\mu}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{T_i = t\}}{\hat{P}(T = t|\mathbf{X}_i)} Y_i.$$

Write the fitted propensity score as $\pi(\mathbf{X}; \hat{\boldsymbol{\gamma}})$. Finally, some estimators use submodels for both the propensity score and the regression functions. We describe a number of such estimators proposed by Robins, Rotnitzky, and others in related missing-data problems. For simplicity, let the estimand be μ_1. Robins, Rotnitzky, and Zhao (1994) proposed the augmented IPW estimator

$$\hat{\mu}_{\mathrm{AIPW,fix}} = \hat{\mu}_{\mathrm{IPW}} - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})} - 1 \right) \Psi(\hat{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X}_i)),$$

and Robins et al. (1995) further considered the estimator

$$\hat{\mu}_{\mathrm{AIPW,est}} = \hat{\mu}_{\mathrm{IPW}} - \hat{\boldsymbol{\beta}}_1^\top \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})} - 1 \right) \Psi(\hat{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X}_i)),$$

where $\hat{\boldsymbol{\beta}}_1$ is the regression coefficient of $\pi^{-1}(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})\, T_i Y_i$ on $(\pi^{-1}(\mathbf{X}_i; \hat{\boldsymbol{\gamma}}) T_i - 1)\Psi(\hat{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X}_i))$ based on asymptotic variance and covariance. Scharfstein, Rotnitzky, and Robins (1999, rejoinder) suggested the estimator

$$\hat{\mu}_{\mathrm{OR,ext}} = \frac{1}{n} \sum_{i=1}^{n} \Psi\left(\tilde{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X}_i) + \tilde{\kappa} \frac{1}{\pi(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})} \right),$$

where $(\tilde{\boldsymbol{\alpha}}_1, \tilde{\kappa})$ solves $\mathbf{0} = \sum_{i=1}^{n} T_i (\mathbf{g}(\mathbf{X}_i), \pi^{-1}(\mathbf{X}_i; \hat{\boldsymbol{\gamma}}))^\top \times (Y_i - \Psi(\boldsymbol{\alpha}_1^\top \mathbf{g}(\mathbf{X}_i) + \kappa \pi^{-1}(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})))$ for the extended model $E(Y|T = 1, \mathbf{X}) = \Psi(\boldsymbol{\alpha}_1^\top \mathbf{g}(\mathbf{X}) + \kappa \pi^{-1}(\mathbf{X}; \hat{\boldsymbol{\gamma}}))$. Rotnitzky and Robins (1995) proposed the estimator

$$\hat{\mu}_{\mathrm{IPW,ext}} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\pi(\mathbf{X}_i; \tilde{\boldsymbol{\gamma}}, \tilde{\theta})} Y_i,$$

where $(\tilde{\boldsymbol{\gamma}}, \tilde{\theta})$ solves $\mathbf{0} = \sum_{i=1}^{n} (\mathbf{f}(\mathbf{X}_i), \pi^{-1}(\mathbf{X}_i; \hat{\boldsymbol{\gamma}})\Psi(\hat{\boldsymbol{\alpha}}_1^\top \times \mathbf{g}(\mathbf{X}_i)))^\top (T_i - \pi(\mathbf{X}_i; \boldsymbol{\gamma}, \theta))$ for the extended model $P(T = 1|\mathbf{X}) = \Pi(\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{X}) + \theta\pi^{-1}(\mathbf{X}; \hat{\boldsymbol{\gamma}})\Psi(\hat{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X})))$. Robins, Rotnitzky, and Bonetti (2001) and Robins (2002a) suggested the estimator

$$\hat{\mu}_{\mathrm{IPW,lim}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(\mathbf{X}_i; \tilde{\boldsymbol{\gamma}}_\infty, \tilde{\theta}_\infty)} Y_i,$$

where $(\tilde{\boldsymbol{\gamma}}_\infty, \tilde{\theta}_\infty)$ is the limit of $(\tilde{\boldsymbol{\gamma}}_k, \tilde{\theta}_k)$ as $k \to \infty$ and $(\tilde{\boldsymbol{\gamma}}_k, \tilde{\theta}_k)$ solves a similar estimating equation for the model $P(T = 1|\mathbf{X}) = \Pi(\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{X}) + \theta\pi^{-1}(\mathbf{X}; \tilde{\boldsymbol{\gamma}}_{k-1}, \tilde{\theta}_{k-1})\Psi(\hat{\boldsymbol{\alpha}}_1^\top \mathbf{g}(\mathbf{X})))$ with $\pi(\mathbf{X}; \tilde{\boldsymbol{\gamma}}_0, \tilde{\theta}_0) = \pi(\mathbf{X}; \hat{\boldsymbol{\gamma}})$ for $k \geq 1$.

All existing propensity score methods are based on estimating equations. The estimator $\hat{\mu}_{\mathrm{OR,ext}}$ appears to be likelihood-based, but the extended model is not supposed to describe the data-generating process or be used to evaluate the estimator (see Robins, Rotnitzky, and van der Laan 2000 for a related discussion). Robins and Ritov (1997) examined implications of the likelihood principle in designed studies with missing data. Here the likelihood is factorized in terms of π and (G_0, G_1), and the estimands $E(Y_0)$ and $E(Y_1)$ are functions of (G_0, G_1) only. The (strict) likelihood principle implies that inference should be the same whatever the knowledge of π. Therefore, likelihood inference can make use of the propensity score π only if some information about the joint distributions (G_0, G_1) is ignored. We develop a nonparametric likelihood method using propensity scores by making explicit what information is ignored and what is retained, and further suggest a closed-form regression estimator that is at least as efficient and robust as existing estimators in the literature.

3.1 Known Propensity Score

To motivate ideas, we treat the case where our knowledge about the propensity score is exact and expressed as a function π^*, referred to as model S0. The case of parametric propensity score is treated in the next section.

Assume that model S0 is correct: π^* agrees with the underlying propensity score π. For likelihood inference, we first want to maximize

$$L_2 = \prod_{i=1}^{n} \left[G_0(\{\mathbf{X}_i, Y_i\})^{1-T_i} G_1(\{\mathbf{X}_i, Y_i\})^{T_i}\right]$$

over distributions (G_0, G_1) with the same marginal on $\mathcal{X}$. But if all constraints (1) were included, then inference would not depend on the propensity score. Robins and Ritov (1997) showed that no such estimator can be uniformly consistent or attain an algebraic rate of convergence if $\mathbf{X}$ has a continuous component. We choose to retain finitely many constraints and ignore other constraints on (G_0, G_1). Keep in mind that such constraints are inherent and different from modeling assumptions. The proposal is similar to the formulation for Monte Carlo integration of Kong et al. (2003), except where using all information leads to a perfect but computationally infeasible estimator.

Let $\mathbf{h}^* = (\pi^*, 1 - \pi^*, h_1^*, \ldots, h_m^*)$ be a vector of real-valued functions including π^* and $1 - \pi^*$ on $\mathcal{X}$. We maximize L_2 over (G_0, G_1) subject to

$$\int \pi^*(\mathbf{x})\, \mathrm{d}G_0 = \int \pi^*(\mathbf{x})\, \mathrm{d}G_1$$

and

$$\int h_j^*(\mathbf{x})\, \mathrm{d}G_0 = \int h_j^*(\mathbf{x})\, \mathrm{d}G_1, \qquad j = 1, \ldots, m.$$

The constraint associated with π^* is such that the marginal probabilities of $T = 1$ and $T = 0$ add to 1, whereas other constraints are included for variance reduction. In addition, we restrict our attention to the G_1's supported on $\{(\mathbf{X}_i, Y_i)\colon i = 1, \ldots, n_1\}$ and G_0's supported on $\{(\mathbf{X}_i, Y_i)\colon i = n_1 + 1, \ldots, n\}$, for technical reasons (see Tan 2004). Theorem 1 provides a formula for the constrained maximum likelihood estimator (MLE). If $\mathcal{X}$ is finite and $\mathbf{h}^*$ spans all functions on $\mathcal{X}$, then $(\hat{G}_0, \hat{G}_1)$ reduces to the usual MLE and does not depend on the propensity score; see Section 4.1.

Theorem 1. Assume that $(1, \ldots, 1)$, $(\pi^*(\mathbf{X}_1), \ldots, \pi^*(\mathbf{X}_n))$, $(h_j^*(\mathbf{X}_1), \ldots, h_j^*(\mathbf{X}_n))$, $j = 1, \ldots, m$, are linearly independent, and that the function $\ell_n : \mathbb{R}^{m+2} \to \mathbb{R} \cup \{-\infty\}$ achieves a maximum at $\hat{\boldsymbol{\lambda}}$,

$$\ell_n(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n_1} \log(\boldsymbol{\lambda}^\top \mathbf{h}^*(\mathbf{X}_i)) + \frac{1}{n}\sum_{i=n_1+1}^{n} \log(1 - \boldsymbol{\lambda}^\top \mathbf{h}^*(\mathbf{X}_i)),$$

where log of 0 or a negative number is $-\infty$. Then the constrained MLE is

$$\hat{G}_1(\{\mathbf{X}_i, Y_i\}) = \frac{n^{-1}}{\hat{\boldsymbol{\lambda}}^\top \mathbf{h}^*(\mathbf{X}_i)}, \qquad i = 1, \ldots, n_1,$$

and

$$\hat{G}_0(\{\mathbf{X}_i, Y_i\}) = \frac{n^{-1}}{1 - \hat{\boldsymbol{\lambda}}^\top \mathbf{h}^*(\mathbf{X}_i)}, \qquad i = n_1 + 1, \ldots, n.$$

The distributions $(\hat{G}_0, \hat{G}_1)$ can be visualized as weighted histograms; see Section 5. By construction, $\int h_j^*(\mathbf{x})\, \mathrm{d}\hat{G}_0$ and $\int h_j^*(\mathbf{x})\, \mathrm{d}\hat{G}_1$ are equal for $j = 1, \ldots, m$. The expectation $E(Y_t)$ can be estimated by

$$\hat{\mu}_t = \int y_t\, \mathrm{d}\hat{G}_t.$$

Similarly, expectations of functions of $(\mathbf{X}, Y_0)$ and $(\mathbf{X}, Y_1)$ can be estimated. The estimator $\hat{\mu}_t$ has the IPW form like $\hat{\mu}_{\mathrm{IPW,ext}}$, but using propensity scores estimated from the linear extended model $P(T = 1|\mathbf{X}) = \boldsymbol{\lambda}^\top \mathbf{h}^*(\mathbf{X})$. The model is fitted by maximum likelihood rather than least squares, and the fitted values lie between 0 and 1 asymptotically.

There is a closed-form estimator related to the likelihood estimator. Let

$$\varrho_1^* = \frac{T}{\pi^*(\mathbf{X})} - 1, \qquad \varrho_0^* = \frac{1 - T}{1 - \pi^*(\mathbf{X})} - 1,$$

$$\boldsymbol{\xi}_1^* = \frac{\mathbf{h}^*(\mathbf{X})}{1 - \pi^*(\mathbf{X})}\varrho_1^*, \qquad \boldsymbol{\xi}_0^* (= -\boldsymbol{\xi}_1^*) = \frac{\mathbf{h}^*(\mathbf{X})}{\pi^*(\mathbf{X})}\varrho_0^*,$$

$$\boldsymbol{\zeta}_1^* = \frac{\mathbf{h}^*(\mathbf{X})}{1 - \pi^*(\mathbf{X})}(1 + \varrho_1^*), \qquad \boldsymbol{\zeta}_0^* = \frac{\mathbf{h}^*(\mathbf{X})}{\pi^*(\mathbf{X})}(1 + \varrho_0^*),$$

$$\eta_1^* = Y(1 + \varrho_1^*), \qquad \text{and} \qquad \eta_0^* = Y(1 + \varrho_0^*).$$

Consider the regression estimator

$$\tilde{\mu}_t = \tilde{E}(\eta_t^*) - \tilde{\boldsymbol{\beta}}_t^\top \tilde{E}(\boldsymbol{\xi}_t^*),$$

where $\tilde{\boldsymbol{\beta}}_t = \tilde{\mathbf{B}}_t^{-1}\tilde{\mathbf{C}}_t$, $\tilde{\mathbf{B}}_t = \tilde{E}(\boldsymbol{\xi}_t^*\boldsymbol{\zeta}_t^{*\top})$, $\tilde{\mathbf{C}}_t = \tilde{E}(\boldsymbol{\xi}_t^*\eta_t^*)$, and $\tilde{E}(\cdot)$ denotes sample average. Theorem 2 says that the regression estimator is a first-order approximation to the likelihood estimator under model S0. This result is similar to those established for Monte Carlo integration by Tan (2004). Throughout, "$\simeq$" denotes a difference of order $o_p(n^{-1/2})$.

Theorem 2. Assume that π^* is strictly between 0 and 1 (i.e., $\pi^* \in [\delta, 1-\delta]$ for some $\delta > 0$), that $h_1^*, \ldots, h_m^*$ are bounded (i.e., $|h_j^*| \le \Delta$ for some $\Delta \ge 1$), and that $1, \pi^*, h_1^*, \ldots, h_m^*$ are linearly independent on $\mathcal{X}$. Under model S0 (i.e., $\pi^* = \pi$), we have

$$\hat{\mu}_t \simeq \tilde{\mu}_t \simeq \tilde{E}(\eta_t^*) - \boldsymbol{\beta}_t\tilde{E}(\boldsymbol{\xi}_t^*),$$

where $\boldsymbol{\beta}_t = \mathbf{B}_t^{-1}\mathbf{C}_t$, $\mathbf{B}_t = E(\boldsymbol{\xi}_t^*\boldsymbol{\zeta}_t^{*\top})$, and $\mathbf{C}_t = E(\boldsymbol{\xi}_t^*\eta_t^*)$. Here $\mathbf{B}_1 = \mathbf{B}_0$ is the variance of $\boldsymbol{\xi}_1^* = -\boldsymbol{\xi}_0^*$, and $\mathbf{C}_1$ or $\mathbf{C}_0$ is the covariance of $\boldsymbol{\xi}_1^*$ and η_1^* or $\boldsymbol{\xi}_0^*$ and η_0^*.

From an estimating equation standpoint, $\tilde{\mu}_t$ is an instance of the method of control variates using estimated optimal coefficients for variance reduction (see Hammersley and Handscomb 1964). The method exploits the fact that $E(\eta_t^*) = \mu_t$ and $E(\boldsymbol{\xi}_t^*) = \mathbf{0}$ under model S0 and defines a class of estimators

$$\tilde{E}(\eta_t^*) - \mathbf{b}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*), \tag{2}$$

where $\mathbf{b}_t$ is an arbitrary vector. The optimal choice of $\mathbf{b}_t$ in minimizing the variance of (2) is given by $\boldsymbol{\beta}_t$. A consistent estimator can be substituted for the unknown $\boldsymbol{\beta}_t$, and the resulting estimator of μ_t achieves the same minimum variance asymptotically. A classical estimator of $\boldsymbol{\beta}_t$ is $\tilde{E}(\boldsymbol{\xi}_t^*\boldsymbol{\xi}_t^{*\top})^{-1}\tilde{E}(\boldsymbol{\xi}_t^*\eta_t^*)$, leading to the estimator $\hat{\mu}_{\mathrm{AIPW,est}}$. The particular estimator $\tilde{\boldsymbol{\beta}}_t$ is proposed here, because the corresponding estimator $\tilde{\mu}_t$ can remain valid even when the propensity score model is wrong; see Theorem 3.

