Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models

Zhiqiang TAN*

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA

Key words and phrases: Causal inference; double robustness; estimating equations; nonparametric likelihood; profile likelihood; propensity score; marginal structural model; nested structural model.

MSC 2000: Primary 63G05; secondary 62J02.

Abstract: This article considers Robins's marginal and nested structural models in the cross-sectional setting and develops likelihood and regression estimators. First, a nonparametric likelihood method is proposed by retaining a finite subset of all inherent and modelling constraints on the joint distributions of potential outcomes and covariates under a correctly specified propensity score model. A profile likelihood is derived by maximizing the nonparametric likelihood over these joint distributions subject to the retained constraints. The maximum likelihood estimator is intrinsically efficient based on the retained constraints and weakly locally efficient. Second, two regression estimators, named hat and tilde, are derived as first-order approximations to the likelihood estimator under the propensity score model. The tilde regression estimator is intrinsically and weakly locally efficient and doubly robust. The methods are illustrated by data analysis for an observational study on right heart catheterization. *The Canadian Journal of Statistics* 38: 609–632; 2010 © 2010 Statistical Society of Canada

Résumé: Cet article considère les modèles structurels emboîtés et marginaux de Robin dans un contexte transversal et il développe des estimateurs du maximum de vraisemblance et de régression. Dans un premier temps, une méthode de vraisemblance non paramétrique est proposée en utilisant un sous-ensemble fini des contraintes inhérentes et de modélisation sur les distributions conjointes des résultats potentiels et des covariables sous un modèle de cote de propension correctement spécifié. Une fonction de vraisemblance de profil est obtenue en maximisant la vraisemblance non paramétrique sous ces distributions conjointes soumises aux contraintes retenues. L'estimateur du maximum de vraisemblance est intrinsèquement efficace sous les contraintes retenues et faiblement localement efficace. Dans un second temps, deux estimateurs de régression, dénommés chapeau et tilde, sont obtenus comme une approximation du premier ordre de l'estimateur du maximum de vraisemblance sous le modèle de cote de propension. L'estimateur tilde est intrinsèquement et faiblement localement efficace et doublement robuste. Les méthodes sont illustrées à l'aide d'une analyse de données provenant d'une étude observationnelle sur le cathétérisme cardiaque droit. *La revue canadienne de statistique* 38: 609–632; 2010 © 2010 Société statistique du Canada

1. INTRODUCTION

Drawing inferences about treatment effects is of interest in economics, epidemiology, and other fields. We adopt the framework of potential outcomes (Neyman, 1923; Rubin, 1974) and consider marginal and nested structural models of Robins (1999) in the cross-sectional setting. Robins and others proposed augmented inverse-probability-weighted (IPW) and G estimators, depending on a propensity score model and an outcome regression model such that they are weakly locally efficient and doubly robust. See Section 2 for a review and Bang & Robins (2005) and Kang

* *Author to whom correspondence may be addressed.*
E-mail: ztan@stat.rutgers.edu

& Schafer (2007) among others for recent works on doubly robust estimation. We develop new estimators in two different but related directions.

First, we propose a nonparametric likelihood method under a correctly specified propensity score model. All previous propensity score methods are based on estimating equations. In fact, *the usual likelihood* is decomposed into a factor depending only on the parameters of the propensity score model and another only on the structural parameters. Then an estimator obeying the likelihood principle should not depend on the propensity score. Robins & Ritov (1997) showed that such estimators cannot in general be uniformly consistent or attain an algebraic rate of convergence. However, our likelihood method circumvents this critique and yields regular consistent estimators under standard regularity conditions. The key idea is that we retain a finite subset of all available constraints and ignore other constraints on the joint distributions of potential outcomes and covariates. Effectively, we obtain an expanded model and *a new likelihood*. Our likelihood method is no longer the type of likelihood methods considered by Robins & Ritov (1997).

The resulting maximum likelihood estimator is intrinsically efficient in the sense that it is asymptotically equivalent to the optimal estimator in the class of estimators that are solutions to linear combinations of estimating equations based on the retained constraints. Moreover, the profile likelihood has asymptotic properties that are prototypical for profile likelihoods in parametric models and also in a number of semiparametric models. See Murphy & van der Vaart (2000) and its discussions.

Second, we derive two regression estimators, named hat and tilde, as first-order approximations to the likelihood estimator under the propensity score model. The likelihood estimator and the two regression estimators are intrinsically and weakly locally efficient. However, the tilde regression estimator is constructed to always be doubly robust, whereas the likelihood estimator and the hat regression estimator are not so (see Section 3.6 for a special case). Therefore, the tilde regression estimator has three desirable asymptotic properties (i.e., intrinsic and weakly local efficiency and double robustness). By intrinsic efficiency, the tilde regression estimator is asymptotically guaranteed to gain efficiency over the nonaugmented IPW or G estimator if the propensity score model is correct and over the augmented IPW (AIPW) or augmented G (AG) estimator if the outcome regression model is misspecified but the propensity score model is correct.

The rest of this article is organized as follows. Section 2 reviews marginal and nested structural models and existing estimators. Section 3 develops the likelihood and regression estimators and related asymptotic results. Section 4 presents a simulation study. Section 5 provides data analysis for an observational study on right heart catheterization (RHC). All proofs are collected in the Appendix in the Supplementary Material.

2. REVIEW

For a population, let T be a treatment variable, Y an outcome variable, and X a vector covarites. Suppose that T takes values in $\mathcal{T} = \{0, 1, \ldots, J-1\}$ with $J \geq 2$, where the value "0" denotes the null treatment (or placebo). For $t \in \mathcal{T}$, let Y_t be the potential outcome that would be observed under treatment t. The average causal effect of treatment t is $\Delta_t = E(Y_t) - E(Y_0)$. Throughout, we make the consistency assumption that $Y = Y_T$ and the assumption of no unmeasured confounding that for each $t \in \mathcal{T}$, Y_t and $1\{T = t\}$ are conditionally independent given X, that is, $P(Y_t | T = t, X) = P(Y_t | X)$.

Suppose that an independent and identically distributed (i.i.d.) sample of n units is selected. The observed data (T_i, Y_i, X_i), $i = 1, \ldots, n$, are i.i.d. from the joint distribution of (T, Y, X). There are two dimension-reduction approaches for consistent estimation of the average causal effect Δ_t. First, the outcome regression approach is based on modelling the conditional distribution

$P(Y|T=t, X)$ or the conditional mean $\mu(t, X) = E(Y|T=t, X)$. Consider a regression model

$$E(Y|T=t, X) = \mu(t, X; \alpha) = \Psi[g_0(X; \alpha_0) + g_1(t, X; \alpha_1)], \tag{1}$$

where Ψ is a link function, $g_0(X; \alpha_0)$ and $g_1(t, X; \alpha_1)$ are known functions such that $g_1(0, X; \alpha_1) = 0$, and $\alpha = (\alpha_0^\top, \alpha_1^\top)^\top$ is a vector of parameters. If outcomes are subject to censoring as in the example in Section 5, then a model more restrictive than (1) is typically needed to achieve identification of α. Consider a regression model with i.i.d. disturbances

$$Y = \mu(T, X; \alpha) + \varepsilon = g_0(X; \alpha_0) + g_1(T, X; \alpha_1) + \varepsilon, \tag{2}$$

where ε has mean 0 and is independent of (T, X). Model (2) is often called an accelerated failure time (AFT) model with Y the log of survival time. Write $\hat{\mu}(t, X) = \mu(t, X; \hat{\alpha})$, where $\hat{\alpha}$ is the least-squares estimator $\hat{\alpha}_{\text{LS}}$ in model (1) or the Buckley & James (1979) estimator $\hat{\alpha}_{\text{BJ}}$ in model (2) in the presence of right-censoring. If model (1) or (2) is correctly specified, then Δ_t can be estimated at rate $n^{-1/2}$ by $\tilde{E}[\hat{\mu}(t, X)] - \tilde{E}[\hat{\mu}(0, X)]$, where $\tilde{E}(\cdot)$ denotes sample average. Second, the propensity score approach is based on modelling the propensity score $\pi(t, X) = P(T = t|X)$ (Rosenbaum & Rubin, 1983). Consider a regression model

$$P(T = t|X) = \pi(t, X; \gamma), \tag{3}$$

where $\pi(t, X; \gamma)$ is a known function, and γ is a vector of parameters. Write $\hat{\pi}(t, X) = \pi(t, X; \hat{\gamma})$, where $\hat{\gamma}$ is the maximum likelihood estimator (MLE) of γ. If model (3) is correctly specified, then Δ_t can be estimated at rate $n^{-1/2}$ by $\tilde{E}[1\{T = t\}Y/\hat{\pi}(t, X)] - \tilde{E}[1\{T = 0\}Y/\hat{\pi}(0, X)]$. See Tan (2006) for other estimators of Δ_t.