Given the functions $\mathbf{h}^*$, the estimators $\hat{\mu}_t$ and $\tilde{\mu}_t$ achieve the lowest asymptotic variance possible by using the control variates $\boldsymbol{\xi}_t^*$. A remaining question is how to choose $\mathbf{h}^*$. Robins and Rotnitzky (1995) and Hahn (1998) showed that $\eta_t^* - \mu_t - E(Y_t|\mathbf{X})\varrho_t^*$ is an efficient influence function (whose variance gives the semiparametric variance bound) under model S0. By this result, it is desirable to choose $\mathbf{h}^*$ such that (condition R)

$$E(Y_1|\mathbf{X} = \mathbf{x}) \text{ is a linear combination of } (1-\pi^*)^{-1}\mathbf{h}^*$$

and

$$E(Y_0|\mathbf{X} = \mathbf{x}) \text{ is a linear combination of } \pi^{*-1}\mathbf{h}^*.$$

Then $\boldsymbol{\beta}_t$ gives the combination coefficient of $E(Y_t|\mathbf{X})$ and $\boldsymbol{\beta}_t^\top\boldsymbol{\xi}_t^*$ gives $E(Y_t|\mathbf{X})\varrho_t^*$ because

$$\begin{aligned}\boldsymbol{\beta}_t &= E^{-1}(\boldsymbol{\xi}_t^*\boldsymbol{\zeta}_t^{*\top})E[\boldsymbol{\xi}_t^*\, Y_t(1+\rho_t^*)] \\ &= E^{-1}(\boldsymbol{\xi}_t^*\boldsymbol{\zeta}_t^{*\top})E\big[\boldsymbol{\xi}_t^*\, E(Y_t|\mathbf{X})(1+\rho_t^*)\big].\end{aligned}$$

It follows that if condition R holds, then $\hat{\mu}_t$ and $\tilde{\mu}_t$ achieve the semiparametric variance bound under model S0. An outcome regression model can be used as a guidance to choose the functions $\mathbf{h}^*$. In fact, condition R is satisfied asymptotically for $\mathbf{h}^* = (\pi^*, 1-\pi^*, \pi^*\Psi(\hat{\boldsymbol{\alpha}}_0^\top\mathbf{g}), (1-\pi^*)\Psi(\hat{\boldsymbol{\alpha}}_1^\top\mathbf{g}))$ if model R holds.

Suppose now that our knowledge about the propensity score can be wrong. Theorem 3 shows the asymptotic behavior of $\tilde{\mu}_t$ whether or not model S0 is correct. By the definition of $\boldsymbol{\beta}_t$, the third term has mean 0 in the expansion. If model S0 is correct, then $\hat{\mu}_t$ and $\tilde{\mu}_t$ are consistent and the third term vanishes because $E(\eta_t^* - \boldsymbol{\beta}_t^\top\boldsymbol{\xi}_t^*) = \mu_t$ and $E(\boldsymbol{\xi}_t^*) = \mathbf{0}$. Otherwise, $\hat{\mu}_t$ and $\tilde{\mu}_t$ become inconsistent in general. However, $\tilde{\mu}_t$ remains consistent if condition R holds. The reason for this is twofold: $\boldsymbol{\beta}_t$ always gives the combination coefficient of $E(Y_t|\mathbf{X})$ under condition R, and the expectation of $\eta_t^* - E(Y_t|\mathbf{X})\varrho_t^*$ remains μ_t even if model S0 is wrong. This robustness can be seen as bias reduction: the asymptotic bias of $\tilde{\mu}_t$ is 0 or close to 0 if condition R holds or approximately so.

Theorem 3. Assume the regularity conditions in Theorem 2. Then

$$\tilde{\mu}_t \simeq \tilde{E}(\eta_t^*) - \boldsymbol{\beta}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*) - E^\top(\boldsymbol{\xi}_t^*)\mathbf{B}_t^{-1}\tilde{E}[\boldsymbol{\xi}_t^*(\eta_t^* - \boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)],$$

where $\boldsymbol{\beta}_t$ and $\mathbf{B}_t$ are defined as in Theorem 2.

It is interesting to compare the new estimators and previously proposed estimators using models S0 and R. For simplicity, let the estimand be μ_1 and take $\mathbf{h}^* = (1-\pi^*, (1-\pi^*)\Psi(\hat{\boldsymbol{\alpha}}_1^\top\mathbf{g}))$. Table 1 presents a comparison in terms of three properties: optimal in using control variates (CV), locally efficient, and doubly robust.

The estimators $\hat{\mu}_1$ and $\hat{\mu}_{\mathrm{IPW,ext}}$ are similar in terms of both construction (through an extended propensity score model) and properties (being locally efficient and optimal in using control variates but not doubly robust). The estimators $\tilde{\mu}_1$ and $\hat{\mu}_{\mathrm{IPW,lim}}$ are similar, deliberately constructed to be doubly robust. However, $\tilde{\mu}_1$ admits a closed-form expression. For $\hat{\mu}_{\mathrm{IPW,lim}}$, the iteration $\tilde{\theta}_{k-1} \mapsto \tilde{\theta}_k$ is so implicitly defined that there is even no closed-form estimating equation for $\tilde{\theta}_\infty$, in contrast to the usual case where an estimator is not in closed form but can be defined by a closed-form estimating equation.

The estimators $\tilde{\mu}_1$, $\hat{\mu}_{\mathrm{AIPW,fix}}$, and $\hat{\mu}_{\mathrm{AIPW,est}}$ belong to the same class (2) with different choices of $\mathbf{b}_1$. In particular, $\tilde{\mu}_1$ and $\hat{\mu}_{\mathrm{AIPW,fix}}$ are equally efficient if models S0 and R are both correct and are robust to misspecification of model S0 if model R holds. However, $\tilde{\mu}_1$ is more efficient if model R is wrong but model S0 is correct, because it achieves the lowest asymptotic variance among the class (2). A lesson is that simply substituting estimates from a working model for unknown quantities can be inefficient if the working model is wrong, even though there is no efficiency loss if the working model is correct.

The choice $\mathbf{h}^* = (\pi^*,\ 1-\pi^*, \pi^*\Psi(\hat{\boldsymbol{\alpha}}_0^\top\mathbf{g}), (1-\pi^*)\Psi(\hat{\boldsymbol{\alpha}}_1^\top\mathbf{g}))$ is more suitable for simultaneous estimation of μ_0 and μ_1 than taking $\mathbf{h}^* = (\pi^*, \pi^*\Psi(\hat{\boldsymbol{\alpha}}_0^\top\mathbf{g}))$ and $(1-\pi^*, (1-\pi^*)\Psi(\hat{\boldsymbol{\alpha}}_1^\top\mathbf{g}))$ separately in the two treatment arms. First, it brings more control variates and leads to better efficiency and robustness asymptotically. Second, it allows the same set of control variates for estimation of μ_0 and μ_1 so that the difference also achieves the lowest asymptotic variance possible by using these control variates.

Table 1. Theoretical Comparison of Estimators

	AIPW fix	*AIPW est*	*OR ext*	*IPW ext*	*IPW lim*	*REG*	*LIK*
Optimal CV	×	✓	×	✓	✓	✓	✓
Locally efficient	✓	✓	✓	✓	✓	✓	✓
Doubly robust	✓	×	✓	×	✓	✓	×

3.2 Parametric Propensity Score

We turn to the case where our knowledge about the propensity score is parametric and expressed as $\pi(\cdot;\boldsymbol{\gamma})$ or model S. Hirano, Imbens, and Ridder (2003) considered nonparametric estimation of the propensity score and showed that the IPW estimator achieves the semiparametric variance bound under suitable smoothing conditions.

White (1982) generalized the theory of maximum likelihood to possibly misspecified models. Let $\boldsymbol{\gamma}^*$ be the value of $\boldsymbol{\gamma}$ minimizing the Kullback–Leibler distance between $\pi(\cdot;\boldsymbol{\gamma})$ and the underlying propensity score π or, equivalently, maximizing $E(\kappa)$ with

$$\kappa(\boldsymbol{\gamma}) = T\log\pi(\mathbf{X};\boldsymbol{\gamma}) + (1-T)\log(1-\pi(\mathbf{X};\boldsymbol{\gamma})).$$

Geometrically, $\pi^* = \pi(\cdot;\boldsymbol{\gamma}^*)$ is the closest element in model S to the truth π. Model S is correct if and only if π^* and π agree. The score function is

$$\mathbf{s} = \frac{\partial\kappa}{\partial\boldsymbol{\gamma}} = \frac{T-\pi(\mathbf{X};\boldsymbol{\gamma})}{\pi(\mathbf{X};\boldsymbol{\gamma})(1-\pi(\mathbf{X};\boldsymbol{\gamma}))}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}.$$

Under standard regularity conditions, the MLE $\hat{\boldsymbol{\gamma}}$ converges to $\boldsymbol{\gamma}^*$ with probability 1 and has the expansion

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \simeq \mathbf{V}^{-1}\tilde{E}(\mathbf{s}^*),$$

where $\mathbf{s}^* = \mathbf{s}(\cdot;\boldsymbol{\gamma}^*)$ satisfies $E(\mathbf{s}^*) = \mathbf{0}$ and $\mathbf{V} = -E(\partial^2\kappa/\partial\boldsymbol{\gamma}^2)|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}$ is nonsingular.

The likelihood is factorized in terms of $\boldsymbol{\gamma}$ and (G_0, G_1). We consider a maximum likelihood procedure in two steps. First, maximize $L_1(\boldsymbol{\gamma})$, that is, fit the propensity score model S. Let $\mathbf{h}^{(1)} = [\pi(\cdot;\boldsymbol{\gamma}), 1-\pi(\cdot;\boldsymbol{\gamma}), h_1(\cdot;\boldsymbol{\gamma}), \ldots, h_m(\cdot;\boldsymbol{\gamma})]$ be a vector of real-valued functions including $\pi(\cdot;\boldsymbol{\gamma})$ and $1-\pi(\cdot;\boldsymbol{\gamma})$ on $\mathcal{X}$ and $\mathbf{h}^{(2)} = \partial\pi(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, and evaluate $\hat{\mathbf{h}}^{(1)} = \mathbf{h}^{(1)}(\cdot;\hat{\boldsymbol{\gamma}})$ and $\hat{\mathbf{h}}^{(2)} = \mathbf{h}^{(2)}(\cdot;\hat{\boldsymbol{\gamma}})$. Second, maximize $L_2(G_0, G_1)$ subject to

$$\int \hat{\pi}(\mathbf{x})\,dG_0 = \int \hat{\pi}(\mathbf{x})\,dG_1,$$

$$\int \hat{h}_j(\mathbf{x})\,dG_0 = \int \hat{h}_j(\mathbf{x})\,dG_1, \qquad j = 1,\ldots,m,$$

and

$$\int \frac{\partial\hat{\pi}}{\partial\gamma_j}(\mathbf{x})\,dG_0 = \int \frac{\partial\hat{\pi}}{\partial\gamma_j}(\mathbf{x})\,dG_1, \qquad j = 1,\ldots,l.$$

The likelihood estimator $\hat{\mu}_t$ is defined as before,

$$\hat{\mu}_t = \int y_t\,d\hat{G}_t.$$

To introduce the regression estimator, let $\mathbf{h} = (\mathbf{h}^{(1)}, \mathbf{h}^{(2)})$, and

$$\varrho_1 = \frac{T}{\pi(\mathbf{X};\boldsymbol{\gamma})} - 1, \qquad \varrho_0 = \frac{1-T}{1-\pi(\mathbf{X};\boldsymbol{\gamma})} - 1,$$

$$\boldsymbol{\xi}_1 = \frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma})}{1-\pi(\mathbf{X};\boldsymbol{\gamma})}\varrho_1, \qquad \boldsymbol{\xi}_0 (= -\boldsymbol{\xi}_1) = \frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma})}{\pi(\mathbf{X};\boldsymbol{\gamma})}\varrho_0,$$

$$\boldsymbol{\zeta}_1 = \frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma})}{1-\pi(\mathbf{X};\boldsymbol{\gamma})}(1+\varrho_1), \qquad \boldsymbol{\zeta}_0 = \frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma})}{\pi(\mathbf{X};\boldsymbol{\gamma})}(1+\varrho_0),$$

$$\eta_1 = Y(1+\varrho_1), \qquad \text{and} \qquad \eta_0 = Y(1+\varrho_0).$$

Define the hat variables $\hat{\varrho}_t$, $\hat{\boldsymbol{\xi}}_t$, $\hat{\boldsymbol{\zeta}}_t$, and $\hat{\eta}_t$ by evaluation at $\hat{\boldsymbol{\gamma}}$ and the limit variables ϱ_t^*, $\boldsymbol{\xi}_t^*$, $\boldsymbol{\zeta}_t^*$, and η_t^* by evaluation at $\boldsymbol{\gamma}^*$. Consider the regression estimator

$$\tilde{\mu}_t = \tilde{E}(\hat{\eta}_t) - \tilde{\boldsymbol{\beta}}_t^\top\tilde{E}(\hat{\boldsymbol{\xi}}_t),$$

where $\tilde{\boldsymbol{\beta}}_t = \tilde{\mathbf{B}}_t^{-1}\tilde{\mathbf{C}}_t$, $\tilde{\mathbf{B}}_t = \tilde{E}(\hat{\boldsymbol{\xi}}_t\hat{\boldsymbol{\zeta}}_t^\top)$, and $\tilde{\mathbf{C}}_t = \tilde{E}(\hat{\boldsymbol{\xi}}_t\hat{\eta}_t)$. As a generalization of Theorem 2, Theorem 4 says that the regression estimator is a first-order approximation to the likelihood estimator under model S.

Theorem 4. In addition to the regularity conditions in Theorem 2, assume that the results of White (1982) hold and that $\partial\pi(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, $\partial h_j(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, $j = 1,\ldots,m$, and $\partial^2\pi(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^2$ are uniformly bounded in a neighborhood of $\boldsymbol{\gamma}^*$. Under model S (i.e., $\pi^* = \pi$), we have

$$\begin{aligned}\hat{\mu}_t - \tilde{\mu}_t &\simeq \tilde{E}(\hat{\eta}_t) - \boldsymbol{\beta}_t^\top\tilde{E}(\hat{\boldsymbol{\xi}}_t)\\ &\simeq \tilde{E}(\eta_t^*) - \boldsymbol{\beta}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*),\end{aligned}$$

where $\boldsymbol{\beta}_t = \mathbf{B}_t^{-1}\mathbf{C}_t$, $\mathbf{B}_t = E(\boldsymbol{\xi}_t^*\boldsymbol{\zeta}_t^{*\top})$, and $\mathbf{C}_t = E(\boldsymbol{\xi}_t^*\eta_t^*)$.

To clarify the two expansions, partition $\hat{\mathbf{h}}$ into $(\hat{\mathbf{h}}^{(1)}, \hat{\mathbf{h}}^{(2)})$ and correspondingly partition $\hat{\boldsymbol{\xi}}_t$ into $(\hat{\boldsymbol{\xi}}_t^{(1)}, \hat{\boldsymbol{\xi}}_t^{(2)})$, $\hat{\boldsymbol{\zeta}}_t$ into $(\hat{\boldsymbol{\zeta}}_t^{(1)}, \hat{\boldsymbol{\zeta}}_t^{(2)})$, and $\tilde{\boldsymbol{\beta}}_t$ into $(\tilde{\boldsymbol{\beta}}_t^{(1)}, \tilde{\boldsymbol{\beta}}_t^{(2)})$. It follows that $\hat{\boldsymbol{\xi}}_1^{(2)} = -\hat{\boldsymbol{\xi}}_0^{(2)} = \hat{\mathbf{s}}$, and $\tilde{E}(\hat{\boldsymbol{\xi}}_t^{(2)}) = \tilde{E}(\pm\hat{\mathbf{s}}) = \mathbf{0}$. The estimator $\tilde{\mu}_t$ becomes $\tilde{E}(\hat{\eta}_t) - \tilde{\boldsymbol{\beta}}_t^{(1)\top}\tilde{E}(\hat{\boldsymbol{\xi}}_t^{(1)})$, and the first expansion in Theorem 4 reduces to

$$\tilde{E}(\hat{\eta}_t) - \boldsymbol{\beta}_t^{(1)\top}\tilde{E}\big(\hat{\boldsymbol{\xi}}_t^{(1)}\big).$$

By Taylor expansions, $\tilde{E}(\hat{\eta}_t)$ and $\tilde{E}(\hat{\boldsymbol{\xi}}_t^{(1)})$ are asymptotically equivalent to $\tilde{E}(\Pi^\perp[\eta_t^*|\mathbf{s}^*])$ and $\tilde{E}(\Pi^\perp[\boldsymbol{\xi}_t^{*(1)}|\mathbf{s}^*])$, where $\Pi^\perp[\cdot|\cdot]$ denotes the residual in the projection of the first variable on the second; see the Appendix for details. Moreover, the coefficient $\boldsymbol{\beta}_t^{(1)}$ of $\boldsymbol{\xi}_t^{*(1)}$ in the full regression of η_t^* on $\boldsymbol{\xi}_t^* = (\boldsymbol{\xi}_t^{*(1)}, \pm\mathbf{s}^*)$ equals the coefficient in the regression of the residual $\Pi^\perp[\eta_t^*|\mathbf{s}^*]$ (from the regression of η_t^* on $\mathbf{s}^*$) on the residual $\Pi^\perp[\boldsymbol{\xi}_t^{*(1)}|\mathbf{s}^*]$ (from the regression of $\boldsymbol{\xi}_t^{*(1)}$ on $\mathbf{s}^*$) by the theory of linear models (Draper and Smith 1981). The second expansion in Theorem 4 follows because the residual from the full regression also equals that from the regression of the residuals.

Theorem 4 implies two optimality properties of $\hat{\mu}_t$ and $\tilde{\mu}_t$. First, $\hat{\mu}_t$ and $\tilde{\mu}_t$ achieve the lowest asymptotic variance among the class

$$\tilde{E}(\hat{\eta}_t) - \mathbf{b}_t^{(1)\top}\tilde{E}\big(\hat{\boldsymbol{\xi}}_t^{(1)}\big), \tag{3}$$

where $\mathbf{b}_t^{(1)}$ is an arbitrary vector, because $\boldsymbol{\beta}_t^{(1)}$ equals the asymptotic covariance of $\hat{\eta}_t$ and $\hat{\boldsymbol{\xi}}_t^{(1)}$ divided by the asymptotic variance of $\hat{\boldsymbol{\xi}}_t^{(1)}$. That is, $\hat{\mu}_t$ and $\tilde{\mu}_t$ make optimal use of the control variates $\hat{\boldsymbol{\xi}}_t^{(1)}$ derived from the functions $\hat{\mathbf{h}}^{(1)}$ for variance reduction. Second, $\hat{\mu}_t$ and $\tilde{\mu}_t$ achieve the semiparametric variance bound under model S if condition R holds [i.e., $E(Y_1|\mathbf{X} = \mathbf{x})$ is a linear combination of $(1-\pi^*)^{-1}\mathbf{h}^*$ and $E(Y_0|\mathbf{X} = \mathbf{x})$ is a linear combination of $\pi^{*-1}\mathbf{h}^*$], because $\boldsymbol{\beta}_t^\top\boldsymbol{\xi}_t^*$ then gives $E(Y_t|\mathbf{X})\varrho_t^*$ and $\eta_t^* - \mu_t - E(Y_t|\mathbf{X})\varrho_t^*$ is an efficient influence function under model S. The effect of variance reduction is overall optimal if condition R holds approximately.

Suppose now that our knowledge about the propensity score can be wrong. The estimators $\hat{\mu}_t$ and $\tilde{\mu}_t$ become inconsistent in

general; however, $\tilde{\mu}_t$ remains consistent if condition R holds. The reason is twofold: $\boldsymbol{\beta}_t^\top \boldsymbol{\xi}_t^*$ always gives $E(Y_t|\mathbf{X})\varrho_t^*$ under condition R, and the expectation of $\eta_t^* - E(Y_t|\mathbf{X})\varrho_t^*$ remains μ_t even if model S is wrong. The asymptotic bias of $\tilde{\mu}_t$ can be substantially reduced as compared with $\hat{\mu}_{\mathrm{IPW}} = \tilde{E}(\hat{\eta}_t)$ if condition R holds approximately. As a generalization of Theorem 3, Theorem 5 shows the asymptotic behavior of $\tilde{\mu}_t$ whether or not model S is correct. The final term arises in the expansion due to the variation of $\hat{\boldsymbol{\gamma}}$ in the hat variables $\hat{\boldsymbol{\xi}}_t$, $\hat{\boldsymbol{\zeta}}_t$, and $\hat{\eta}_t$.

Theorem 5. Assume the regularity conditions in Theorem 4. Then

$$\begin{aligned}\tilde{\mu}_{\mathrm{REG}} \simeq{}& \tilde{E}(\eta_t^*) - \boldsymbol{\beta}_t^\top \tilde{E}(\boldsymbol{\xi}_t^*) - E^\top(\boldsymbol{\xi}_t^*)\mathbf{B}_t^{-1}\tilde{E}[\boldsymbol{\xi}_t^*(\eta_t^* - \boldsymbol{\beta}_t^\top \boldsymbol{\zeta}_t^*)] \\ &+ \{E(\partial\eta_t^*/\partial\boldsymbol{\gamma}) - \boldsymbol{\beta}_t^\top E(\partial\boldsymbol{\xi}_t^*/\partial\boldsymbol{\gamma}) \\ &\quad - E^\top(\boldsymbol{\xi}_t^*)\mathbf{B}_t^{-1}E[\partial(\boldsymbol{\xi}_t^*(\eta_t^* - \boldsymbol{\beta}_t^\top \boldsymbol{\zeta}_t^*))/\partial\boldsymbol{\gamma}]\} \\ &\times \mathbf{V}^{-1}\tilde{E}(\mathbf{s}^*),\end{aligned}$$

where $\boldsymbol{\beta}_t$ and $\mathbf{B}_t$ are defined as in Theorem 4.

The estimators $\hat{\mu}_t$ and $\tilde{\mu}_t$ are systematic generalizations of those in Section 3.1. Here $\hat{\mu}_t$ is defined by maximum likelihood subject to an extra set of constraints associated with $\hat{\mathbf{h}}^{(2)}$, and $\tilde{\mu}_t$ is defined by regressing on an extra set of control variates $\pm\hat{\mathbf{s}}$. These pieces accommodate the variation of $\hat{\boldsymbol{\gamma}}$ and are necessary for asymptotic optimality. However, there can be a trade-off in finite samples. The variation of $\tilde{\boldsymbol{\beta}}_t$ is negligible in large samples under model S but can become substantial in small to medium samples due to the extra regressors $\hat{\mathbf{s}}$. Given this consideration, we can drop the extra constraints in $\hat{\mu}_t$ and the extra control variates in $\tilde{\mu}_t$. Specifically, define the likelihood estimator

$$\hat{\mu}_t^{(\mathrm{m})} = \int y_t \, \mathrm{d}\hat{G}_t^{(\mathrm{m})},$$

where $L_2(G_0, G_1)$ is maximized subject to the constraints associated with $\hat{\mathbf{h}}^{(1)}$ and define the regression estimator

$$\tilde{\mu}_t^{(\mathrm{m})} = \tilde{E}(\hat{\eta}_t) - \tilde{\boldsymbol{\beta}}_t^{(\mathrm{m})\top}\tilde{E}\big(\hat{\boldsymbol{\xi}}_t^{(1)}\big),$$

where $\tilde{\boldsymbol{\beta}}_t^{(\mathrm{m})} = (\tilde{\mathbf{B}}_t^{(\mathrm{m})})^{-1}\tilde{\mathbf{C}}_t^{(\mathrm{m})}$, $\tilde{\mathbf{B}}_t^{(\mathrm{m})} = \tilde{E}(\hat{\boldsymbol{\xi}}_t^{(1)}\hat{\boldsymbol{\zeta}}_t^{(1)\top})$, and $\tilde{\mathbf{C}}_t^{(\mathrm{m})} = \tilde{E}(\hat{\boldsymbol{\xi}}_t^{(1)}\hat{\eta}_t)$. The two estimators are simple generalizations of $\hat{\mu}_t$ and $\tilde{\mu}_t$ in Section 3.1, because $\hat{\pi}$ and $\hat{\mathbf{h}}^{(1)}$ are substituted for π^* and $\mathbf{h}^*$ everywhere. Define condition $\mathrm{R}^{(\mathrm{m})}$ as the case where $E(Y_1|\mathbf{X} = \mathbf{x})$ is a linear combination of $(1 - \pi^*)^{-1}\mathbf{h}^{*(1)}$ and $E(Y_0|\mathbf{X} = \mathbf{x})$ is a linear combination of $\pi^{*-1}\mathbf{h}^{*(1)}$. Theorem 6 summarizes the asymptotic behaviors of $\hat{\mu}_t^{(m)}$ and $\tilde{\mu}_t^{(m)}$.

Theorem 6. (a) In addition to the regularity conditions in Theorem 2, assume that White's (1982) results hold, and that $\partial\pi(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$ and $\partial h_j(\cdot;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, $j = 1, \ldots, m$, are uniformly bounded in a neighborhood of $\boldsymbol{\gamma}^*$. Under model S (i.e., $\pi^* = \pi$), we have

$$\begin{aligned}\hat{\mu}_t^{(\mathrm{m})} &\simeq \tilde{\mu}_t^{(\mathrm{m})} \simeq \tilde{E}(\hat{\eta}_t) - \boldsymbol{\beta}_t^{(\mathrm{m})\top}\tilde{E}\big(\hat{\boldsymbol{\xi}}_t^{(1)}\big) \\ &\simeq \tilde{E}(\Pi^\perp[\eta_t^*|\mathbf{s}^*]) - \boldsymbol{\beta}_t^{(\mathrm{m})\top}\tilde{E}(\Pi^\perp[\boldsymbol{\xi}_t^*|\mathbf{s}^*]),\end{aligned}$$

where $\boldsymbol{\beta}_t^{(\mathrm{m})} = \mathbf{B}_t^{(\mathrm{m})-1}\mathbf{C}_t^{(\mathrm{m})}$, $\mathbf{B}_t^{(\mathrm{m})} = E(\boldsymbol{\xi}_t^{*(1)}\boldsymbol{\zeta}_t^{*(1)\top})$, and $\mathbf{C}_t^{(\mathrm{m})} = E(\boldsymbol{\xi}_t^{*(1)}\eta_t^*)$.

(b) Assume the regularity conditions in part (a). Then $\tilde{\mu}_t^{(\mathrm{m})}$ has the expansion as in Theorem 5 with $\boldsymbol{\xi}_t^{*(1)}$, $\boldsymbol{\zeta}_t^{*(1)}$, $\boldsymbol{\beta}_t^{(\mathrm{m})}$, and $\mathbf{B}_t^{(\mathrm{m})}$ in place of $\boldsymbol{\xi}_t^*$, $\boldsymbol{\zeta}_t^*$, $\boldsymbol{\beta}_t$, and $\mathbf{B}_t$.

The estimators $\hat{\mu}_t$ and $\tilde{\mu}_t$ are fully efficient among the class (3) in large samples, whereas $\hat{\mu}_t^{(\mathrm{m})}$ and $\tilde{\mu}_t^{(\mathrm{m})}$ are expected to perform well in small to medium samples. For $\hat{\mathbf{h}}^{(1)} = (\hat{\pi}, 1 - \hat{\pi}, \hat{\pi}\Psi(\hat{\boldsymbol{\alpha}}_0^\top\mathbf{g}), (1 - \hat{\pi})\Psi(\hat{\boldsymbol{\alpha}}_1^\top\mathbf{g}))$, conditions R and $\mathrm{R}^{(\mathrm{m})}$ are satisfied asymptotically if model R holds. The four estimators are locally efficient (i.e., achieve the semiparametric variance bound under model S if models S and R are both correct). The regression estimators are also doubly robust (i.e., remain consistent and asymptotically normal if either model S or R is correct). From here on, we write $\hat{G}_t$, $\hat{\mu}_t$, and $\tilde{\mu}_t$ for the modified estimators as well as the original estimators if no distinction is made between the two kinds of estimators.

Given the theoretical results, we discuss two issues for data analysis. We illustrate our method in detail in Section 5. First, propensity score models and outcome regression models play different roles, even though double robustness refers equally to the two models. The estimators are derived under the assumption that model S is correct, and then examined in the situation where model R is also correct or model S is misspecified but model R is correct. Although both models must be considered carefully, our strategy for data analysis is to build and check propensity score models as a starting point, and incorporate outcome regression models for variance and bias reduction.

Second, propensity score models can be checked with the following idea. Pick up a collection of test functions $\hat{h}_j$'s on $\mathcal{X}$ and calculate

$$\tilde{E}\left[\hat{h}_j(\mathbf{X})\left(\frac{T}{\hat{\pi}(\mathbf{X})} - \frac{1 - T}{1 - \hat{\pi}(\mathbf{X})}\right)\right]. \tag{4}$$

The statistic gives the average difference in $\hat{h}_j(\mathbf{X})$ between treated and untreated groups after propensity score weighting. A test function in $\hat{\mathbf{h}}^{(1)}$ corresponds to a component of $\tilde{E}(\hat{\boldsymbol{\xi}}_t^{(1)})$. If model S is correct, then the sample averages relative to standard errors (or z-ratios) should be statistically nonsignificant from 0 (or standard normal). Moreover, the regression estimator $\tilde{\mu}_t$ and the IPW estimator $\tilde{E}(\hat{\eta}_t)$ are expected to yield similar values relative to standard errors, indicating a zero bias reduction. Examination of z-ratios against the standard normal can reveal possible misspecification of model S.

The statistic (4) can be written as $\tilde{E}[\phi(\mathbf{X})(T - \hat{\pi}(\mathbf{X}))]$, where $\phi(\mathbf{X}) = \hat{h}_j(\mathbf{X})/[\hat{\pi}(\mathbf{X})(1 - \hat{\pi}(\mathbf{X}))]$. It gives the covariance between a function of the covariates and the residual, thereby representing one piece of information on the patterns of the residual against the covariates. We plan to investigate this approach to model checking for general regression models in future work. Robins and Rotnitzky (2001) discussed goodness-of-fit tests by comparing $\hat{\mu}_{\mathrm{OR}}$ based on model R, $\hat{\mu}_{\mathrm{IPW}}$ based on model S, and a doubly robust estimator, say $\hat{\mu}_{\mathrm{AIPW,fix}}$, based on both models. The difference between $\hat{\mu}_{\mathrm{IPW}}$ and $\hat{\mu}_{\mathrm{IPW,fix}}$ is $\tilde{E}[\phi(\mathbf{X})(T - \hat{\pi}(\mathbf{X}))]$ with $\phi(\mathbf{X}) = \hat{\pi}^{-1}(\mathbf{X})\hat{E}(Y|T = 1, \mathbf{X})$. The difference between $\hat{\mu}_{\mathrm{OR}}$ and $\hat{\mu}_{\mathrm{IPW,fix}}$ is $\tilde{E}[T\phi(\mathbf{X})(Y - \hat{E}(Y|T = 1, \mathbf{X}))]$ with $\phi(\mathbf{X}) = \hat{\pi}^{-1}(\mathbf{X})$, a generalization of the statistic (4) to outcome regression models. Therefore, these goodness-of-fit tests are well connected.

3.3 Simulation Study

Assume that the marginal probability of $T = 1$ is $1/2$, and that X has a truncated normal distribution in each treatment group: $X|T = 1 \sim \mathrm{N}(d, \sigma_1^2)$ and $X|T = 0 \sim \mathrm{N}(-d, \sigma_0^2)$ truncated on the interval $(-a, a)$. Then the marginal distribution of X is a mixture of two truncated normal distributions, and the propensity score is

$$\begin{aligned}
&\operatorname{logit} P(T = 1|X) \\
&\quad = -\log\left[\Phi\left(\frac{a-d}{\sigma_1}\right) - \Phi\left(\frac{-a-d}{\sigma_1}\right)\right] \\
&\qquad - \log\left[\Phi\left(\frac{a+d}{\sigma_0}\right) - \Phi\left(\frac{-a+d}{\sigma_0}\right)\right] \\
&\qquad - \log\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{d^2}{2}\left(-\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) \\
&\qquad + d\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)X + \frac{1}{2}\left(-\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)X^2.
\end{aligned}$$

Assume that, conditional on X, each response Y_t has a normal distribution with mean $E(Y_t|X)$ and variance τ^2 $(t = 0, 1)$. Note that if $a = \infty$ and $\sigma_1^2 \le \sigma_0^2/2$, the variance of $T/\pi(X)$ is infinite. For simplicity, $a = 5/2$ is fixed throughout.