To estimate average causal effects over subpopulations classified by a subvector of X, Robins and others proposed marginal and nested structural models and developed appropriate estimators in cross-sectional and longitudinal studies (e.g., van der Laan & Robins, 2003). We review these models and estimators in the cross-sectional setting.

2.1. Marginal and Nested Structural Models

Let V be a subvector of X. A marginal structural model imposes restrictions on the (marginal) distribution of Y_t in subpopulations classified by V, and hence is a regression model of Y_t against V. For example, a marginal structural mean model is

$$E(Y_t|V) = \nu(t, V; \theta) = \Psi[c_0(V; \theta_0) + c_1(t, V; \theta_1)], \tag{4}$$

where Ψ is a link function, $c_0(V; \theta_0)$ and $c_1(t, V; \theta_1)$ are known functions such that $c_1(0, V; \theta_1) = 0$, and $\theta = (\theta_0^\top, \theta_1^\top)^\top$ is a vector of parameters. A marginal structural distribution model with the identity link (or marginal structural AFT model) is $Y_t = c_0(V; \theta_0) + c_1(t, V; \theta_1) + \delta_t$, where δ_t ($t \in \mathcal{T}$) are identically distributed with mean 0 and independent of V.

A nested structural model parameterizes a link between the distribution of Y_t and that of Y_0, but otherwise placing no restriction on either distribution, in subpopulations classified by V. For example, an additive nested structural mean model is

$$E(Y_t|V) = E(Y_0|V) + c_1(t, V; \theta_1), \tag{5}$$

and a multiplicative nested structural mean model is

$$E(Y_t|V) = E(Y_0|V) \exp[c_1(t, V; \theta_1)], \tag{6}$$

where $c_1(t, V; \theta_1)$ is a known function such that $c_1(0, V; \theta_1) = 0$, indicating additive or multiplicative causal effects, and θ_1 is a vector of parameters. Model (5) or (6) can be equivalently expressed as $E(Y_t|V) = \Psi[\theta_0(V) + c_1(t, V; \theta_1)]$, where Ψ is the identity or exponential link, and $\theta_0(V)$ is an unknown function. An additive nested structural distribution model (or nested structural AFT model) is

$$P(Y_t|V) = P(Y_0 + c_1(t, V; \theta_1)|V). \tag{7}$$

That is, Y_t has the same distribution as Y_0 shifted by $c_1(t, V; \theta_1)$ given V. No restrictions other than (7) are imposed on the conditional distributions of Y_t, $t \in \mathcal{T}$.

2.2. Inverse Weighting and G Estimation

To estimate the parameters of structural models, a fundamental difficulty is that the full data $[(Y_t)_{t\in\mathcal{T}}, X]$ are not completely observed, but T as a coarsening variable and (Y, X) as coarsened data are observed. The methods of inverse weighting and G estimation address this difficulty by providing mappings from estimating functions based on the full data to estimating functions based on the observed data.

Let ϑ be a $d \times 1$ vector of parameters in a structural model, such as θ in the marginal structural mean model (4) or θ_1 in the nested structural mean model (5) or (6). Assume that a vector of unbiased full-data estimating functions is available in the form

$$\tau_{\text{full}}(\vartheta) = \sum_{t\in\mathcal{T}} \eta(t, Y_t, X; \vartheta),$$

where $\eta(t, Y_t, X; \vartheta)$ is a $d \times 1$ vector of functions such that $\sum_{t\in\mathcal{T}} E[\eta(t, Y_t, X; \vartheta)] = 0$. See three examples of η presented later in this section. Given a propensity score model (3), the vector of observed-data estimating functions initially mapped from $\tau_{\text{full}}(\vartheta)$ by inverse weighting is

$$\hat{\tau}_{\text{init}}(\vartheta) = \frac{\eta(T, Y, X; \vartheta)}{\hat{\pi}(T, X)},$$

and the initial, nonaugmented estimator $\hat{\vartheta}_{\text{init}}$ is a solution to $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta)] = 0$, where $\hat{\pi}(t, X) = \pi(t, X; \hat{\gamma})$ defined below display (3). Propensity score weighting can be augmented by outcome regression to enhance efficiency and robustness. Assume that a parametric or semiparametric model is specified for $P(Y|T = t, X)$ such as model (1) or (2). The vector of augmented observed-data estimating functions is

$$\hat{\tau}_{\text{aug}}(\vartheta) = \hat{\tau}_{\text{init}}(\vartheta) - \left\{ \frac{\hat{E}(\eta|T, X)}{\hat{\pi}(T, X)} - \sum_{t\in\mathcal{T}} \hat{E}(\eta|T = t, X) \right\},$$

and the augmented estimator $\hat{\vartheta}_{\text{aug}}$ is a solution to $\tilde{E}[\hat{\tau}_{\text{aug}}(\vartheta)] = 0$, where $\hat{E}(\eta|T = t, X)$ is an estimator of $E(\eta|T = t, X)$ based on the outcome regression model. A general guideline for obtaining $\hat{E}(\eta|T = t, X)$ is to express $E(\eta|T = t, X)$ as a function of the parameters of the outcome regression model and then substitute consistent estimators for those parameters. For the structural mean models (4)–(6), η is linear in Y_t. Model (1) can be used with α estimated by $\hat{\alpha}_{\text{LS}}$. For the nested structural AFT model (7) and right-censored outcomes, η is not linear in Y_t. Model (2) can be used with α estimated by $\hat{\alpha}_{\text{BJ}}$ and the distribution of ε estimated by the Kaplan & Meier (1958) estimator based on the residuals.

Under regularity conditions (Robins, 1999), $\hat{\vartheta}_{\text{aug}}$ is consistent and asymptotically normal if the propensity score model (3) is correctly specified. The estimator $\hat{\vartheta}_{\text{aug}}$ is weakly locally efficient in the following sense. If both the propensity score model and the outcome regression model are correctly specified, then it achieves the minimum asymptotic variance in the class of estimators of ϑ that are solutions to estimating equations of the form

$$\tilde{E}\left[\hat{\tau}_{\text{init}}(\vartheta) - \left\{\frac{q(T, X)}{\hat{\pi}(T, X)} - \sum_{t \in \mathcal{T}} q(t, X)\right\}\right] = 0, \tag{8}$$

where $q(t, X)$ is a $d \times 1$ vector of functions but $\eta(t, Y_t, X; \vartheta)$ is fixed. This property is weaker than that of local efficiency, which says that if the two models are correctly specified, then an estimator achieves the semiparametric variance bound or, equivalently here, the minimum asymptotic variance based on estimating equation (8) over all possible $\eta(t, Y_t, X; \vartheta)$ and $q(t, X)$. Moreover, $\hat{\vartheta}_{\text{aug}}$ is doubly robust, that is, it remains consistent if either the propensity score model or the outcome regression model is correctly specified.

For each of the structural models considered in Section 2.1, there exists a class of unbiased full-data estimating functions in the form $\tau_{\text{full}}(\vartheta)$ (e.g., van der Laan & Robins, 2003). The following examples will be used in Sections 4 and 5.