We consider factorial combinations of the following conditions: (a) $d = 1/4$ or $1/2$, (b) $(\sigma_0^2, \sigma_1^2) = (1, 1)$, $(4/3, 2/3)$, or $(2/3, 4/3)$, (c) $E(Y_0|X) = 1 + X$, e^X, $1 + X$, or $1 + X$, and $E(Y_1|X) = 3 + X$, $2 + \mathrm{e}^X$, $3 + 2X$, or $2 + \mathrm{e}^X$, and (d) $\tau^2 = 1/2$. Values of d between $1/8$ and $1/2$ and of σ_1^2/σ_0^2 between $1/2$ and 2 are typical of those that might occur in practice (Rubin 1973). Two possible propensity score models are

(S1) $\operatorname{logit} \pi(X; \gamma) = \gamma_1 + \gamma_2 X$

and

(S2) $\operatorname{logit} \pi(X; \gamma) = \gamma_1 + \gamma_2 X + \gamma_3 X^2$.

Model S2 is correct under all of the specifications of (σ_0^2, σ_1^2), whereas model S1 is correct under only the first specification: $\sigma_0^2 = \sigma_1^2$. The first two specifications in (c) give parallel response curves, whereas the other two give nonparallel response curves. A response model is

(R) $E(Y_0|X) = \alpha_{00} + \alpha_{01} X$ and $E(Y_1|X) = \alpha_{10} + \alpha_{11} X$.

This model is correct under the first and third specifications of $E(Y_t|X)$, but not under the other two. We compare 10 estimators, including SUB, Rosenbaum and Rubin's (1984) subclassification estimator with 5 subclasses; IPW (lim), the estimator $\hat{\mu}_{\mathrm{IPW,lim}}$ with the number of iterations fixed at 50; REG, the regression estimator $\tilde{\mu}_t^{(\mathrm{m})}$; and LIK, the likelihood estimator $\hat{\mu}_t^{(\mathrm{m})}$ with $\hat{\mathbf{h}}^{(1)} = (\hat{\pi}, 1 - \hat{\pi}, \hat{\pi}X, (1 - \hat{\pi})X)$.

Similar comparisons are obtained for different values of d and (σ_0^2, σ_1^2). Table 2 summarizes estimates of $\mu_1 - \mu_0$ for $d = 1/2$ and $(\sigma_0^2, \sigma_1^2) = (4/3, 2/3)$. The OR estimator performs well if the response model used is correct but has serious bias otherwise. The IPW estimator has negligible bias if the propensity score model used is correct but has serious bias otherwise. The SUB estimator has considerable bias even if the propensity score model used is correct, indicating that 5 subclasses are not sufficient here.

Table 2. Numerical Comparison of Estimators

	OR	*SUB*	*IPW*	*AIPW fix*	*AIPW est*	*OR ext*	*IPW ext*	*IPW lim*	*REG*	*LIK*
Quadratic propensity score model										
L	.00045	.086	−.0036	.00023	−.00058	−.0029	−.011	−.0080	−.000030	.00076
I	.0500	.0612	.0776	.0607	.0578	.152	.0642	.0712	.0627	.0613
NP			.0752						.0541	.0530
E	−.38	.054	−.0098	−.024	−.064	.89	−.035	−.025†	−.021	−.020
X	.113	.095	.129	.217	.111	1.50	.0969	.101†	.0755	.0779
PP			.124						.0650	.0622
L	−.00031	.15	.0098	−.00053	.0012	−.0037	.023	.024	−.00079	.0084
I	.0662	.0848	.124	.0742	.0703	.158	.0820	.0931	.0761	.0758
NN			.113						.0708	.0706
E	−.39	.047	−.013	−.027	−.067	.88	−.039	−.033†	−.022	−.012
X	.0813	.0793	.144	.230	.127	1.50	.108	.114†	.0818	.0773
PN			.135						.0724	.0710
Linear propensity score model										
L	.00045	.080	−.23	.00031	.0087	−.00022	−.037	−.15	.000030	−.0019
I	.0500	.0602	.105	.0521	.0513	.0620	.0714	.119	.0611	.0633
NP			.108						.0559	.0527
E	−.38	.022	−.74	−.58	−.37	.29	−.051	−.44	.042	−.0056
X	.113	.106	.306	.207	.114	.230	.527	.413	.0664	.0767
PP			.305						.0615	.0582
L	−.00031	.14	−.11	−.00045	.014	−.00098	−.058	−.078	−.00073	.0029
I	.0662	.0827	.102	.0678	.0664	.0754	.303	.207	.0748	.0784
NN			.106						.0721	.0760
E	−.39	.034	−.35	−.25	−.23	.29	−.035	−.23	−.0011	−.0084
X	.0813	.0786	.126	.0900	.0865	.236	.531	.271	.0753	.0778
PN			.130						.0725	.0709

NOTE: LINP, EXPP, LINN, and EXPN correspond to the four specifications of $E(Y_t|X)$ in the text. The results are based on 5,000 Monte Carlo samples each of size 500. Each cell gives the bias (upper) and the standard deviation (middle) of the point estimates, and $\sqrt{\text{mean of the variance estimates}}$ (lower). Each cell of IPW (lim) labeled with † excludes two Monte Carlo samples in which the iteration diverges grossly.

The REG and AIPW (fix) estimators have small biases if either the propensity score or the response model is correct. If the propensity model is correct, then both estimators have similar variances when the response model is correct, but the REG estimator has variance reduced by a factor of 1–8 when the response model is wrong. The AIPW (est) and IPW (ext) estimators have small variances if the propensity score model is correct but have considerable biases otherwise even if the response model is correct. The OR (ext) and IPW (lim) estimators have overall poor performances.

The LIK estimator has similar mean squared error as the REG estimator if the propensity score model is correct. Otherwise, the LIK estimator appears to have small mean squared error whether or not the response model is correct, even though it may not be consistent in theory. The REG and LIK estimators have overall the smallest mean squared errors. The square root of the mean of variance estimates agrees reasonably well with the corresponding Monte Carlo standard deviation for the two estimators.

4. SENSITIVITY ANALYSIS

Assume that the assignment mechanism can be confounded given measured covariates $T \not\perp (Y_0, Y_1)|\mathbf{X}$. The likelihood of $\{(\mathbf{X}_i, Y_i, T_i)\}$ is

$$L_1 \times L_2 = \prod_{i=1}^{n} \left[(1-\pi(\mathbf{X}_i))^{1-T_i}\pi(\mathbf{X}_i)^{T_i}\right] \times \prod_{i=1}^{n} \left[H_0(\{\mathbf{X}_i, Y_i\})^{1-T_i} H_1(\{\mathbf{X}_i, Y_i\})^{T_i}\right],$$

where H_0 is the composite distribution $P(Y_0|T=0,\mathbf{X})P(\mathbf{X})$ and H_1 is the composite distribution $P(Y_1|T=1,\mathbf{X})P(\mathbf{X})$ with the same marginal distribution on $\mathcal{X}$. The likelihood shares a similar structure as that in Section 3 with (H_0, H_1) in place of (G_0, G_1). The propensity score $\pi(\mathbf{X}) = P(T=1|\mathbf{X})$ states how treatment status T depends on covariates $\mathbf{X}$ regardless of the no-confounding assumption. A propensity score model can be fit to the data $\{(\mathbf{X}_i, T_i)\}$ as usual. On the other hand, (H_0, H_1) are different from (G_0, G_1), the joint distributions of $(\mathbf{X}, Y_0)$ and $(\mathbf{X}, Y_1)$, except under no confounding. An outcome regression model about $E(Y_t|\mathbf{X})$ is not identifiable from the data $\{(\mathbf{X}_i, Y_i, T_i)\}$; nevertheless, a model about $E(Y_t|T=t,\mathbf{X})$ can be identified from the data. Consider the outcome regression model (model Z)

$$E(Y_t|T=t,\mathbf{X}) = \Psi(\boldsymbol{\alpha}_t^\top \mathbf{g}(\mathbf{X})),$$

where $E(Y_t|T=t,\mathbf{X}) = E(Y|T=t,\mathbf{X})$ is the regression function of observed outcome Y on covariates $\mathbf{X}$ for subjects with treatment t. Models Z and R agree with each other under no confounding, but not in general.

The estimation method in Section 3 can be extended such that (H_0, H_1) plays the role of (G_0, G_1). The estimators $\hat{G}_t$, $\hat{\mu}_t$, and $\tilde{\mu}_t$ are defined exactly as before. Theorem 7 summarizes asymptotic results (see Robins 1999, sec. 2.9, for related results). If model S is correct, then $(\hat{G}_0, \hat{G}_1)$ is a consistent estimator of (H_0, H_1), and $\hat{\mu}_t$ is a consistent estimator of $E[E(Y_t|T=t,\mathbf{X})]$, which is the expectation of Y_t given $T=t$ and $\mathbf{X}$ and averaged over the distribution of $\mathbf{X}$. The regression estimator $\tilde{\mu}_t$ is a first-order approximation to the likelihood estimator $\hat{\mu}_t$ if model S is correct, and remains consistent for $E[E(Y_t|T=t,\mathbf{X})]$ if model Z is correct and $\hat{\mathbf{h}}^{(1)}$ includes $(\hat{\pi}, 1-\hat{\pi}, \hat{\pi}\,\Psi(\boldsymbol{\alpha}_0^\top \mathbf{g}), (1-\hat{\pi})\Psi(\boldsymbol{\alpha}_1^\top \mathbf{g}))$. In retrospect, these results clarify what is estimated $[(H_0, H_1)]$ and what is assumed [the equality of (H_0, H_1) to (G_0, G_1)] in no-confounding inference.

Theorem 7. (a) Assume the regularity conditions in Theorem 4 or 6(a). If model S is correct, then $\hat{\mu}_t$ and $\tilde{\mu}_t$ or $\hat{\mu}_t^{(\mathrm{m})}$ and $\tilde{\mu}_t^{(\mathrm{m})}$ are consistent for $E[E(Y_t|T=t,\mathbf{X})]$ and have the same expansions as in Theorem 4 or 6(a).

(b) Assume the regularity conditions in Theorem 5 or 6(b). Then $\tilde{\mu}_t$ or $\tilde{\mu}_t^{(\mathrm{m})}$ has the expansion as in Theorem 5 or 6(b).

The distributions $P(Y_0|T=0,\mathbf{X})$ and $P(Y_1|T=1,\mathbf{X})$ are identifiable from observed data, whereas the distributions $P(Y_0|\mathbf{X})$ and $P(Y_1|\mathbf{X})$ are not. However, there is a mixture structure for these distributions,

$$P(Y_0|\mathbf{X}) = (1-\pi(\mathbf{X}))P(Y_0|T=0,\mathbf{X}) + \pi(\mathbf{X})P(Y_0|T=1,\mathbf{X}) \tag{5}$$

and

$$P(Y_1|\mathbf{X}) = (1-\pi(\mathbf{X}))P(Y_1|T=0,\mathbf{X}) + \pi(\mathbf{X})P(Y_1|T=1,\mathbf{X}). \tag{6}$$

The components $P(Y_0|T=1,\mathbf{X})$ and $P(Y_1|T=0,\mathbf{X})$ are not identifiable because Y_0 (or Y_1) cannot be observed on subjects with $T=1$ (or $T=0$). The gaps between $P(Y_t|T=1-t,\mathbf{X})$ and $P(Y_t|T=t,\mathbf{X})$ indicate unmeasured confounding (i.e., systematic differences between treated and untreated groups in their outcomes that would be observed even if both groups received the same treatment after controlling for measured covariates). In Section 4.2 we propose a sensitivity analysis method by postulating a relationship between the unidentifiable distribution $P(Y_t|T=1-t,\mathbf{X})$ and the identifiable distribution $P(Y_t|T=t,\mathbf{X})$. Next, the distributional equations (5) and (6) imply a corresponding structure for the expectations,

$$E(Y_0|\mathbf{X}) = (1-\pi(\mathbf{X}))E(Y_0|T=0,\mathbf{X}) + \pi(\mathbf{X})E(Y_0|T=1,\mathbf{X}) \tag{7}$$

and

$$E(Y_1|\mathbf{X}) = (1-\pi(\mathbf{X}))E(Y_1|T=0,\mathbf{X}) + \pi(\mathbf{X})E(Y_1|T=1,\mathbf{X}). \tag{8}$$

It follows that the biases $E(Y_0|T=0,\mathbf{X}) - E(Y_0|\mathbf{X})$ and $E(Y_1|T=1,\mathbf{X}) - E(Y_1|\mathbf{X})$ are given by $\pi(\mathbf{X})[E(Y_0|T=0,\mathbf{X}) - E(Y_0|T=1,\mathbf{X})]$ and $(1-\pi(\mathbf{X}))[E(Y_1|T=1,\mathbf{X}) - E(Y_1|T=0,\mathbf{X})]$. This result can be used for sensitivity analysis by postulating a relationship between the unidentifiable expectation $E(Y_t|T=1-t,\mathbf{X})$ and the identifiable expectation $E(Y_t|T=t,\mathbf{X})$ (see Brumback et al. 2004; Robins 1999). We discuss this approach briefly in Section 4.3.

4.1 Illustration

We illustrate our ideas using a simple hypothetical example. Suppose that an iid sample is selected from a male population

Table 3. Binary Covariate

	$T=1$	$T=0$
$X=1$	(52, 28)	(11, 9)
$X=0$	(30, 10)	(37, 23)

NOTE: Each cell gives the numbers of $Y=1$ (left) and $Y=0$ (right).

of age 50–70, and their treatment status T (drinking alcohol or not), binary outcome Y (having headaches or not), and binary covariate X (high or low income) are recorded; see Table 3. For simplicity, imagine that the numbers are in thousands, so that we can ignore sampling errors. The task is to study the effect of drinking alcohol on experiencing headaches in this population.

The raw probability of experiencing headaches is $P(Y_1 = 1|T = 1) = 68.3\%$ among drinkers and $P(Y_0 = 0|T = 0) = 60.0\%$ among nondrinkers. Comparing the two probabilities would give the average causal effect if people chose to drink at random. However, two out of three drinkers (80/120) and one out of four nondrinkers (20/80) earn a high income. It is necessary to adjust for the income level to estimate the effect of drinking.

Because the covariate X is binary, two methods can be used with equivalent results. The first method estimates the regression means given X and T, $E(Y_1|X = 1, T = 1) = 65.0\%$, $E(Y_1|X = 0, T = 1) = 75.0\%$, $E(Y_0|X = 1, T = 0) = 55.0\%$, and $E(Y_1|X = 0, T = 0) = 61.7\%$, and then calculates the sample averages over the distribution of X, $E[E(Y_1|X, T = 1)] = (65.0\% + 75.0\%)/2 = 70.0\%$ and $E[E(Y_0|X, T = 0)] = (55.0\% + 61.7\%)/2 = 58.3\%$. The second method, which is generalized in this article, weights each drinker by $w_1(1)$ if $X = 1$ or $w_1(0)$ if $X = 0$ such that the weighted probabilities of $X = 1$ and $X = 0$ among drinkers match the probabilities of $X = 1$ and $X = 0$ in the whole sample [$80w_1(1) = 1/2$ and $40w_1(0) = 1/2$], and then estimates $E[E(Y_1|X, T = 1)]$ by the weighted probability of $Y_1 = 1$ [$52w_1(1) + 30w_1(0) = 70.0\%$]. Similarly, it weights each nondrinker by $w_0(1)$ if $X = 1$ or $w_0(0)$ if $X = 0$ such that $20w_0(1) = 1/2$ and $60w_0(0) = 1/2$, and then estimates $E[E(Y_0|X, T = 0)]$ by the weighted probability of $Y_0 = 1$ [$11w_0(1) + 37w_0(0) = 58.3\%$]. It is easy to see that $w_1(X) = 1/[200\pi(X)]$ and $w_0(X) = 1/[200(1 - \pi(X))]$, where the propensity score $\pi(X) = P(T = 1|X)$ is .8 for $X = 1$ and .4 for $X = 0$. If both high-income and low-income persons chose to drink at random (i.e., there were no confounding given the income level), then $E[E(Y_1|X, T = 1)]$ would become $P(Y_1 = 1)$ and $E[E(Y_0|X, T = 0)]$ would become $P(Y_0 = 1)$, so that comparing 70.0% and 58.3% would give the average causal effect.