(i) For the marginal structural mean model (4),

$$\eta(t, Y_t, X; \theta) = \phi(t, V; \theta)(Y_t - \nu(t, V; \theta)),$$

where $\phi(t, V; \theta)$ is a $\dim(\theta) \times 1$ vector of functions. The initial and augmented estimators of θ are called IPW and AIPW estimators, respectively. Let $\varpi(t, V) = P(T = t|V)$. A computationally convenient, but not fully efficient, choice of ϕ is $\phi_{\text{conv}}(t, V; \theta) = \hat{\varpi}(t, V)\partial\nu(t, V; \theta)/\partial\theta$, where $\hat{\varpi}(t, V)$ is an estimator of $\varpi(t, V)$ based on a possibly misspecified model for $\varpi(t, V)$. The asymptotic distribution of $\hat{\vartheta}_{\text{init}}$ or $\hat{\vartheta}_{\text{aug}}$ is not affected by the fact that $\hat{\varpi}(t, V)$ is estimated (Robins, 1999, Section 4.3).

(ii) For the additive nested mean structural model (5),

$$\eta(t, Y_t, X; \theta) = \varphi(t, V; \theta_1)(Y_t - c_1(t, V; \theta_1)),$$

where $\varphi(t, V; \theta_1)$ is a $\dim(\theta_1) \times 1$ vector of functions such that $\sum_{t \in \mathcal{T}} \varphi(t, V; \theta_1) = 0$. Similarly, for the multiplicative nested mean structural model (6), $\eta(t, Y_t, X; \theta) = \varphi(t, V; \theta_1)(Y_t / \exp[c_1(t, V; \theta_1)])$. The initial and augmented estimators of θ_1 are called G and AG estimators, respectively. For model (5), a computationally convenient, but not fully efficient, choice of φ is $\varphi_{\text{conv}}(t, V; \theta_1) = \hat{\varpi}(t, V)[\partial c_1(t, V; \theta_1)/\partial\theta_1 - \sum_{j \in \mathcal{T}} \hat{\varpi}(j, V)\partial c_1(j, V; \theta_1)/\partial\theta_1]$.

(iii) For the nested structural AFT model (7),

$$\eta(t, Y_t, X; \theta_1) = \varphi(t, V; \theta_1)\kappa(Y_t - c_1(t, V; \theta_1), V; \theta_1)$$

where $\kappa(Y_0, V; \theta_1)$ is a scalar function, and $\varphi(t, V; \theta_1)$ is a $\dim(\theta_1) \times 1$ vector of functions such that $\sum_{t \in \mathcal{T}} \varphi(t, V; \theta_1) = 0$. Suppose that Y is censored by a known constant C as in the case where survival time is censored at the administrative end of follow-up. Then κ should be chosen such that $\kappa(Y_t - c_1(t, V; \theta_1), V; \theta_1)$ depends on Y_t only through $Y_t \wedge C$. Two examples of such functions are

$$\kappa(Y_0, V; \theta_1) = Y_0 \wedge (C - \max_{t\in\mathcal{T}} c_1(t, V; \theta_1)), \tag{9}$$

$$\text{or} \quad 1\{Y_0 \leq C - \max_{t\in\mathcal{T}} c_1(t, V; \theta_1)\}. \tag{10}$$

An uncensored value of $Y_t - c_1(t, V; \theta_1)$ may become censored after the transformation by κ. This technique is called artificial censoring by Robins (1989) and others.

3. NONPARAMETRIC LIKELIHOOD AND ESTIMATING EQUATIONS

In the absence of a structural model and in the case of binary treatments, Tan (2006) developed a nonparametric likelihood method for estimating the distributions of (Y_t, X), and derived a regression estimator of $E(Y_t)$ that is locally efficient, doubly robust, and asymptotically at least as efficient as all previously proposed estimators under a correctly specified propensity score model. We extend the likelihood and regression estimators to structural models with discrete treatments (binary or nonbinary).

Let γ^* be the value of γ minimizing the Kullback–Leibler distance between $\pi(\cdot;\gamma)$ and the true propensity score π. Model (3) is correct if and only if $\pi^* = \pi(\cdot;\gamma^*)$ and π are identical. Let $s(\cdot;\gamma) = [\partial\pi(T, X;\gamma)/\partial\gamma]/\pi(T, X;\gamma)$ and $s^* = s(\cdot;\gamma^*)$. Let $\Gamma = -E(\partial s/\partial\gamma)|_{\gamma=\gamma^*}$. Under regularity conditions including the nonsingularity of Γ (White, 1982), the MLE $\hat{\gamma}$ converges to γ^* in probability and has the expansion

$$\hat{\gamma} - \gamma^* \simeq \Gamma^{-1}\tilde{E}(s^*).$$

Throughout, $\simeq$ denotes equality up to order $o_p(n^{-1/2})$.

3.1. Profile Likelihood and the Likelihood Estimator

Throughout Sections 3.1 and 3.2, assume that the propensity score model (3) is correctly specified. The nonparametric likelihood (Keifer & Wolfowitz, 1956) is

$$L_1 \times L_2 = \prod_{i=1}^{n} \pi(T_i, X_i;\gamma) \times \prod_{i=1}^{n} G_{T_i}(\{Y_i, X_i\}),$$

where G_t is the joint distribution of (Y_t, X), $t \in \mathcal{T}$, and hence a probability measure. The likelihood is a product of two factors, L_1 involving γ only and L_2 involving G_t, $t \in \mathcal{T}$, only. Each distribution G_t can be factorized as the marginal distribution $P(X)$ times the conditional distribution $P(Y_t|X)$, but such a factorization is avoided here. This likelihood formulation allows us to clarify all available constraints under a structural model, which deals with $P(Y_t|V)$ rather than $P(Y_t|X)$.

The distributions G_t, $t \in \mathcal{T}$, satisfy two different types of constraints. First, by definition, the distributions G_t, $t \in \mathcal{T}$, induce the same marginal distribution of X and hence satisfy the "inherent constraint"

$$\sum_{t\in\mathcal{T}} \int h(t, x)\, \mathrm{d}G_t(y_t, x) = 0 \tag{11}$$

for each bounded function $h(t, X)$ such that $\sum_{t\in\mathcal{T}} h(t, X) = 0$. Such constraints always hold regardless of any structural model. Second, under a structural model, the distributions G_t, $t \in \mathcal{T}$, are further constrained by full-data estimating equations as discussed in Section 2.2 and satisfy the "modelling constraint"

$$\sum_{t\in\mathcal{T}} \int \eta(t, y_t, x; \vartheta)\, \mathrm{d}G_t(y_t, x) = 0 \tag{12}$$

for each function $\eta(t, Y_t, X; \vartheta)$ such that $\sum_{t\in\mathcal{T}} E[\eta(t, Y_t, X; \vartheta)] = 0$. Such constraints arise from modelling assumptions in a specific structural model.

The key to our likelihood method is that we choose to retain a finite subset of all constraints (11) and (12) and ignore the remaining constraints. On one hand, it is standard to retain only finitely many modelling constraints (12), as in the estimation of conditional mean models with fully observed data by the approaches of estimating functions (McCullagh & Nelder, 1989) and empirical likelihood (Owen, 2001). On the other hand, it is unconventional to retain only finitely many inherent constraints (11). Effectively, we obtain an expanded model and a *new likelihood* (see Section 3.4). The idea of ignoring inherent constraints is related to the formulation of Kong et al. (2003) for Monte Carlo integration. The key to that formulation is to ignore part of all available information about the baseline measure.

Let $\eta(t, Y_t, X; \vartheta)$ be a $p \times 1$, $p \geq d$, vector of functions, to be used in constraints (12), such that $\sum_{t\in\mathcal{T}} \eta(t, Y_t, X; \vartheta)$ is a vector of full-data unbiased estimating functions. We refer to the case where $p = d$ as just-determined and where $p > d$ as over-determined (e.g., Owen, 2001). Let $h = (h^{(0)\top}, h^{(1)\top}, h^{(2)\top})^\top$ where

$h^{(0)}(t, X; \gamma) = [1\{t = j\} - \pi(t, X; \gamma)]^\top_{j\in\mathcal{T}}$ satisfies $\sum_{t\in\mathcal{T}} h^{(0)}(t, X; \gamma) \equiv 0$,
$h^{(1)}(t, X; \gamma)$ is a vector of functions such that $\sum_{t\in\mathcal{T}} h^{(1)}(t, X; \gamma) \equiv 0$,
$h^{(2)}(t, X; \gamma) = \partial\pi(t, X; \gamma)/\partial\gamma$ satisfies $\sum_{t\in\mathcal{T}} h^{(2)}(t, X; \gamma) \equiv 0$.

The vector $h^{(0)}(t, X; \gamma)$ consists of J functions $1\{t = 0\} - \pi(t, X; \gamma), \ldots,$ and $1\{t = J - 1\} - \pi(t, X; \gamma)$. Two choices of $h^{(1)}(t, X; \gamma)$ are suggested in Section 3.3 by taking $\hbar^{(1)}(t, X; \gamma)$ as (17) and (18) in Equation (14). The functions in $\hat{h}(t, X) = h(t, X; \hat{\gamma})$ are to be used as $h(t, X)$ in constraints (11). We use a hat above a function of γ and other variables to indicate that the function is evaluated at $\hat{\gamma}$, and use a star to indicate that the function is evaluated at γ^*. For example, $\partial\hat{\pi}/\partial\gamma = \partial\pi/\partial\gamma|_{\gamma=\hat{\gamma}}$ and $\partial\pi^*/\partial\gamma = \partial\pi/\partial\gamma|_{\gamma=\gamma^*}$.

Our likelihood method consists of two steps. First, maximize L_1, that is, fit the propensity score model by maximum likelihood. Second, compute the likelihood estimator $\hat{\vartheta}_{\text{lik}}$ as a maximizer of $\text{pl}_n(\vartheta)$, and the profile likelihood as

$$\text{pl}_n(\vartheta) = \max_{G_t : t\in\mathcal{T}} L_2$$

subject to the constraints that $\int \mathrm{d}G_t(y_t, x) = 1$, $t \in \mathcal{T}$, and

$$\text{(C0)} \qquad \sum_{t\in\mathcal{T}} \int \hat{\pi}(t, x)\, \mathrm{d}G_t(y_t, x) = 1,$$

$$\text{(C1)} \qquad \sum_{t\in\mathcal{T}} \int \hat{h}^{(1)}(t, x)\, \mathrm{d}G_t(y_t, x) = 0,$$

$$\text{(C2)} \qquad \sum_{t\in\mathcal{T}} \int \frac{\partial\hat{\pi}}{\partial\gamma}(t, x)\, \mathrm{d}G_t(y_t, x) = 0,$$

$$\text{(C3)} \qquad \sum_{t\in\mathcal{T}} \int \eta(t, y_t, x; \vartheta)\, \mathrm{d}G_t(y_t, x) = 0.$$

See Section 3.4 for a discussion on roles of inherent constraints C0–C2. Constraint C0 is equivalent to $\sum_{t\in\mathcal{T}} \int \hat{h}^{(0)}(t, x)\, \mathrm{d}G_t(y_t, x) = 0$ given $\int \mathrm{d}G_t(y_t, x) = 1$. In addition, we require in the maximization of L_2 that G_t be a probability measure supported on $\{(Y_i, X_i) : T_i = t\}$, $t \in \mathcal{T}$. Theorem

1 provides a formula for computing $\mathrm{pl}_n(\vartheta)$, and shows that maximizing L_2 with high-dimensional unknowns $(G_t)_{t\in\mathcal{T}}$ can be solved through maximizing a certain function ℓ_n with low-dimensional unknowns (λ, ϱ).

Theorem 1. *Assume that the row vectors in the matrix* $[\hat h(T_1, X_1), \ldots, \hat h(T_n, X_n)]$ *and in the matrix* $[\eta(T_1, Y_1, X_1; \vartheta), \ldots, \eta(T_n, Y_n, X_n; \vartheta)]$ *are linearly independent. Consider the following function*

$$\ell_n(\lambda, \varrho; \vartheta) = \tilde E[\log(\hat\pi(T, X) + \lambda^\top \hat h(T, X) + \varrho^\top \eta(T, Y, X; \vartheta))],$$

where log of 0 *or a negative number is* $-\infty$. *For fixed* ϑ, *if* ℓ_n *achieves a maximum at* $\hat\lambda = \hat\lambda(\vartheta)$ *and* $\hat\varrho = \hat\varrho(\vartheta)$, *then* L_2 *achieves the constrained maximum at*

$$\hat G_t(\{Y_i, X_i\}; \vartheta) = \frac{n^{-1}}{\hat\pi(t, X_i) + \hat\lambda^\top \hat h(t, X_i) + \hat\varrho^\top \eta(t, Y_i, X_i; \vartheta)} \quad \text{if } T_i = t.$$

In the just-determined case $p = d$, $\hat\vartheta_{\mathrm{lik}}$ can be computed by solving

$$\sum_{t\in\mathcal{T}} \int \eta(t, y_t, x; \vartheta)\, \mathrm{d}\hat G_t(y_t, x) = 0,$$

where $\hat G_t$, $t \in \mathcal{T}$, are jointly a maximizer of L_2 subject to inherent constraints C0–C2, without modelling constraint C3, as in Tan (2006). Therefore, our likelihood method recovers the full-data distributions G_t, $t \in \mathcal{T}$, from observed data by using the propensity score model alone, and then derives estimating equations from the full-data estimating functions $\sum_{t\in\mathcal{T}} \eta(t, Y_t, X; \vartheta)$ and the estimated distributions $\hat G_t$, $t \in \mathcal{T}$.

The following theorem and corollary summarize the asymptotic properties of the profile likelihood and the likelihood estimator. The results extend similar results on the empirical likelihood estimation of conditional mean models in the absence of missing data (Owen, 2001; Qin & Lawless, 1994), and agree with a general theory of profile likelihood for semiparametric models (Murphy & van der Vaart, 2000).

Theorem 2. *Let*

$$\tau_{\mathrm{init}}(\vartheta) = \frac{\eta(T, Y, X; \vartheta)}{\pi(T, X; \gamma)}, \quad \xi = \frac{h(T, X; \gamma)}{\pi(T, X; \gamma)}.$$

Assume that the regularity conditions in the Supplementary Appendix hold. If the propensity score model (3) *is correctly specified, then the following results hold.*

(i) $\hat\vartheta_{\mathrm{lik}}$ *converges to the true value* ϑ_0 *in probability and has the expansion*

$$\hat\vartheta_{\mathrm{lik}} - \vartheta_0 \simeq -\mathcal{V}^{-1} E^\top\left(\frac{\partial\tau^*_{\mathrm{init}}(\vartheta_0)}{\partial\vartheta}\right) \mathrm{var}^{-1}[\tau^*_{\mathrm{reg}}(\vartheta_0)]\tilde E[\tau^*_{\mathrm{reg}}(\vartheta_0)],$$

where $\tau^*_{\mathrm{reg}}(\vartheta) = \tau^*_{\mathrm{init}}(\vartheta) - \beta^\top(\vartheta)\xi^*$ *with* $\beta(\vartheta) = E^{-1}(\xi^*\xi^{*\top})E[\xi^*\tau^{*\top}_{\mathrm{init}}(\vartheta)]$, *and*

$$\mathcal{V} = E^\top\left(\frac{\partial\tau^*_{\mathrm{init}}(\vartheta_0)}{\partial\vartheta}\right) \mathrm{var}^{-1}[\tau^*_{\mathrm{reg}}(\vartheta_0)] E\left(\frac{\partial\tau^*_{\mathrm{init}}(\vartheta_0)}{\partial\vartheta}\right);$$

(ii) $-n^{-1}\partial^2 \log pl_n(\vartheta)/\partial\vartheta^2|_{\vartheta=\hat\vartheta_{\mathrm{lik}}}$ *is a consistent estimator of* $\mathcal{V}$;

(iii) $-2\log[pl_n(\vartheta_0)/pl_n(\hat\vartheta_{\mathrm{lik}})]$ *is asymptotically chi-squared with d degrees of freedom.*