A sensitivity analysis asks how the estimates might change without the assumption of no confounding. In that case, people are not equally likely to drink alcohol at the same level of income, and those experiencing headaches might be more inclined to drink due to some unmeasured characteristics. The overall probability of drinking alcohol is $\pi(1) = 80\%$ among high-income people and $\pi(0) = 40\%$ among low-income people. Our method postulates that the actual odds of drinking alcohol could differ from .8/.2 among high-income people and .4/.6 among low-income people by at most a factor of some number $\Lambda \geq 1$. Alternatively, unmeasured confounding says that being a drinker might be related to some unmeasured characteristic that increases the probability of experiencing headaches at each level of income. For example, the probability of experiencing headaches for high-income drinkers could be larger than that of experiencing headaches if high-income nondrinkers did drink. It is equivalent to postulate that the counterfactual probabilities could differ from the corresponding observed probabilities by at most a factor of Λ.

For a fixed value of Λ, our method proceeds as follows. First, it weights each drinker i by $\lambda_{1i}w_1(X_i)$, $1 \leq i \leq 120$, and each nondrinker i by $\lambda_{0i}w_0(X_i)$, $121 \leq i \leq 200$, such that the weighted probabilities of X match those in the sample,

$$\sum_{i=1}^{120} \lambda_{1i}w_1(1)X_i = \frac{1}{2}, \qquad \sum_{i=1}^{120} \lambda_{1i}w_1(0)(1 - X_i) = \frac{1}{2},$$

$$\sum_{i=121}^{200} \lambda_{0i}w_0(1)X_i = \frac{1}{2}, \quad \text{and} \qquad \sum_{i=121}^{200} \lambda_{0i}w_0(0)(1 - X_i) = \frac{1}{2},$$

where $\Lambda^{-1} \leq \lambda_{1i}, \lambda_{0i} \leq \Lambda$ are real numbers. Second, it obtains lower and upper bounds for $E[E(Y_1|X, T = 0)]$ and $E[E(Y_0|X, T = 1)]$, by minimizing and maximizing the weighted probabilities of $Y_1 = 1$ and $Y_0 = 1$ over the unknowns λ_{1i}'s and λ_{0i}'s,

$$\text{min or max} \quad \sum_{i=1}^{120} \lambda_{1i}w_1(X_i)Y_i$$

and

$$\text{min or max} \quad \sum_{i=121}^{200} \lambda_{0i}w_0(X_i)Y_i.$$

Similarly, bounds can be obtained for $E(Y_1)$ and $E(Y_0)$ by solving

$$\text{min or max} \quad \sum_{i=1}^{120} \big[\pi(X_i) + (1 - \pi(X_i))\lambda_{1i}\big]w_1(X_i)\,Y_i$$

and

$$\text{min or max} \quad \sum_{i=121}^{200} \big[\pi(X_i)\lambda_{0i} + (1 - \pi(X_i))\big]w_0(X_i)Y_i.$$

Let $\Lambda = 1.5$. The lower bound for $E[E(Y_1|X, T = 0)]$ is 55.0% and that for $E(Y_1)$ is 64.5%, both achieved at $\lambda_{1i} = \lambda_1(X_i, Y_i)$ where $\lambda_1(X, Y) = 1/1.4$, 1.5, $1/1.2$, and 1.5 for $(X, Y) = (1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$. The upper bound for $E[E(Y_0|X, T = 1)]$ is 72.2% and that for $E(Y_0)$ is 66.9%, both achieved at $\lambda_{0i} = \lambda_0(X_i, Y_i)$ where $\lambda_0(X, Y) = 1.3$, $1/1.5$, 1.2, and $1/1.5$ for $(X, Y) = (1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$. The observed income-adjusted difference in experiencing headaches, 70.0% versus 58.3%, between drinkers and nondrinkers could be explained away by some unmeasured characteristics such that people experiencing headaches are more inclined to drink by a factor of $1.5^2 = 2.25$ in odds than those experiencing no headaches at each level of income.

4.2 Distributional Specification

A sensitivity analysis investigates how inferences might change given unmeasured confounding of various magnitudes.

We continue the line of reasoning from equations (5) and (6) and develop a sensitivity analysis method based on $(\hat{G}_0, \hat{G}_1)$.

Assume that $P(Y_t|T = 1 - t, \mathbf{X})$ (i.e., the distribution of Y_t that would be observed for subjects with treatment $1 - t$ and covariates $\mathbf{X}$) is absolutely continuous with respect to $P(Y_t|T = t, \mathbf{X})$ (i.e. the distribution of Y_t that is observed for subjects with treatment t and covariates $\mathbf{X}$), $t = 0, 1$. Define the Radon–Nikodym derivative or the ratio of densities with respect to a baseline measure,

$$\lambda_0(\mathbf{X}, Y_0) = \frac{\mathrm{d}P(Y_0|T = 1, \mathbf{X})}{\mathrm{d}P(Y_0|T = 0, \mathbf{X})} \qquad \text{and}$$

$$\lambda_1(\mathbf{X}, Y_1) = \frac{\mathrm{d}P(Y_1|T = 0, \mathbf{X})}{\mathrm{d}P(Y_1|T = 1, \mathbf{X})}.$$

The case where $\lambda_0 = \lambda_1 = 1$ corresponds to no confounding, whereas deviations of λ_0 and λ_1 from 1 indicate unmeasured confounding. For example, $\lambda_1(\mathbf{X}, Y_1)$ is <1 for larger values of Y_1 and >1 for smaller values of Y_1 when treated subjects tend to have larger outcomes than untreated subjects with the same $\mathbf{X}$ would have under the treatment. By Bayes's rule, λ_0 and λ_1 can be expressed as odds ratios,

$$\lambda_0(\mathbf{X}, Y_0) = \frac{(1 - \pi(\mathbf{X}))P(T = 1|Y_0, \mathbf{X})}{\pi(\mathbf{X})P(T = 0|Y_0, \mathbf{X})} \qquad \text{and}$$

$$\lambda_1(\mathbf{X}, Y_1) = \frac{\pi(\mathbf{X})P(T = 0|Y_1, \mathbf{X})}{(1 - \pi(\mathbf{X}))P(T = 1|Y_1, \mathbf{X})}.$$

For the foregoing example, it is equivalent to say that subjects with the same $\mathbf{X}$ but having larger outcomes under the treatment are more likely to be treated.

The characterization of unmeasured confounding through λ_t is related to selection odds models of Robins et al. (1999). Let $q_t(\mathbf{x}, y_t)$ be a known function to be varied in sensitivity analysis ($t = 0, 1$). A nonparametric selection odds model is

$$\operatorname{logit} P(T = t|\mathbf{X}, Y_t) = \log b_t(\mathbf{X}) - \log q_t(\mathbf{X}, Y_t),$$

where $b_t(\mathbf{x})$ is an unknown function on $\mathcal{X}$. Equivalently, the model says that $\lambda_t(\mathbf{x}, y_t)$ is proportional to $q_t(\mathbf{x}, y_t)$ at each $\mathbf{x} \in \mathcal{X}$,

$$\lambda_t(\mathbf{X}, Y_t) = \frac{q_t(\mathbf{X}, Y_t)}{a_t(\mathbf{X})},$$

where $a_t(\mathbf{x}) = \int q_t(\mathbf{x}, y_t)\,\mathrm{d}P(y_t|T = t, \mathbf{x})$. If $\mathbf{X}$ is high-dimensional with continuous components, then parametric assumptions are needed on a_t or b_t to avoid the curse of dimensionality in estimation. In this case, Rotnitzky, Robins, and Scharfstein's (1998) augmented IPW estimators can be inconsistent in either direction. Moreover, Robins et al. (1999, sec. 7.2) showed that the estimators can be incompatible with any joint distribution of $(\mathbf{X}, Y_0, Y_1, T)$, because implications from the two treatment arms conflict. Consequently, we avoid parametric specifications and work with bounds on λ_t.

Suppose that the odds of receiving the treatment for subjects with covariates $\mathbf{X}$ could be different from $\pi(\mathbf{X})/(1 - \pi(\mathbf{X}))$ by at most a factor of Λ,

$$\Lambda^{-1} \le \lambda_t(\mathbf{X}, Y_t) \le \Lambda, \tag{9}$$

where $\Lambda \ge 1$ indicates the degree of departure from no confounding ($t = 0, 1$). The model is similar to the model of Rosenbaum (2002a, sec. 4.2) but has a slightly different parameterization. Rosenbaum's model says that the odds ratio of receiving the treatment for two subjects with $\mathbf{X}_i = \mathbf{X}_j$ is at most Γ,

$$\Gamma^{-1} \le \frac{\pi(\mathbf{X}_i, U_i)(1 - \pi(\mathbf{X}_j, U_j))}{(1 - \pi(\mathbf{X}_i, U_i))\pi(\mathbf{X}_j, U_j)} \le \Gamma,$$

where $\Gamma \ge 1$ is a sensitivity parameter, U is an unmeasured covariate such that $T \perp (Y_0, Y_1)|(\mathbf{X}, U)$, and $\pi(\mathbf{X}, U) = P(T = 1|\mathbf{X}, U)$. If $\Lambda = \sqrt{2}$ or $\Gamma = 2$, then two subjects who appear similar, with the same $\mathbf{X}$, could differ in their odds of receiving the treatment by as much as a factor of 2.

The distributions (H_0, H_1) can be estimated consistently by $(\hat{G}_0, \hat{G}_1)$; see Theorem 7. Let H_1^c be the composite distribution $P(Y_0|T = 1, \mathbf{X})P(\mathbf{X})$ and let H_0^c be the composite distribution $P(Y_1|T = 0, \mathbf{X})P(\mathbf{X})$. By the definition of λ_t, the distributions (H_0^c, H_1^c) are related to (H_0, H_1) by

$$\mathrm{d}H_1^c = \lambda_0(\mathbf{X}, Y_0)\,\mathrm{d}H_0 \qquad \text{and} \qquad \mathrm{d}H_0^c = \lambda_1(\mathbf{X}, Y_1)\,\mathrm{d}H_1.$$

By (5) and (6), the distributions (G_0, G_1) are related to (H_0, H_1) by

$$\mathrm{d}G_0 = (1 - \pi(\mathbf{X}))\,\mathrm{d}H_0 + \pi(\mathbf{X})\,\mathrm{d}H_1^c$$

and

$$\mathrm{d}G_1 = (1 - \pi(\mathbf{X}))\,\mathrm{d}H_0^c + \pi(\mathbf{X})\,\mathrm{d}H_1.$$

If λ_0 and λ_1 were known, then (H_0^c, H_1^c) could be estimated by substituting $(\hat{G}_0, \hat{G}_1)$ for (H_0, H_1), and (G_0, G_1) could be estimated similarly. The estimated distributions can be examined in various ways. The marginal distributions of $\mathbf{X}$ from H_{1-t}^c and H_t should be similar. Comparison of the marginal distributions of Y_t from H_{1-t}^c and H_t reveals hidden bias. Moreover, comparison of the marginal distribution of Y_{1-t} from H_t^c and that of Y_t from H_t has a causal interpretation.

For sensitivity analysis, we consider several values of Λ and find bounds on $E[E(Y_t|T = 1 - t, \mathbf{X})] = \int y_t \lambda_t\,\mathrm{d}H_t$ under model (9) given each Λ. Although the model assumes no functional forms on λ_t, we need to take into account the fact that λ_t is a Radon–Nikodym derivative at each $\mathbf{x} \in \mathcal{X}$. By definition, $\lambda_t\,\mathrm{d}H_t$ and H_t induce the same marginal distribution on $\mathcal{X}$. Equivalently, λ_t satisfies

$$\int h(\mathbf{x})\lambda_t\,\mathrm{d}H_t(\mathbf{x}, y_t) = \int h(\mathbf{x})\,\mathrm{d}H_t(\mathbf{x}, y_t) \tag{10}$$

for each bounded function h on $\mathcal{X}$. If $\mathbf{X}$ has a continuous component, then infinitely many constraints are required to guarantee the marginal equality of $\lambda_t\,\mathrm{d}H_t$ and H_t on $\mathcal{X}$. Our method is to take a finite collection of constraints and obtain conservative bounds. The idea of using constraints (10) is similar to using constraints (1) in Section 3, but with an important difference. Here the constraint associated with $\hat{\pi}$ is on a more equal footing with others. The bounds depend on all of the constraints included and become no wider as more constraints are included in finite samples and asymptotically.

Let $\hat{\mathbf{h}}^c = (\hat{\pi}, 1 - \hat{\pi}, \hat{h}_1, \ldots, \hat{h}_{m^c})$ be $m^c + 2$ real-valued functions on $\mathcal{X}$. For a fixed value of Λ, we obtain bounds on $\int y_t \lambda_t\,\mathrm{d}H_t$ by solving the linear programming (LP)

$$\min \text{ or } \max \quad \int y_t \lambda_t\,\mathrm{d}\hat{G}_t$$

subject to

$$\int \lambda_t \mathrm{d}\hat{G}_t = 1,$$

$$\int \hat{\pi}(\mathbf{x})\lambda_t \,\mathrm{d}\hat{G}_t = \int \hat{\pi}(\mathbf{x})\,\mathrm{d}\hat{G}_t,$$

$$\int \hat{h}_j(\mathbf{x})\lambda_t \,\mathrm{d}\hat{G}_t = \int \hat{h}_j(\mathbf{x})\,\mathrm{d}\hat{G}_t, \qquad j = 1, \ldots, m^c,$$

and

$$\Lambda^{-1} \le \lambda_t \le \Lambda.$$

Each of these integrals is a finite sum because $\hat{G}_t$ is a discrete distribution supported on $\{(\mathbf{X}_i, Y_i)\colon i = 1, \ldots, n_1\}$ or $\{(\mathbf{X}_i, Y_i)\colon i = n_1 + 1, \ldots, n\}$. The unknowns are the values of λ_t on observed data, $\lambda_1(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n_1$, or $\lambda_0(\mathbf{X}_i, Y_i)$, $i = n_1 + 1, \ldots, n$. We evaluate the expectation $E(Y_t) = \int y_t \,\mathrm{d}G_t$ at the same extreme values of λ_t. Alternatively, bounds can be obtained by solving the corresponding linear programming problem, and the extreme values of λ_t can be different; see Section 4.1 for a special case.

Let $\hat{\lambda}_t$ be the solution to the sample LP. It is straightforward to show that $\hat{\lambda}_t$ takes at most $m^c + 2$ other values than Λ^{-1} and Λ. Let $\mathbf{h}^{*c} = (\pi^*, 1 - \pi^*, h_1^*, \ldots, h_{m^c}^*)$ be the asymptotic limit of $\hat{\mathbf{h}}^c$, and consider the linear programming

$$\text{min or max} \quad \int y_t \lambda_t \,\mathrm{d}H_t$$

subject to

$$\int \lambda_t \,\mathrm{d}H_t = 1,$$

$$\int \pi^*(\mathbf{x})\lambda_t \,\mathrm{d}H_t = \int \pi^*(\mathbf{x})\,\mathrm{d}H_t,$$

$$\int h_j^*(\mathbf{x})\lambda_t \,\mathrm{d}H_t = \int h_j^*(\mathbf{x})\,\mathrm{d}H_t, \qquad j = 1, \ldots, m^c,$$

and

$$\Lambda^{-1} \le \lambda_t \le \Lambda.$$

Let λ_t^* be the solution to the population LP. We conjecture that $\hat{\lambda}_t \,\mathrm{d}\hat{G}_t$ converges weakly to $\lambda_t^* \,\mathrm{d}H_t$ at rate $n^{-1/2}$, and plan to investigate the asymptotic theory in future work. The marginal distribution of $\lambda_t^* \,\mathrm{d}H_t$ on $\mathcal{X}$ can be different from the underlying $P(\mathbf{X})$, and $\lambda_0^* \,\mathrm{d}H_0$ and $\lambda_1^* \,\mathrm{d}H_1$ can be different from each other on $\mathcal{X}$. However, the minimum (or maximum) $\int y_t \lambda_t^* \,\mathrm{d}H_t$ is at least as small (or large) as what is strictly implied by model (9), because the constraints are contained in model (9). Our method generally involves incompatible distributions like that of Robins et al. (1999) but yields conservative bounds asymptotically.

The method in this section allows us to calculate bounds for a range of values of Λ in a study. There are two approaches to interpreting the bounds (see Robins 2002b; Rosenbaum 2002b). One approach is to find the smallest value Λ_s such that the qualitative conclusions are altered, and report that the study becomes sensitive to unmeasured confounding at Λ_s. This approach exploits the fact that Λ is on the common scale of odds ratio and centers on an objective measure calculated from the data without reference to substantive meanings. The other approach is to decide a cutoff Λ_c below which the values of Λ are considered plausible, and examine how the inferences might change for Λ over the range of plausible values. It is essential to judge what values of Λ are plausible from substantive knowledge. But this task can be enormously difficult, because the meaning of Λ depends on the covariates $\mathbf{X}$ that are conditioned on in the analysis, as discussed by Scharfstein et al. (1999) and Robins (2002b) for related methods. We recommend carefully implementing either approach with its own caveat.

4.3 Mean Specification

Brumback et al. (2004) and Robins (1999) proposed a method for sensitivity analysis by quantifying unmeasured confounding or hidden bias through

$$c_t(\mathbf{X}) = E(Y_t|T = 1 - t, \mathbf{X}) - E(Y_t|T = t, \mathbf{X})$$

and generalized it to longitudinal studies. Here we consider a different implementation by specifying bounds rather than exact forms for this function, and highlight an interesting phenomenon associated with this method.

Suppose that unmeasured confounding is such that $\Delta_t^- \le c_t(\mathbf{X}) \le \Delta_t^+$, where Δ_t^- and Δ_t^+ are sensitivity parameters ($t = 0, 1$). This specification contains the functional forms c_1, c_2, and c_3 of Brumback et al. (2004, sec. 3.1). For example, $\Delta_1^+ = 0$ and $\Delta_1^- < 0$ indicate that treated subjects on average have larger outcomes by as much as $|\Delta_1^-|$ than untreated subjects would have under the treatment at each level of $\mathbf{X}$. By the bias formula following (7) and (8), $E(Y_t)$ is bounded by $E[E(Y_t|T = t, \mathbf{X})] + \Delta_t^{\pm} E[P(T = 1 - t|\mathbf{X})]$, which can be estimated by

$$\tilde{\mu}_t + \Delta_t^- \tilde{p}_{1-t} \le \mu_t \le \tilde{\mu}_t + \Delta_t^+ \tilde{p}_{1-t},$$

where $\tilde{p}_1 = n^{-1} \sum_{i=1}^n T_i$ and $\tilde{p}_0 = 1 - \tilde{p}_1$. Similarly, bounds can be obtained for $\mu_1 - \mu_0$. Consider the case where $\Delta_0^- = \Delta_1^+ = 0$ and $\Delta_0^+ = -\Delta_1^- = \Delta \ge 0$ (i.e., treated subjects tend to have larger outcomes under either treatment). It follows that the inequality

$$\{\tilde{\mu}_1 - \tilde{\mu}_0 - z_{\alpha/2}\,\mathrm{se}(\tilde{\mu}_1 - \tilde{\mu}_0)\} - \Delta \le \mu_1 - \mu_0 \le \tilde{\mu}_1 - \tilde{\mu}_0 + z_{\alpha/2}\,\mathrm{se}(\tilde{\mu}_1 - \tilde{\mu}_0)$$

holds with probability $\ge 1 - \alpha$ asymptotically. There is significant evidence for a positive treatment effect if the lower confidence bound (in the curly brackets) is positive at $\alpha = 95\%$ and if no confounding is assumed ($\Delta = 0$). Moreover, the evidence would persist as long as hidden bias Δ were considered smaller than the lower confidence bound.

The distributional and mean specifications represent two different views on unmeasured confounding. The function $\lambda_t(\mathbf{X}, Y_t)$ is on a unit-free scale with two useful interpretations as density ratio and odds ratio, whereas the function $c_t(\mathbf{X})$ directly gives hidden bias in the scale of outcomes. Choosing plausible values is difficult for both cases.

5. DATA ANALYSIS

Right heart catheterization (RHC) is a medical procedure performed daily in hospitals since the 1970s. Many physicians

believe that RHC leads to better patient outcomes; however, the benefit of RHC has not been demonstrated in a randomized clinical trial. Physicians could not ethically participate in such a trial or encourage a patient to participate if convinced that the procedure is beneficial. In this context, the observational study of Connors et al. (1996) was influential, raising the concern that RHC may not benefit patients and may in fact cause harm. The original analysis used propensity score matching and Rosenbaum and Rubin's (1983b) method. We analyze their data using the new methods and illustrate the values of our approach to causal inference.

The study of Connors et al. included 5,735 critically ill patients admitted to the intensive care units of 5 medical centers. For each patient, treatment status T (=1 if RHC was used within 24 hours of admission and 0 otherwise), health outcome Y (survival time up to 30 days), and a comprehensive list of 75 covariates $\mathbf{X}$ (specified by a panel of 7 specialists in critical care) were recorded. Comparing the covariates between the 2,184 patients managed with RHC and the 3,551 patients without RHC suggests that the two groups differ significantly in many of the covariates; see Figure 2 for the histograms of three covariates discussed in Connors et al. (1996).

First, we fit the propensity score (i.e., the association of RHC with 75 covariates) using logit regression. The first model includes a constant term and the main effects of 75 covariates. As noted in Section 3, we calculate the statistic (4) over its standard error to check the adequacy of the model. Here the $\hat{h}_j$'s are taken to be the components of $\hat{\pi}(\mathbf{X})(1, \mathbf{X})$ and $(1 - \hat{\pi}(\mathbf{X}))(1, \mathbf{X})$; see the left and right boxplots above model 1 in Figure 1. The z-ratios are overly large in absolute value, indicating that the first model does not give a good fit. A sequence of models are then constructed stepwise by adding interactions of 75 covariates; Figure 1 shows the z-ratios from these models. The distributions of z-ratios indicate a gradual improvement. For the final model, the z-ratios appear to be reasonably distributed as standard normal, indicating that the model gives a satisfactory fit. Overall, patients managed with RHC have higher propensity scores (i.e., higher probabilities of receiving RHC) than those without RHC. Nevertheless, there is a considerable overlap in these probabilities between the two groups.

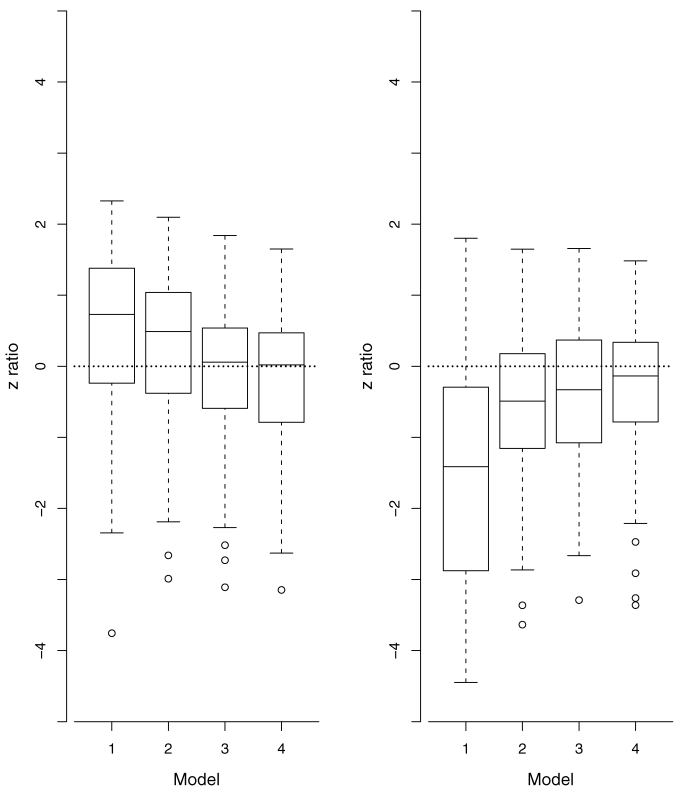

Figure 1. Checking Propensity Score Models.

The fitted propensity score varies from patient to patient. Therefore, the distribution of observed outcome from the treated (or untreated) group is a distortion of that of potential outcome that would be observed had each patient received RHC (or none had received RHC). For the simple choice $\hat{\mathbf{h}}^{(1)} = (\hat{\pi}, 1 - \hat{\pi})$, the method in Section 3 is used to recover the joint distribution of covariates and each potential outcome. Two copies of the marginal distribution of each covariate can be extracted on the treated and untreated groups; see Figure 2 for the weighted histograms of three covariates. Each pair of corresponding histograms agree with one another, indicating that the covariates are balanced across the treated and untreated groups after weighting. The marginal distribution of each potential outcome can also be extracted; Figure 2 presents the weighted survival curves together with the raw survival curves. It is interesting that the survival curve with RHC is always lower than that without RHC even after adjustment.

We consider other choices of $\hat{\mathbf{h}}$ for variance and bias reduction (see Hirano and Imbens 2002 for a related analysis). First, two linear regression models of 30-day survival (i.e., $Y \geq 30$) on 75 covariates are fit to the treated group and the untreated group separately. Six most significant covariates are identified from the two models: 2-month predicted survival probability (X_1), Duke activity status index (X_2), do-not-resuscitate status (X_3), bilirubin (X_4), primary disease category coma (X_5), and second disease category cirrhosis (X_6). Second, two logit regression models are fit to the two groups separately. With $\hat{\mathbf{h}}^{(1)} = (\hat{\pi}, 1 - \hat{\pi}, \hat{\pi}\mathbf{g}_0, (1 - \hat{\pi})\mathbf{g}_1)$, the average causal effect of RHC on 30-day survival is estimated for each of the following choices: (a) $X_1, \ldots, X_6$ are sequentially added to $\mathbf{g}_0$ and $\mathbf{g}_1$, (b) the fitted values from the two linear models are taken to be $\mathbf{g}_0$ and $\mathbf{g}_1$, and (c) the fitted values from the two logit models are taken to be $\mathbf{g}_0$ and $\mathbf{g}_1$; see Table 4. Different choices of $\hat{\mathbf{h}}^{(1)}$ lead to similar estimates. The standard errors show a minor reduction after $\hat{\mathbf{h}}^{(1)}$ includes $(\hat{\pi}, 1 - \hat{\pi})$ but are considerably smaller than that of the IPW estimate.

In the foregoing estimation, it is assumed that two patients with the same measured covariates (or, equivalently, the same propensity score) have the same chance of receiving RHC regardless of their unmeasured characteristics. We use the method in Section 4 to investigate the range of estimates under various deviations from this assumption. In such cases, the distribution of $(\mathbf{X}, Y_t)$ estimated previously converges not to $P(Y_t|\mathbf{X})P(\mathbf{X})$, but rather to $H_t = P(Y_t|T = t, \mathbf{X})P(\mathbf{X})$. For several values of Λ, extreme values of the expectation of Y_t from $H^c_{1-t} = P(Y_t|T = 1 - t, \mathbf{X})P(\mathbf{X})$ are obtained by linear programming with $\hat{\mathbf{h}}^c = (\hat{\pi}, 1 - \hat{\pi}, \hat{\pi}\mathbf{X}, (1 - \hat{\pi})\mathbf{X})$; see Table 5. The marginal distribution of each covariate from H^c_{1-t} (or, more relevantly, H^c_t) can be compared with the corresponding one from H_t; see Figure 3 for the weighted histograms of three covariates.

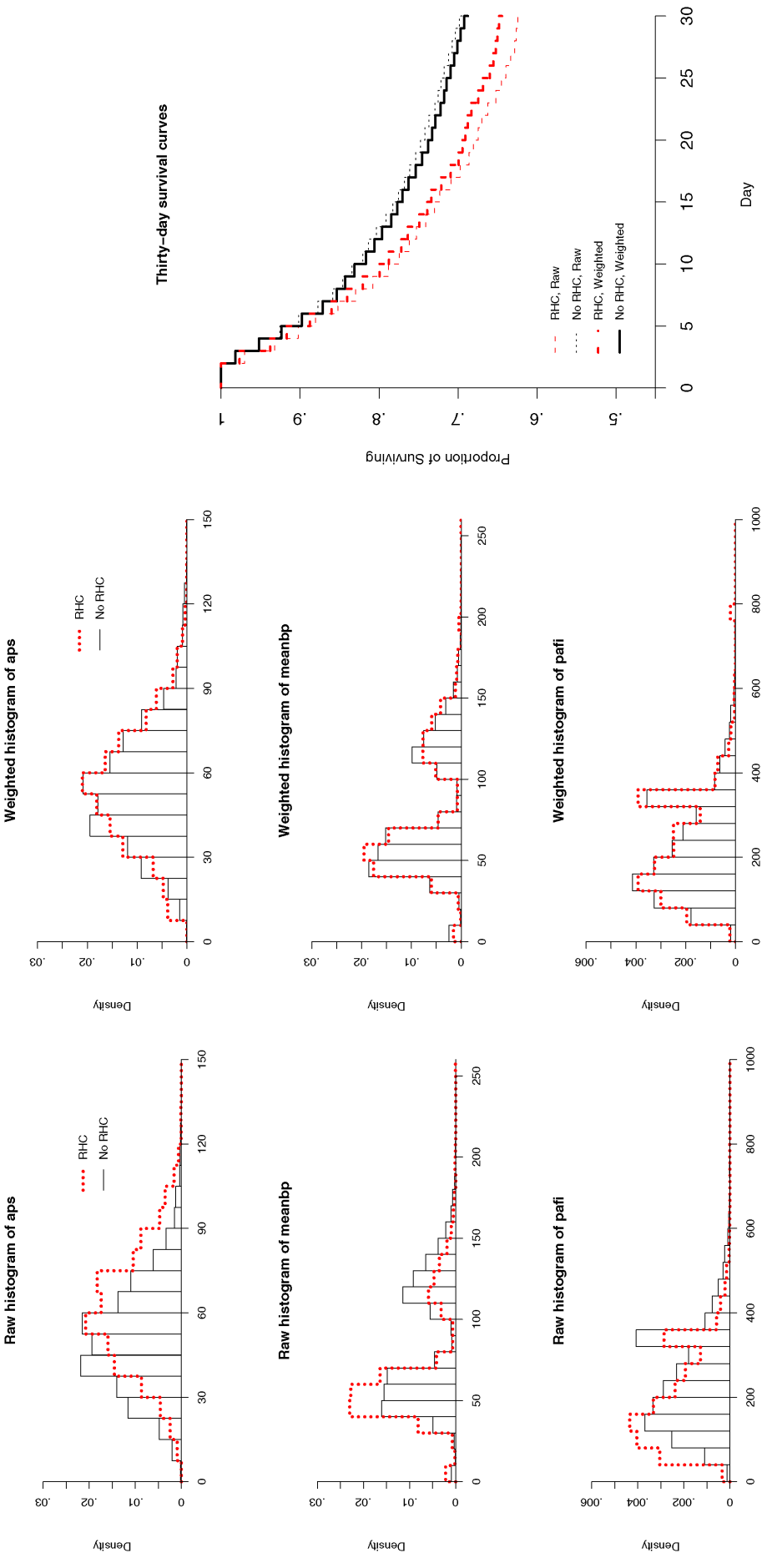

Figure 2. No-Confounding Estimation.