Corollary 1. *The likelihood estimator $\hat{\vartheta}_{\text{lik}}$ is asymptotically equivalent, up to the first order, to the optimal estimator in the class of estimators of ϑ that are solutions to d estimating equations of the form*

$$\begin{aligned} 0 &= a^\top \tilde{E}\left[\hat{\tau}_{\text{init}}(\vartheta) - b^\top \hat{\xi}\right] \qquad (13)\\ &= a^\top \tilde{E}\left[\hat{\tau}_{\text{init}}(\vartheta) - (b^{(0)\top}, b^{(1)\top})(\hat{\xi}^{(0)\top}, \hat{\xi}^{(1)\top})^\top\right], \end{aligned}$$

where a is a $p \times d$ matrix and b is a $\dim(h) \times p$ matrix. Recall that $\hat{\tau}_{\text{init}}(\vartheta) = \hat{\eta}/\hat{\pi}$ and $\hat{\xi} = \hat{h}/\hat{\pi}$ are defined by evaluating $\tau_{\text{init}}(\vartheta)$ and ξ at $\hat{\gamma}$. The vector $\hat{\xi} = (\hat{\xi}^{(0)\top}, \hat{\xi}^{(1)\top}, \hat{\xi}^{(2)\top})^\top$ and the matrix $b = (b^{(0)\top}, b^{(1)\top}, b^{(2)\top})^\top$ are partitioned according to $h = (h^{(0)\top}, h^{(1)\top}, h^{(2)\top})^\top$. Simplification of (13) *holds because $\hat{\xi}^{(2)} = \hat{s} = \hat{h}^{(2)}/\hat{\pi}$ and hence $\tilde{E}(\hat{\xi}^{(2)}) = \tilde{E}(\hat{s}) = 0$.*

By Corollary 1, $\hat{\vartheta}_{\text{lik}}$ is asymptotically equivalent to the optimal estimator based on estimating functions $\hat{\tau}_{\text{init}}(\vartheta)$ and $\hat{\xi}$, which are mapped by inverse weighting from, respectively, the retained constraints (11) and (12) associated with $\hat{\eta}$ and $\hat{h}$. Therefore, $\hat{\vartheta}_{\text{lik}}$ asymptotically attains the best possible efficiency by using and only using the retained constraints. We call $\hat{\vartheta}_{\text{lik}}$ intrinsically efficient for fixed $\hat{\eta}$ and $\hat{h}$ or simply intrinsically efficient.

An optimal member in the class of estimating equations (13) can be identified with $b = \beta$ and $a = \text{var}^{-1}(\tau^*_{\text{reg}})E[\partial\tau^*_{\text{init}}/\partial\vartheta]$. There are two steps in which the optimality is realized. First, $\hat{\tau}_{\text{init}}(\vartheta)$ are optimally augmented by using $\hat{\xi}$. In fact, $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta) - \beta^\top\hat{\xi}]$ is asymptotically equivalent to $\tilde{E}[\tau^*_{\text{init}}(\vartheta) - \beta^\top\xi^*]$ and achieves the minimum asymptotic variance among $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta) - b^\top\hat{\xi}]$ for arbitrary b. See the proof of Theorem 2(i). Particularly, $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta) - \beta^\top\hat{\xi}]$ always has asymptotic variance no greater than $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta)]$. We refer to $\hat{\xi}$ as control variates in that they have mean 0 asymptotically and are used for variance reduction, as in the method of control variates for Monte Carlo integration (Hammersley & Handscomb, 1964). Second, $\hat{\tau}_{\text{init}}(\vartheta) - \beta^\top\hat{\xi}$ are optimally combined into d estimating functions. In fact, $(n\mathcal{V})^{-1}$ is precisely the minimum asymptotic variance for the class of estimators that are solutions to (13) with $b = \beta$ (Hansen, 1982; McCullagh & Nelder, 1989). The second step is relevant only in the over-determined case $p > d$.

Qin & Lawless (1994) developed empirical likelihood estimation under the assumption that *observed-data* estimating functions are available. In contrast, our likelihood method assumes that *full-data* estimating functions are available. Nevertheless, there is an interesting derivation of $\hat{\vartheta}_{\text{lik}}$ by using empirical likelihood. A prerequisite for this derivation is to *recognize* $\hat{\tau}_{\text{init}}(\vartheta)$ and $\hat{\xi}$ as asymptotically unbiased estimating functions based on the observed data (T, Y, X), although our nonparametric likelihood derivation relies on characterizing constraints (11) and (12) based on the full data $[(Y_t)_{t\in\mathcal{T}}, X]$. Conceptually, the relationship between the two derivations is parallel to that between observed-data estimating functions $\hat{\tau}_{\text{init}}(\vartheta)$ and $\hat{\xi}$ and full-data constraints (11) and (12) associated with $\hat{\eta}$ and $\hat{h}$. To our knowledge, the estimator $\hat{\vartheta}_{\text{lik}}$ and both of the derivations are new. See Qin & Zhang (2007) for a different application of empirical likelihood for estimation of $E(Y_t)$ and potentially also for estimation of ϑ in the present setting. Their estimator of $E(Y_t)$ is locally efficient and doubly robust, but is not intrinsically efficient even if $\pi(t, x)$ is known and substituted for $\hat{\pi}(t, x)$. The resulting profile likelihood does not possess asymptotic properties such as Theorem 2(ii)–(iii).

The empirical likelihood of (T_i, Y_i, X_i), $i = 1, \ldots, n$, is

$$L = \prod_{i=1}^{n} P(\{T_i, Y_i, X_i\}),$$

where P is the joint distribution of (T, Y, X). For fixed ϑ, the profile empirical likelihood is $\max_P L$, where P is a discrete distribution supported on $\{(T_i, Y_i, X_i) : i = 1, \ldots, n\}$ and satisfies the constraints $\int \hat{\xi}(t, x)\,\mathrm{d}P(t, y, x) = 0$ and $\int \hat{\tau}_{\text{init}}(t, y_t, x; \vartheta)\,\mathrm{d}P(t, y, x) = 0$. Given $\int \mathrm{d}P(t, y, x) = 1$, these constraints can be written in a form similar to C0–C3:

$$\int \frac{1\{t = j\}}{\hat{\pi}(t, x)}\,\mathrm{d}P(t, y, x) = 1, \quad j \in \mathcal{T},$$

$$\int \frac{\hat{h}^{(1)}(t, x)}{\hat{\pi}(t, x)}\,\mathrm{d}P(t, y, x) = 0,$$

$$\int \frac{\frac{\partial \hat{\pi}}{\partial \gamma}(t, x)}{\hat{\pi}(t, x)}\,\mathrm{d}P(t, y, x) = 0,$$

$$\int \frac{\eta(t, y_t, x; \vartheta)}{\hat{\pi}(t, x)}\,\mathrm{d}P(t, y, x) = 0.$$

By standard calculations, L is maximized subject to the above constraints at

$$\hat{P}(\{T_i, Y_i, X_i\}; \vartheta) = \frac{n^{-1}}{1 + \hat{\lambda}^\top \frac{\hat{h}(T_i, X_i)}{\hat{\pi}(T_i, X_i)} + \hat{\varrho}^\top \frac{\eta(T_i, Y_i, X_i; \vartheta)}{\hat{\pi}(T_i, X_i)}}.$$

where $(\hat{\lambda}, \hat{\varrho})$ is a maximizer of ℓ_n as described in Theorem 1. Therefore, the profile empirical likelihood function of ϑ is proportional to $\mathrm{pl}_n(\vartheta)$, and the empirical likelihood estimator is identical to $\hat{\vartheta}_{\text{lik}}$ as a maximizer of $\mathrm{pl}_n(\vartheta)$. Theorem 2 and Corollary 1 can also be shown by applying asymptotic theory for empirical likelihood in Qin & Lawless (1994), although additional work is required to accommodate the variation of $\hat{\gamma}$.

The generalized method of moments (GMM) (Hansen, 1982) provides an alternative method for estimation of ϑ based on the estimating functions $\hat{\tau}_{\text{init}}(\vartheta)$ and $\hat{\xi}$. A GMM estimator can be obtained as the optimal estimator in the class of estimators that are solutions to (13), with the optimal a and b consistently estimated. The estimator is asymptotically equivalent to the optimal estimator and hence to $\hat{\vartheta}_{\text{lik}}$, up to the first order. The derivation of $\hat{\vartheta}_{\text{lik}}$ as an empirical likelihood estimator suggests an interesting, open question whether $\hat{\vartheta}_{\text{lik}}$ has better higher order asymptotic properties than GMM estimators, even though $\hat{\tau}_{\text{init}}(\vartheta)$ and $\hat{\xi}$ involve the estimator $\hat{\gamma}$. See Newey & Smith (2004) for higher order asymptotic results about empirical likelihood and GMM estimators in the case where estimating functions are free of estimated nuisance parameters. However, this comparison depends on the assumption that the propensity score model is correctly specified. If the propensity score model is misspecified, then $\hat{\vartheta}_{\text{lik}}$ is generally inconsistent. In Section 3.2, we derive a doubly robust GMM estimator, which remains consistent if either the propensity score model or the outcome regression model is correctly specified.