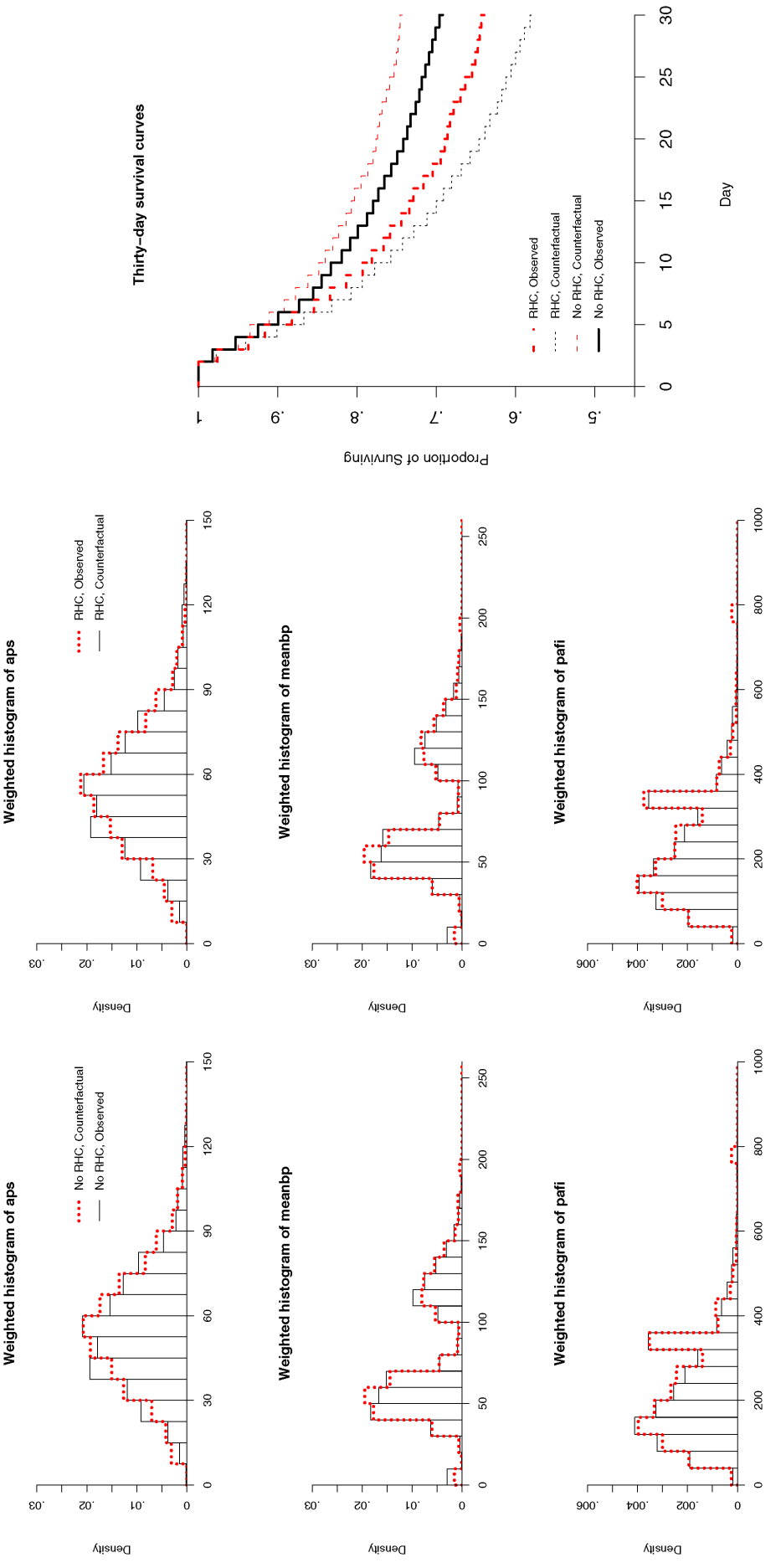

Figure 3. Sensitivity Analysis.

Table 4. No-Confounding Estimation

	REG		LIK				
	Diff	SE	Diff	SE		Diff	SE
1	−.0433	.0159	−.0433	.0158	IPW	−.0461	.0268
$1 + X_1$	−.0400	.0150	−.0406	.0158	Linear		
$1 + X_1 + X_2$	−.0404	.0149	−.0407	.0157	REG	−.0507	.0150
$1 + X_1 + X_2 + X_3$	−.0465	.0159	−.0487	.0155	LIK	−.0504	.0150
$1 + X_1 + \cdots + X_4$	−.0474	.0160	−.0491	.0154	Logit		
$1 + X_1 + \cdots + X_5$	−.0476	.0159	−.0494	.0153	REG	−.0510	.0149
$1 + X_1 + \cdots + X_6$	−.0476	.0159	−.0485	.0153	LIK	−.0522	.0148

NOTE: The estimates are based on $\hat{\mu}_t^{(m)}$ and $\tilde{\mu}_t^{(m)}$, with standard errors from Theorem 6(a) and (b). The raw estimate is −.0736. The estimates based on fitted values are −.0626 from linear regression and −.0642 from logit regression.

For $\Lambda = 1.5$, the odds of receiving RHC could differ from that implied by the propensity score by at most a factor of 1.5. The counterfactual probability of 30-day survival might be as small as 57.5% had RHC been withheld from patients with RHC (from H_1^c), and as large as 74.3% had RHC been applied to patients without RHC (from H_0^c). In comparison, the observed probability of 30-day survival is estimated as 64.0% for patients with RHC (from H_1) and 69.2% for patients without RHC (from H_0). Figure 3 compares the survival curves from H_0^c and H_1^c with those from H_0 and H_1. The curve from H_1 is below that from H_0, but H_1 is above H_1^c and H_0^c is above H_0. The appearance of a harmful effect of RHC could be explained away by unmeasured confounding such that sicker patients are more likely to receive RHC by a factor of $\Lambda^2 = 2.25$ in odds than healthier patients with the same covariates. This result is based on fewer parametric assumptions and is slightly more conservative than the results of Connors et al. (1996) and Lin et al. (1998).

6. SUMMARY

We have adopted Rubin's potential outcomes framework for causal inference and proposed two methods serving complementary purposes. One of these methods can be used to estimate average causal effects, assuming no confounding given measured covariates. The other can be used to assess how the estimates might change under various departures from no confounding. Both methods are developed from a nonparametric likelihood perspective. We illustrate the methods by analyzing the data from the observational study of Connors et al. (1996).

In this article we focus on the setting where treatment is dichotomous, the average causal effect over the population is sought, the outcome and covariates are completely recorded, and the sample is iid. It remains for future work to relax these assumptions and develop appropriate methods. For example, the average causal effects over subpopulations give more specific information. Given a discrete covariate, the conditional expectation of each potential outcome can be estimated by the ratio of unconditional expectations. However, the direct estimates are unsatisfactory for subpopulations defined by a continuous covariate or several covariates. It is desirable to incorporate parametric or semiparametric specifications along the line of Robins's (1999) marginal structural models.

APPENDIX: PROOFS

Proof of Theorem 1

The set of $\boldsymbol{\lambda}$'s such that $\ell_n(\boldsymbol{\lambda})$ is finite is nonempty and open. On this set, $\ell_n(\boldsymbol{\lambda})$ is strictly concave, because the log function is strictly concave and the vectors $(1, \ldots, 1)$, $(\pi^*(\mathbf{X}_1), \ldots, \pi^*(\mathbf{X}_n))$, $(h_j^*(\mathbf{X}_1), \ldots, h_j^*(\mathbf{X}_n))$, $j = 1, \ldots, m$, are linearly independent. By assumption, ℓ_n achieves a unique maximum at $\hat{\boldsymbol{\lambda}}$. Thus its first-order derivatives of $\ell_n(\boldsymbol{\lambda})$ must be 0 at $\hat{\boldsymbol{\lambda}}$. It follows that the positive numbers

$$\hat{w}_{1i} = \frac{1}{\hat{\boldsymbol{\lambda}}^\top \mathbf{h}^*(\mathbf{X}_i)}, \qquad i = 1, \ldots, n_1,$$

and

$$\hat{w}_{0i} = \frac{1}{1 - \hat{\boldsymbol{\lambda}}^\top \mathbf{h}^*(\mathbf{X}_i)}, \qquad i = n_1 + 1, \ldots, n,$$

satisfy the constraints

$$\sum_{i=1}^{n_1} w_{1i} = 1, \qquad \sum_{i=n_1+1}^{n} w_{0i} = 1,$$

$$\sum_{i=1}^{n_1} \pi^*(\mathbf{X}_i) w_{1i} = \sum_{i=n_1+1}^{n} \pi^*(\mathbf{X}_i) w_{0i},$$

and

$$\sum_{i=1}^{n_1} h_j^*(\mathbf{X}_i) w_{1i} = \sum_{i=n_1+1}^{n} h_j^*(\mathbf{X}_i) w_{0i}, \qquad j = 1, \ldots, m.$$

Moreover, Jensen's inequality implies that for any positive numbers w_{1i}, $i = 1, \ldots, n_1$, and w_{0i}, $i = n_1 + 1, \ldots, n$, satisfying the foregoing

Table 5. Sensitivity Analysis

Λ	$Y_1/T = 1, \boldsymbol{X}$	$Y_1/T = 0, \boldsymbol{X}$	$Y_0/T = 1, \boldsymbol{X}$	$Y_0/T = 0, \boldsymbol{X}$	Y_1	Y_0
1	.640	.640	.692	.692	.640	.692
1.2	.640	[.585, .690]	[.641, .738]	.692	[.608, .669]	[.672, .709]
1.5	.640	[.514, .743]	[.575, .789]	.692	[.567, .700]	[.646, .729]
2	.640	[.421, .801]	[.482, .843]	.692	[.514, .734]	[.612, .750]

NOTE: The results are based on $(\hat{G}_0^{(m)}, \hat{G}_1^{(m)})$, where the fitted values from the logit regression models are used in $\hat{\mathbf{h}}^{(1)}$. Standard errors of differences, say .575 versus .640 discussed in the text, are approximately .014–.022.

constraints,

$$\frac{1}{n}\sum_{i=1}^{n_1}\log\frac{w_{1i}}{\hat{w}_{1i}}+\frac{1}{n}\sum_{i=n_1+1}^{n}\log\frac{w_{0i}}{\hat{w}_{0i}}$$
$$\le\log\left[\frac{1}{n}\sum_{i=1}^{n_1}\frac{w_{1i}}{\hat{w}_{1i}}+\frac{1}{n}\sum_{i=n_1+1}^{n}\frac{w_{0i}}{\hat{w}_{0i}}\right]$$
$$=\log(1)=0.$$

The equality holds if and only if $w_{1i}=\hat{w}_{1i}$, $i=1,\ldots,n_1$, and $w_{0i}=\hat{w}_{0i}$, $i=n_1+1,\ldots,n$. Thus these weights give the unique constrained MLE $(\hat{G}_0,\hat{G}_1)$.

Proof of Theorem 2

For convenience, we show the equivalent results in which $\mathbf{h}^*$ is defined as $(1,\pi^*,h_1^*,\ldots,h_m^*)$. Consider the criterion function

$$\ell(\boldsymbol{\lambda})=E\big[T\log(\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))+(1-T)\log(1-\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))\big].$$

Then $\ell(\boldsymbol{\lambda})$ is finite in a neighborhood Λ_0 of $\boldsymbol{\lambda}_0=(0,1,0,\ldots,0)^\top$ such that $\boldsymbol{\lambda}^\top\mathbf{h}^*$ is sufficiently close to $\boldsymbol{\lambda}_0^\top\mathbf{h}^*=\pi^*$ and contained in $[\delta/2,1-\delta/2]$ on $\mathcal{X}$. For each fixed value of $\mathbf{X}$, $T\log(\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))+(1-T)\log(1-\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))$ is concave in $\boldsymbol{\lambda}$. Under model S0 (i.e., $\pi^*=\pi$), $\ell(\boldsymbol{\lambda})$ achieves a unique maximum at $\boldsymbol{\lambda}_0$ due to Jensen's inequality

$$E\left[\log\left\{\frac{(\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))^T(1-\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))^{1-T}}{\pi(\mathbf{X})^T(1-\pi(\mathbf{X}))^{1-T}}\right\}\right]\le\log(1)=0.$$

The equality holds if and only if $\boldsymbol{\lambda}=\boldsymbol{\lambda}_0$, because the functions in $\mathbf{h}^*$ are linearly independent on $\mathcal{X}$. Clearly, $\hat{\boldsymbol{\lambda}}$ is defined by maximizing the sample version $\ell_n(\boldsymbol{\lambda})$, which converges to $\ell(\boldsymbol{\lambda})$ with probability 1. By direct calculations, we have

$$\frac{\partial\ell_n}{\partial\boldsymbol{\lambda}}=\tilde{E}\left[\frac{T\mathbf{h}^*(\mathbf{X})}{\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X})}-\frac{(1-T)\mathbf{h}^*(\mathbf{X})}{1-\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X})}\right]$$

and

$$\frac{\partial^2\ell_n}{\partial\boldsymbol{\lambda}^2}=-\tilde{E}\left[\frac{T\mathbf{h}^*(\mathbf{X})\mathbf{h}^{*\top}(\mathbf{X})}{(\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))^2}+\frac{(1-T)\mathbf{h}^*(\mathbf{X})\mathbf{h}^{*\top}(\mathbf{X})}{(1-\boldsymbol{\lambda}^\top\mathbf{h}^*(\mathbf{X}))^2}\right].$$

The second-order derivatives are uniformly bounded in the neighborhood Λ_0 because $\mathbf{h}^*$ is bounded by Δ and $\boldsymbol{\lambda}^\top\mathbf{h}^*$ is contained in $[\delta/2,\ 1-\delta/2]$ on $\mathcal{X}$. Moreover, the negative Hessian matrix converges with probability 1,

$$-\frac{\partial^2\ell_n}{\partial\boldsymbol{\lambda}^2}(\boldsymbol{\lambda}_0)\to E\left[\frac{\mathbf{h}^*(\mathbf{X})(T-\pi^*(\mathbf{X}))^2\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})(1-\pi^*(\mathbf{X}))^2}\right]=\mathbf{B}_1.$$

By the asymptotic theory of M-estimators from convex minimization (Niemiro 1992), $\hat{\boldsymbol{\lambda}}$ converges to $\boldsymbol{\lambda}_0$ with probability 1, and has the expansion

$$\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0=\mathbf{B}_1^{-1}\left[\frac{\partial}{\partial\boldsymbol{\lambda}}\ell_n(\boldsymbol{\lambda}_0)\right]+o_p\big(n^{-1/2}\big)$$
$$=\mathbf{B}_1^{-1}\tilde{E}\left[\frac{\mathbf{h}^*(\mathbf{X})(T-\pi^*(\mathbf{X}))}{\pi^*(\mathbf{X})(1-\pi^*(\mathbf{X}))}\right]+o_p\big(n^{-1/2}\big).$$

We show the results for $\hat{\mu}_1$ and $\tilde{\mu}_1$; those for $\hat{\mu}_0$ and $\tilde{\mu}_0$ follow similarly. Note that $\hat{\boldsymbol{\lambda}}^\top\mathbf{h}^*$ converges to π^* $(\ge\delta)$ on $\mathcal{X}$ with probability 1. Then $\hat{\boldsymbol{\lambda}}^\top\mathbf{h}^*/\pi^*$ converges uniformly to 1 on $\mathcal{X}$, and the right side of the inequality

$$\left|\tilde{E}\left[\frac{YT}{\hat{\boldsymbol{\lambda}}^\top\mathbf{h}^*(\mathbf{X})}\right]-\tilde{E}\left[\frac{YT}{\pi^*(\mathbf{X})}\right]\right|\le\left\|\frac{\pi^*}{\hat{\boldsymbol{\lambda}}^\top\mathbf{h}^*}-1\right\|_{\sup}\tilde{E}\left[\left|\frac{YT}{\pi^*(\mathbf{X})}\right|\right]$$

tends to 0 with probability 1. Thus $\hat{\mu}_1$ converges to μ_1 with probability 1. By a Taylor expansion about $\boldsymbol{\lambda}_0$, we obtain

$$\hat{\mu}_1=\tilde{E}\left[\frac{YT}{\pi^*(\mathbf{X})}\right]-\tilde{E}\left[\frac{YT\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})}\right](\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0)+o_p\big(n^{-1/2}\big)$$
$$=\tilde{E}\left[\frac{YT}{\pi^*(\mathbf{X})}\right]-\tilde{E}\left[\frac{YT\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})}\right]\mathbf{B}_1^{-1}\tilde{E}\left[\frac{\mathbf{h}^*(\mathbf{X})(T-\pi^*(\mathbf{X}))}{\pi^*(\mathbf{X})(1-\pi^*(\mathbf{X}))}\right]$$
$$+o_p\big(n^{-1/2}\big).$$

The remainder term in the first equation is

$$(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0)^\top\tilde{E}\left[\frac{YT\mathbf{h}^*(\mathbf{X})\mathbf{h}^{*\top}(\mathbf{X})}{(\boldsymbol{\lambda}^{\dagger\top}\mathbf{h}^*(\mathbf{X}))^3}\right](\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0)=o_p\big(n^{-1/2}\big),$$

where $\boldsymbol{\lambda}^\dagger$ lies between $\hat{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}_0$. Note that with probability 1,

$$\tilde{E}\left[\frac{YT\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})}\right]\to E\left[\frac{YT\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})}\right]=\mathbf{C}_1.$$

The first-order term is a regression estimator with a different estimator of $\boldsymbol{\beta}_1$ than $\tilde{\boldsymbol{\beta}}_1$. We conclude the proof because two consistent estimators of $\boldsymbol{\beta}_1$ yield equivalent regression estimators to first order.