3.2. Regression Estimators

The vector of functions $h(t, X; \gamma)$ such that $\sum_{t \in \mathcal{T}} h(t, X; \gamma) = 0$ used in our likelihood method can be written as

$$h(t, X; \gamma) = \hbar(t, X; \gamma) - \pi(t, X; \gamma) \sum_{j \in \mathcal{T}} \hbar(j, X; \gamma), \tag{14}$$

where $\hbar(t, X; \gamma)$ is a vector of functions. Partition $\hbar = (\hbar^{(0)\top}, \hbar^{(1)\top}, \hbar^{(2)\top})^\top$ and similar vectors in the same way as $h = (h^{(0)\top}, h^{(1)\top}, h^{(2)\top})^\top$. We take $\hbar^{(0)}(t, X; \gamma) = [1\{t = j\}]_{j \in \mathcal{T}}^\top$ and

$\hbar^{(2)}(t, X; \gamma) = \partial\pi(t, X; \gamma)/\partial\gamma$, and leave $\hbar^{(1)}(t, X; \gamma)$ to be specified [see (17) and (18) in Section 3.3]. Let

$$\zeta(T, X; \gamma) = \frac{\hbar(T, X; \gamma)}{\pi(T, X; \gamma)},$$

Then ξ defined in Theorem 2 is related to $\hbar$ or ζ by

$$\begin{aligned}\xi(T, X; \gamma) &= \frac{\hbar(T, X; \gamma)}{\pi(T, X; \gamma)} - \sum_{t \in \mathcal{T}} \hbar(t, X; \gamma) \\ &= \zeta(T, X; \gamma) - E[\zeta(T, X; \gamma)|X] \quad \text{under model (3).}\end{aligned}$$

Therefore, ξ is the projection of ζ on the space of square-integrable functions $\varrho(T, X)$ such that $E[\varrho(T, X)|X] = 0$. This space is the tangent space in the nonparametric model for the propensity score (e.g., van der Laan & Robins, 2003).

In the just-determined case $p = d$, define the regression estimators $\hat{\vartheta}_{\text{reg}}$ and $\tilde{\vartheta}_{\text{reg}}$ as solutions to $\tilde{E}[\hat{\tau}_{\text{reg}}(\vartheta)] = 0$ and $\tilde{E}[\tilde{\tau}_{\text{reg}}(\vartheta)] = 0$, respectively, where

$$\begin{aligned}\hat{\tau}_{\text{reg}}(\vartheta) &= \hat{\tau}_{\text{init}}(\vartheta) - \hat{\beta}^{\top}(\vartheta)\hat{\xi} \quad \text{with} \quad \hat{\beta}(\vartheta) = \tilde{E}^{-1}(\hat{\xi}\hat{\xi}^{\top})\tilde{E}[\hat{\xi}\hat{\tau}_{\text{init}}^{\top}(\vartheta)], \\ \tilde{\tau}_{\text{reg}}(\vartheta) &= \hat{\tau}_{\text{init}}(\vartheta) - \tilde{\beta}^{\top}(\vartheta)\hat{\xi} \quad \text{with} \quad \tilde{\beta}(\vartheta) = \tilde{E}^{-1}(\hat{\xi}\hat{\zeta}^{\top})\tilde{E}[\hat{\xi}\hat{\tau}_{\text{init}}^{\top}(\vartheta)].\end{aligned}$$

The two estimators can be seen as GMM estimators, with a fixed at a nonsingular constant matrix and b replaced by $\hat{\beta}$ and $\tilde{\beta}$ in (13). It is important to notice the difference between $\hat{\beta}$ and $\tilde{\beta}$. The coefficient $\hat{\beta}$ is the least-squares estimator in the linear regression of $\hat{\tau}_{\text{init}}(\vartheta)$ on $\hat{\xi}$. See Robins, Rotnitzky & Zhao (1995) for the use of $\hat{\beta}$ in related missing data problems. In contrast, $\tilde{\beta}$ is constructed by substituting $\tilde{E}(\hat{\xi}\hat{\zeta}^{\top})$ for $\tilde{E}(\hat{\xi}\hat{\xi}^{\top})$ in $\hat{\beta}$, as in the regression estimator of $E(Y_t)$ in Tan (2006). See Goetgeluk, Vansteelandt & Goetghebeur (2009) for a related application of this idea to structural nested direct effects models. If the propensity score model is correctly specified, then $\hat{\beta}$ and $\tilde{\beta}$ have the same limit β in probability, and hence $\hat{\vartheta}_{\text{reg}}$ and $\tilde{\vartheta}_{\text{reg}}$ are consistent and asymptotically equivalent up to the first order (see Theorem 3). On the other hand, if the propensity score model is misspecified, then $\hat{\beta}$ and $\tilde{\beta}$ generally have different limits in probability, and hence $\hat{\vartheta}_{\text{reg}}$ and $\tilde{\vartheta}_{\text{reg}}$ have different asymptotic properties. The estimator $\hat{\vartheta}_{\text{reg}}$ is not doubly robust, whereas $\tilde{\vartheta}_{\text{reg}}$ is doubly robust due to the construction of $\tilde{\beta}$ (see Theorem 4).

The following theorem shows that both regression estimators are first-order approximations to the likelihood estimator under a correctly specified propensity score model. This result is an extension of the results for the estimators of $E(Y_t)$ in Tan (2006), and is similar to those established for Monte Carlo integration in Tan (2004).

Theorem 3. *Suppose that the regularity conditions hold as in Theorem* 2. *If $p = d$ and the propensity score model* (3) *is correctly specified, then*

$$\begin{aligned}\hat{\vartheta}_{lik} - \vartheta_0 \simeq \hat{\vartheta}_{\text{reg}} - \vartheta_0 &\simeq \tilde{\vartheta}_{\text{reg}} - \vartheta_0 \\ &\simeq -E^{-1}\left(\frac{\partial\tau_{\text{init}}^{*}(\vartheta_0)}{\partial\vartheta}\right)\tilde{E}[\tau_{\text{reg}}^{*}(\vartheta_0)].\end{aligned}$$

The asymptotic expansion for $\hat{\vartheta}_{\text{lik}}$ agrees with that in Theorem 2.

In the over-determined case $p > d$, $\hat{\vartheta}_{\text{reg}}$ and $\tilde{\vartheta}_{\text{reg}}$ can be extended as GMM estimators that solutions to (13), with the optimal a consistently estimated in addition to the fact that b is replaced by $\hat{\beta}$ and $\tilde{\beta}$. The resulting estimating functions on the right-hand side of (13) are optimal linear combinations of $\hat{\tau}_{\mathrm{reg}}(\vartheta)$ and $\tilde{\tau}_{\mathrm{reg}}(\vartheta)$. It remains future work to fully investigate such GMM estimators in this general case.

3.3. Local Efficiency and Double Robustness

Throughout Sections 3.3–3.7, assume that $p = d$. In this section, we discuss choices of $\hbar(t, X; \gamma)$, depending on the outcome regression model (1) or (2), such that $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ are weakly locally efficient, and $\tilde{\vartheta}_{\mathrm{reg}}$ is doubly robust.

Suppose that the propensity score model (3) is correctly specified. The asymptotic efficiency of $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ depends on the choices of $\hbar(t, X; \gamma)$ and $\eta(t, Y_t, X; \vartheta)$. The semiparametric theory of Robins (1999) can be used to find the optimal choices of these functions that lead to semiparametric efficiency. In general, the optimal choice of $\eta(t, Y_t, X; \vartheta)$ involves complicated integral equations. Therefore, the computationally convenient choices ϕ_{conv} and φ_{conv} are recommended for practical use (see Section 2.1).

For fixed $\eta(t, Y_t, X; \vartheta)$, consider the following condition for $\hbar(t, X; \gamma)$:

$$E[\eta(T, Y, X; \vartheta) | T = t, X] = c^{\top} \hbar^{*}(t, X) \text{ almost surely} \tag{15}$$

for some $\dim(\hbar) \times d$ matrix c. If $\hbar(t, X; \gamma)$ satisfies condition (15), then β equals

$$\begin{aligned} \beta &= E^{-1}(\xi^{*} \zeta^{*\top}) E(\xi^{*} \tau_{\mathrm{init}}^{*\top}) \\ &= E^{-1}\left[\xi^{*} \frac{\hbar^{*\top}(T, X)}{\pi^{*}(T, X)}\right] E\left[\xi^{*} \frac{E^{\top}(\eta | T, X)}{\pi^{*}(T, X)}\right] = c, \end{aligned} \tag{16}$$

and therefore $\tau^{*}_{\mathrm{reg}}(\vartheta)$ equals

$$\tau^{*}_{\mathrm{aug}}(\vartheta) = \tau^{*}_{\mathrm{init}}(\vartheta) - \left\{ \frac{E(\eta | T, X)}{\pi^{*}(T, X)} - \sum_{t \in \mathcal{T}} E(\eta | T = t, X) \right\}.$$

By Theorem 3, the influence function of $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ is $-E^{-1}[\partial \tau^{*}_{\mathrm{init}} / \partial \vartheta] \tau^{*}_{\mathrm{reg}}(\vartheta)$. By Robins (1999), $-E^{-1}[\partial \tau^{*}_{\mathrm{init}} / \partial \vartheta] \tau^{*}_{\mathrm{aug}}(\vartheta)$ is the efficient influence function for estimators of ϑ that are solutions to estimating Equations (8). Therefore, if condition (15) holds, then $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ achieve the minimum asymptotic variance over all possible $\hbar(t, X; \gamma)$ and moreover in the class of estimators that are solutions to (8). The estimators $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ are weakly locally efficient as is $\hat{\vartheta}_{\mathrm{aug}}$ discussed in Section 2.2.

The outcome regression model (1) or (2) can be used as a guidance to choose $\hbar(t, X; \gamma)$. Consider the following choices (free of γ)

$$\hbar^{(1)}(t, X) = \hat{E}(\eta | T = t, X) \tag{17}$$

$$\text{or } = [\, 1\{t = j\} \hat{E}^{\top}(\eta | T = j, X)]^{\top}_{j \in \mathcal{T}}, \tag{18}$$

which are column vectors of dimension d and dJ, respectively. If the outcome regression model is correctly specified, then condition (15) is satisfied asymptotically. Therefore, with choice (17) or (18), $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$ are weakly locally efficient.

Now suppose that the propensity score model (3) can be wrong. Theorem 4 shows the asymptotic behaviour of $\tilde{\vartheta}_{\mathrm{reg}}$ without assuming that model (3) is correctly specified.

Theorem 4. *Redefine*

$$\tau^{*}_{\mathrm{reg}}(\vartheta) = \tau^{*}_{\mathrm{init}}(\vartheta) - \beta(\vartheta)^{\top} \xi^{*} \quad \textit{with} \quad \beta(\vartheta) = E^{-1}(\xi^{*} \zeta^{*\top}) E[\xi^{*} \tau^{*\top}_{\mathrm{init}}(\vartheta)].$$

Assume that there exists a unique value ϑ^* such that $E[\tau^*_{\mathrm{reg}}(\vartheta^*)] = 0$, and additional regularity conditions in the Supplementary Appendix hold. Then $\tilde{\vartheta}_{\mathrm{reg}}$ converges to ϑ^* in probability and has the expansion

$$\tilde{\vartheta}_{\mathrm{reg}} - \vartheta^* \simeq -E^{-1}\left(\frac{\partial \tau^*_{\mathrm{reg}}(\vartheta^*)}{\partial \vartheta}\right)\Big\{\tilde{E}[\tau^*_{\mathrm{reg}}(\vartheta^*) - (\tau^*_{\mathrm{init}}(\vartheta^*) - \beta^\top(\vartheta^*)\zeta^*)\xi^{*\top}\rho]$$
$$- E\left[\frac{\partial}{\partial \gamma}\left(\tau^*_{\mathrm{reg}}(\vartheta^*) - (\tau^*_{\mathrm{init}}(\vartheta^*) - \beta^\top(\vartheta^*)\zeta^*)\xi^{*\top}\rho\right)\right]\Gamma^{-1}\tilde{E}(s^*)\Big\}.$$

where $\rho = E^{-1}(\zeta^*\xi^{*\top})E(\xi^*)$.

If the propensity score model is correct, then $E[\tau^*_{\mathrm{reg}}(\vartheta_0)] = 0$ and hence $\vartheta^* = \vartheta_0$, that is, $\tilde{\vartheta}_{\mathrm{reg}}$ is consistent as are $\hat{\vartheta}_{\mathrm{lik}}$ and $\hat{\vartheta}_{\mathrm{reg}}$. The asymptotic expansion for $\tilde{\vartheta}_{\mathrm{reg}}$ reduces to that in Theorem 3 for the following reasons. First, $\rho = 0$ and $E[\partial\tau^*_{\mathrm{reg}}/\partial\vartheta] = E[\partial\tau^*_{\mathrm{init}}/\partial\vartheta]$, because $E(\xi^*) = 0$. Second, $E[\partial\tau^*_{\mathrm{reg}}/\partial\gamma] = 0$, because $E[\partial\tau^*_{\mathrm{reg}}(\vartheta)/\partial\gamma] = E[\tau^*_{\mathrm{reg}}(\vartheta)s^{*\top}]$ algebraically and $\tau^*_{\mathrm{reg}}(\vartheta)$ and $\xi^{*(2)} = s^*$ are uncorrelated by the construction of $\tau^*_{\mathrm{reg}}(\vartheta)$.

If the propensity score model is wrong, then generally $E[\tau^*_{\mathrm{reg}}(\vartheta_0)] \neq 0$ and $\tilde{\vartheta}_{\mathrm{reg}}$ is inconsistent as are $\hat{\vartheta}_{\mathrm{lik}}$ and $\hat{\vartheta}_{\mathrm{reg}}$. However, if condition (15) holds, then $E[\tau^*_{\mathrm{reg}}(\vartheta_0)] = 0$ and hence $\vartheta^* = \vartheta_0$ (i.e., $\tilde{\vartheta}_{\mathrm{reg}}$ is consistent) for the following reasons. First, under condition (15), β still gives c due to (16) and hence $\tau^*_{\mathrm{reg}}(\vartheta)$ equals $\tau^*_{\mathrm{aug}}(\vartheta)$. Second, $E[\tau^*_{\mathrm{aug}}(\vartheta_0)]$ remains 0 even if the propensity score model is wrong (e.g., Scharfstein, Rotnitzky & Robins, 1999). For choice (17) or (18), $\tilde{\vartheta}_{\mathrm{reg}}$ is doubly robust, that is, it remains consistent if either the propensity score model or the outcome regression model is correctly specified.

3.4. Constraints and Marginalized Estimators

To retain constraints C0–C3 is tantamount to postulating the following model for the data (T_i, Y_i, X_i). First, T is generated according to $P(T = t) = \int \pi \, \mathrm{d}G_t$, $t \in \mathcal{T}$. Second, (Y, X) given $T = t$ is generated from the biased-sampling distribution (Vardi, 1985)

$$\frac{\pi(t, x)}{\int \pi(t, x')\mathrm{d}G_t(y'_t, x')}\,\mathrm{d}G_t(y_t, x),$$

where G_t, $t \in \mathcal{T}$, are probability distributions that satisfy C0–C3 but may not induce the same marginal distribution of X. This model expands the original model in which G_t, $t \in \mathcal{T}$, are probability distributions that induce the same marginal distribution of X. The nonparametric likelihood, even though still equal to $L_1 \times L_2$, is strictly a function defined for G_t, $t \in \mathcal{T}$, in the expanded parameter space, and therefore differs from the original nonparametric likelihood given at the beginning of Section 3.1.