Proof of Theorem 3

By direct calculations, we have

$$\tilde{\mu}_t=\tilde{E}(\eta_t^*)-\boldsymbol{\beta}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*)-(\tilde{\boldsymbol{\beta}}_t^\top-\boldsymbol{\beta}_t^\top)\tilde{E}(\boldsymbol{\xi}_t^*)$$
$$=\tilde{E}(\eta_t^*)-\boldsymbol{\beta}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*)-\tilde{E}[(\eta_t^*-\boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)\boldsymbol{\xi}_t^{*\top}]\tilde{E}^{-1}(\boldsymbol{\zeta}_t^*\boldsymbol{\xi}_t^{*\top})\tilde{E}(\boldsymbol{\xi}_t^*)$$
$$=\tilde{E}(\eta_t^*)-\boldsymbol{\beta}_t^\top\tilde{E}(\boldsymbol{\xi}_t^*)-\tilde{E}[(\eta_t^*-\boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)\boldsymbol{\xi}_t^{*\top}]E^{-1}(\boldsymbol{\zeta}_t^*\boldsymbol{\xi}_t^{*\top})E(\boldsymbol{\xi}_t^*)$$
$$+o_p\big(n^{-1/2}\big).$$

The second equation follows from

$$\tilde{\boldsymbol{\beta}}_t^\top-\boldsymbol{\beta}_t^\top=\tilde{E}(\eta_t^*\boldsymbol{\xi}_t^{*\top})\tilde{E}^{-1}(\boldsymbol{\zeta}_t^*\boldsymbol{\xi}_t^{*\top})-\boldsymbol{\beta}_t^\top$$
$$=\tilde{E}[(\eta_t^*-\boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)\boldsymbol{\xi}_t^{*\top}]\tilde{E}^{-1}(\boldsymbol{\zeta}_t^*\boldsymbol{\xi}_t^{*\top}),$$

and the third equation follows from $\tilde{E}^{-1}(\boldsymbol{\zeta}_t^*\boldsymbol{\xi}_t^{*\top})\tilde{E}(\boldsymbol{\xi}_t^*)-E^{-1}(\boldsymbol{\zeta}_t^*\times\boldsymbol{\xi}_t^{*\top})E(\boldsymbol{\xi}_t^*)=o_p(1)$ and $\tilde{E}[(\eta_t^*-\boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)\boldsymbol{\xi}_t^{*\top}]=O_p(n^{-1/2})$ because $E[(\eta_t^*-\boldsymbol{\beta}_t^\top\boldsymbol{\zeta}_t^*)\boldsymbol{\xi}_t^{*\top}]=0$.

Proof of Theorem 4

From the regularity conditions, it follows that $(\hat{\pi},\hat{\mathbf{h}})$ converges uniformly to $(\pi^*,\mathbf{h}^*)$ on $\mathcal{X}$ with probability 1. The proof is similar to that of Theorem 2. Although $\ell(\boldsymbol{\lambda})$ remains the same, the sample version $\ell_n(\boldsymbol{\lambda})$ becomes

$$\ell_n(\boldsymbol{\lambda})=\tilde{E}\big[T\log(\boldsymbol{\lambda}^\top\hat{\mathbf{h}}(\mathbf{X}))+(1-T)\log(1-\boldsymbol{\lambda}^\top\hat{\mathbf{h}}(\mathbf{X}))\big].$$

By the uniform convergence of $(\hat{\pi},\hat{\mathbf{h}})$, it follows that $\ell_n(\boldsymbol{\lambda})$ converges to $\ell(\boldsymbol{\lambda})$, the second-order derivatives of $\ell_n(\boldsymbol{\lambda})$ are uniformly bounded in a neighborhood of $\boldsymbol{\lambda}_0$, and

$$-\frac{\partial^2\ell_n}{\partial\boldsymbol{\lambda}^2}(\boldsymbol{\lambda}_0)\to E\left[\frac{\mathbf{h}^*(\mathbf{X})(T-\pi^*(\mathbf{X}))^2\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})(1-\pi^*(\mathbf{X}))^2}\right]=\mathbf{B}_1.$$

By the asymptotic theory of M-estimators from convex minimization (Niemiro 1992), $\hat{\boldsymbol{\lambda}}$ converges to $\boldsymbol{\lambda}_0$ with probability 1, and has the expansion

$$\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0=\mathbf{B}^{-1}\left[\frac{\partial}{\partial\boldsymbol{\lambda}}\ell_n(\boldsymbol{\lambda}_0)\right]+o_p\big(n^{-1/2}\big)$$
$$=\mathbf{B}^{-1}\tilde{E}\left[\frac{\hat{\mathbf{h}}(\mathbf{X})(T-\hat{\pi}(\mathbf{X}))}{\hat{\pi}(\mathbf{X})(1-\hat{\pi}(\mathbf{X}))}\right]+o_p\big(n^{-1/2}\big).$$

Note that $\hat{\boldsymbol{\lambda}}^\top \hat{\mathbf{h}}$ converges uniformly to π^* ($\geq \delta$) with probability 1. Then $\hat{\boldsymbol{\lambda}}^\top \hat{\mathbf{h}}/\pi^*$ converges uniformly to 1, and the right side of the inequality

$$\left|\tilde{E}\left[\frac{YT}{\hat{\boldsymbol{\lambda}}^\top \hat{\mathbf{h}}(\mathbf{X})}\right] - \tilde{E}\left[\frac{YT}{\pi^*(\mathbf{X})}\right]\right| \leq \left\|\frac{\pi^*}{\hat{\boldsymbol{\lambda}}^\top \hat{\mathbf{h}}} - 1\right\|_{\sup} \tilde{E}\left[\left|\frac{YT}{\pi^*(\mathbf{X})}\right|\right]$$

tends to 0 with probability 1. Thus $\hat{\mu}_1$ converges to μ_1 with probability 1. By a Taylor expansion about $\boldsymbol{\lambda}_0$, we obtain

$$\hat{\mu}_1 = \tilde{E}\left[\frac{YT}{\hat{\pi}(\mathbf{X})}\right] - \tilde{E}\left[\frac{YT\hat{\mathbf{h}}^\top(\mathbf{X})}{\hat{\pi}^2(\mathbf{X})}\right](\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + o_p\big(n^{-1/2}\big).$$

Note that with probability 1,

$$\tilde{E}\left[\frac{YT\hat{\mathbf{h}}^\top(\mathbf{X})}{\hat{\pi}^2(\mathbf{X})}\right] \to E\left[\frac{YT\mathbf{h}^{*\top}(\mathbf{X})}{\pi^{*2}(\mathbf{X})}\right] = \mathbf{C}_1.$$

Thus $\hat{\mu}_1$ has the first given expansion. It remains to show the asymptotic expansions

$$\tilde{E}(\hat{\eta}_1) = \tilde{E}(\Pi^\perp[\eta_1^*|\mathbf{s}^*]) + o_p\big(n^{-1/2}\big)$$

and

$$\tilde{E}(\hat{\boldsymbol{\xi}}_1) = \tilde{E}(\Pi^\perp[\boldsymbol{\xi}_1^*|\mathbf{s}^*]) + o_p\big(n^{-1/2}\big),$$

where $\Pi^\perp[\eta_1^*|\mathbf{s}^*] = \eta_1^* - \mathbf{W}_1\mathbf{V}^{-1}\mathbf{s}^*$, $\Pi^\perp[\boldsymbol{\xi}_1^*|\mathbf{s}^*] = \boldsymbol{\xi}_1^* - \mathbf{U}_1\mathbf{V}^{-1}\mathbf{s}^*$, $\mathbf{W}_1 = E[\eta_1^*\mathbf{s}^{*\top}]$, and $\mathbf{U}_1 = E[\boldsymbol{\xi}_1^*\mathbf{s}^{*\top}]$. By Taylor expansions about $\boldsymbol{\gamma}^*$, we obtain

$$\tilde{E}(\hat{\eta}_1) = \tilde{E}(\eta_1^*) - \tilde{E}\left[\frac{YT}{\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right](\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + o_p\big(n^{-1/2}\big)$$

and

$$\tilde{E}(\hat{\boldsymbol{\xi}}_1) = \tilde{E}(\boldsymbol{\xi}_1^*) - \tilde{E}\left[\frac{\mathbf{h}^*(\mathbf{X})T}{(1-\pi^*(\mathbf{X}))\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right](\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + o_p\big(n^{-1/2}\big).$$

The second expansion is simplified because the extra coefficient of $(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ is negligible,

$$\tilde{E}\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\left(\frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma}^*)}{1-\pi(\mathbf{X};\boldsymbol{\gamma}^*)}\right)\left(\frac{T}{\pi^*(\mathbf{X})} - 1\right)\right] \to E\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\left(\frac{\mathbf{h}(\mathbf{X};\boldsymbol{\gamma}^*)}{1-\pi(\mathbf{X};\boldsymbol{\gamma}^*)}\right)\left(\frac{T}{\pi^*(\mathbf{X})} - 1\right)\right] = 0.$$

Note that with probability 1,

$$\tilde{E}\left[\frac{YT}{\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right] \to E\left[\frac{YT}{\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right] = \mathbf{W}_1$$

and

$$\tilde{E}\left[\frac{\mathbf{h}^*(\mathbf{X})T}{(1-\pi^*(\mathbf{X}))\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right] \to E\left[\frac{\mathbf{h}^*(\mathbf{X})T}{(1-\pi^*(\mathbf{X}))\pi^{*2}(\mathbf{X})}\frac{\partial\pi(\mathbf{X};\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right] = \mathbf{U}_1.$$

It follows from the discussion in the text that $\hat{\mu}_1$ has the second given expansion.

Proof of Theorem 5

As in the proof of Theorem 3, we have

$$\begin{aligned}\tilde{\mu}_t &= \tilde{E}(\hat{\eta}_t) - \boldsymbol{\beta}_t^\top \tilde{E}(\hat{\boldsymbol{\xi}}_t) - \tilde{E}[(\hat{\eta}_t - \boldsymbol{\beta}_t^\top \hat{\boldsymbol{\xi}}_t)\hat{\boldsymbol{\xi}}_t^\top]\tilde{E}^{-1}(\hat{\boldsymbol{\xi}}_t\hat{\boldsymbol{\xi}}_t^\top)\tilde{E}(\hat{\boldsymbol{\xi}}_t)\\ &= \tilde{E}(\hat{\eta}_t) - \boldsymbol{\beta}_t^\top \tilde{E}(\hat{\boldsymbol{\xi}}_t) - \tilde{E}[(\hat{\eta}_t - \boldsymbol{\beta}_t^\top \hat{\boldsymbol{\xi}}_t)\hat{\boldsymbol{\xi}}_t^\top]E^{-1}(\boldsymbol{\xi}_t^*\boldsymbol{\xi}_t^{*\top})E(\boldsymbol{\xi}_t^*)\\ &\quad + o_p\big(n^{-1/2}\big).\end{aligned}$$

The result follows from the expansion

$$\begin{aligned}\tilde{E}(\hat{\eta}_t) &= \tilde{E}(\eta_t^*) + \tilde{E}(\partial\eta_t^*/\partial\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + o_p\big(n^{-1/2}\big)\\ &= \tilde{E}(\eta_t^*) + E(\partial\eta_t^*/\partial\boldsymbol{\gamma})\mathbf{V}^{-1}\tilde{E}(\mathbf{s}^*) + o_p\big(n^{-1/2}\big)\end{aligned}$$

and other similar expansions for $\tilde{E}(\hat{\boldsymbol{\xi}}_t)$ and $\tilde{E}[(\hat{\eta}_t - \boldsymbol{\beta}_t^\top \hat{\boldsymbol{\xi}}_t)\hat{\boldsymbol{\xi}}_t^\top]$.

Proof of Theorem 7

The proof is similar to those of Theorems 4 and 5. It is sufficient to show that under model S (i.e., $\pi^* = \pi$),

$$E\left[\frac{YT}{\pi^*(\mathbf{X})}\right] = E[E(Y_1|T=1,\mathbf{X})].$$

It follows from two applications of the rule of iterated expectations, $E(YT|\mathbf{X}) = E(Y\cdot 1|T=1,\mathbf{X})P(T=1|\mathbf{X}) + E(Y\cdot 0|T=0,\mathbf{X}) \times P(T=0|\mathbf{X}) = E(Y_1|T=1,\mathbf{X})\pi(\mathbf{X})$ and $E[YT/\pi^*(\mathbf{X})] = E[E(YT|\mathbf{X})/\pi^*(\mathbf{X})]$.

[Received August 2004. Revised October 2005.]

REFERENCES

Brumback, B. A., Hernan, M. A., Haneuse, S. J. P. A., and Robins, J. M. (2004), "Sensitivity Analysis for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures," *Statistics in Medicine*, 23, 749–767.

Connors, A. F., Speroff, T., Dawson, N. V., et al. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, 276, 889–897.

Cornfield, J., Haenszel, W., Hammond, E. C., Liliefeld, A., Shimkin, M., and Wynder, E. L. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: Wiley.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hammersley, J. M., and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Methuen.

Hirano, K., and Imbens, G. W. (2002), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores," *Econometrica*, 71, 1161–1189.

Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–970.

Holland, P. W., and Rubin, D. B. (1988), "Causal Inference in Retrospective Studies," *Evaluation Review*, 12, 203–231.

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

——— (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003), "A Theory of Statistical Models for Monte Carlo Integration" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 65, 585–618.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," translated in *Statistical Science*, 5, 465–480.

Niemiro, W. (1992), "Asymptotics for M-Estimators Defined by Convex Minimization," *The Annals of Statistics*, 20, 1514–1533.

Robins, J. M. (1999), "Association, Causation, and Marginal Structural Models," *Synthese*, 121, 151–179.

——— (2002a), Comment on "Using Inverse Weighting and Predictive Inference to Estimate the Effects of Time-Varying Treatments on the Discrete-Time Hazard," by R. Dawson and P. W. Lavori, *Statistics in Medicine*, 21, 1663–1680.

——— (2002b), Comment on "Covariance Adjustment in Randomized Experiments and Observational Studies," by P. R. Rosenbaum, *Statistical Science*, 17, 309–321.

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semiparametric Models," *Statistics in Medicine*, 16, 285–319.

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

——— (2001), Comment on "Inference for Semiparametric Models: Some Questions and an Answer," by P. J. Bickel and J. Kwon, *Statistica Sinica*, 11, 920–936.

Robins, J. M., Rotnitzky, A., and Bonetti, M. (2001), Comment on "Addressing an Idiosyncrasy in Estimating Survival Curves Using Double Sampling in the Presence of Self-Selected Right Censoring," by C. E. Frangakis and D. B. Rubin, *Biometrics*, 57, 343–347.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology*, eds. E. M. Halloran and D. Berry, New York: Springer-Verlag, pp. 1–92.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000), Comment on "On Profile Likelihood," by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association*, 95, 431–435.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

——— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rosenbaum, P. R. (1986), "Dropping Out of High School in the United States: An Observational Study," *Journal of Educational Statistics*, 11, 207–224.

——— (2002a), *Observational Studies* (2nd ed.), New York: Springer-Verlag.

——— (2002b), "Covariance Adjustment in Randomized Experiments and Observational Studies" (with discussion), *Statistical Science*, 17, 286–327.

Rosenbaum, P. R., and Rubin, D. B. (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1983b), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

Rotnitzky, A., and Robins, J. M. (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika*, 82, 805–820.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339.

Rotnitzky, A., Scharfstein, D. O., Su, T.-L., and Robins, J. M. (2001), "Methods for Conducting Sensitivity Analysis of Trials With Potentially Nonignorable Competing Causes of Censoring," *Biometrics*, 57, 103–113.

Rubin, D. B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203.

——— (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1977), "Assignment of Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 7, 34–58.

——— (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, 75, 591–593.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models" (with discussion), *Journal of the American Statistical Association*, 94, 1096–1146.

Tan, Z. (2004), "On a Likelihood Approach for Monte Carlo Integration," *Journal of the American Statistical Association*, 99, 1027–1036.

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer-Verlag.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.