From this perspective, we discuss different roles of inherent constraints C0–C2 in our methods, under the assumption that the propensity score model (3) is correctly specified. First, the constraint $\sum_{t\in\mathcal{T}}\int \hat{\pi}\,\mathrm{d}G_t = 1$ is needed to ensure that the marginal probabilities of T sum to 1. If this constraint is removed, then $\hat{\vartheta}_{\mathrm{lik}}$ is generally inconsistent whatever other constraints are included. Second, the constraints $\int \mathrm{d}G_t = 1$, or equivalently $\sum_{t\in\mathcal{T}}\int \hat{h}^{(0)}\,\mathrm{d}G_t = 0$ given $\sum_{t\in\mathcal{T}}\int \hat{\pi}\,\mathrm{d}G_t = 1$, are needed to make G_t a proper probability distribution. If these constraints are removed, then the total mass of $\hat{G}_t$ differs from 1 but converges to 1 in probability, and $\hat{\vartheta}_{\mathrm{lik}}$ remains consistent but has increased asymptotic variance. In fact, if all the constraints except $\sum_{t\in\mathcal{T}}\int \hat{\pi}\,\mathrm{d}G_t = 1$ are removed, then $\hat{G}_t$ places mass $n^{-1}/\hat{\pi}(t, X_i)$ at (Y_i, X_i) with $T_i = t$ for $t \in \mathcal{T}$. Third, the constraints $\sum_{t\in\mathcal{T}}\int \hat{h}^{(1)}\,\mathrm{d}G_t = 0$ and $\sum_{t\in\mathcal{T}}\int \hat{h}^{(2)}\,\mathrm{d}G_t = 0$ are included for variance reduction. If these constraints are removed, then $\hat{\vartheta}_{\text{lik}}$ is still consistent but generally has increased asymptotic variance. The control variates $\hat{\xi} = \hat{h}/\hat{\pi}$ are mapped by inverse weighting from exactly the inherent constraints (11) associated with $\hat{h}$. Therefore, $\hat{\xi}^{(0)}$, $\hat{\xi}^{(1)}$, and $\hat{\xi}^{(2)}$ play similar roles in $\hat{\vartheta}_{\text{reg}}$ and $\tilde{\vartheta}_{\text{reg}}$ as the corresponding constraints play in $\hat{\vartheta}_{\text{lik}}$.

The constraints C2 associated with $\hat{h}^{(2)} = \partial\hat{\pi}/\partial\gamma$ and the control variates $\hat{\xi}^{(2)} = \hat{s}$ are included to accommodate the variation of $\hat{\gamma}$ asymptotically. If these constraints and control variates are removed, then $\hat{\vartheta}_{\text{lik}}$, $\hat{\vartheta}_{\text{reg}}$, and $\tilde{\vartheta}_{\text{reg}}$ are still consistent but lack the asymptotic optimality described in Corollary 1. However, there can be a trade-off in finite samples. The variations of $\hat{\beta}$ and $\tilde{\beta}$ are negligible in large samples, but can become substantial in small to medium samples due to the extra regressors $\hat{s}$. To reduce these variations, we consider marginalized estimators by removing (or marginalizing) the constraints C2 associated with $\hat{h}^{(2)}$ and the control variates $\hat{s}$ from the original estimators.

Define the likelihood estimator $\hat{\vartheta}^{(\text{m})}_{\text{lik}}$ as a solution to

$$\sum_{t\in\mathcal{T}} \int \eta(t, y_t, x; \vartheta)\, \mathrm{d}\hat{G}^{(\text{m})}_t(y_t, x) = 0,$$

where $\hat{G}^{(\text{m})}_t$, $t \in \mathcal{T}$, are obtained by maximizing L_2 subject to constraints C0–C1, but not C2. Define the regression estimators $\hat{\vartheta}^{(\text{m})}_{\text{reg}}$ and $\tilde{\vartheta}^{(\text{m})}_{\text{reg}}$ as solutions to $\tilde{E}[\hat{\tau}^{(\text{m})}_{\text{reg}}(\vartheta)] = 0$ and $\tilde{E}[\tilde{\tau}^{(\text{m})}_{\text{reg}}(\vartheta)] = 0$, respectively, where $\hat{\tau}^{(\text{m})}_{\text{reg}}(\vartheta)$ and $\tilde{\tau}^{(\text{m})}_{\text{reg}}(\vartheta)$ are the same as $\hat{\tau}_{\text{reg}}(\vartheta)$ and $\tilde{\tau}_{\text{reg}}(\vartheta)$ with $\hat{\xi}$ and $\hat{\zeta}$ replaced by $\hat{\xi}^{(\text{m})} = (\hat{\xi}^{(0)\top}, \hat{\xi}^{(1)\top})^\top$ and $\hat{\zeta}^{(\text{m})} = (\hat{\zeta}^{(0)\top}, \hat{\zeta}^{(1)\top})^\top$. The asymptotic expansion in Theorem 4 remains valid with ξ^* and ζ^* replaced by $\xi^{*(\text{m})} = (\xi^{*(0)\top}, \xi^{*(1)\top})^\top$ and $\zeta^{*(\text{m})} = (\zeta^{*(0)\top}, \zeta^{*(1)\top})^\top$ throughout. The asymptotic expansions in Theorem 3 need to be modified, beyond the aforementioned replacements, as

$$\begin{aligned}\hat{\vartheta}^{(\text{m})}_{\text{lik}} - \vartheta_0 &\simeq \hat{\vartheta}^{(\text{m})}_{\text{reg}} - \vartheta_0 \simeq \tilde{\vartheta}^{(\text{m})}_{\text{reg}} - \vartheta_0 \\ &\simeq -E^{-1}\left(\frac{\partial\tau^*_{\text{init}}(\vartheta_0)}{\partial\vartheta}\right)\{\tilde{E}[\tau^{*(\text{m})}_{\text{reg}}(\vartheta_0)] - E[\tau^{*(\text{m})}_{\text{reg}}(\vartheta_0)s^{*\top}]\Gamma^{-1}\tilde{E}(s^*)\},\end{aligned}$$

where $\tau^{*(\text{m})}_{\text{reg}}(\vartheta) = \tau^*_{\text{init}}(\vartheta) - \beta^{(\text{m})\top}\xi^{*(\text{m})}$ and $\beta^{(\text{m})} = E^{-1}[\xi^{*(\text{m})}\xi^{*(\text{m})\top}]E[\xi^{*(\text{m})}\tau^{*\top}_{\text{init}}(\vartheta)]$. The marginalized estimators lack the asymptotic optimality in Corollary 1. The term in the curly bracket differs from $\tilde{E}[\tau^*_{\text{init}}(\vartheta_0) - \beta^\top\xi^*]$ and does not achieve the minimum asymptotic variance among $\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta_0) - b^\top\hat{\xi}]$ for arbitrary b. See the proof of Theorem 2(i).

3.5. Computation and Simplified Estimators

Computation of $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ is straightforward by using a Newton–Raphson type algorithm, except for one subtle aspect. The suggested choice (17) or (18) for $\hat{h}^{(1)}(t, X)$ involves the unknown parameters ϑ through $\eta(t, Y_t, X; \vartheta)$, and hence so do $\hat{G}_t$, $\hat{\beta}$, and $\tilde{\beta}$. Then we need to update $\hat{G}_t$, $\hat{\beta}$, and $\tilde{\beta}$ at the current value of ϑ in each Newton–Raphson iteration. For example, to solve $\tilde{E}[\tilde{\tau}_{\text{reg}}(\vartheta)] = 0$ and obtain $\tilde{\vartheta}_{\text{reg}}$, the resulting Newton–Raphson algorithm is to start with an initial estimate $\tilde{\vartheta}^{(1)}$, and compute for $k = 1, 2, \ldots,$

$$\begin{aligned}\tilde{\beta}^{(k)} &= \tilde{E}^{-1}(\hat{\xi}\hat{\zeta}^\top)\tilde{E}[\hat{\xi}\hat{\tau}_{\text{init}}(\vartheta)]|_{\vartheta=\hat{\vartheta}^{(k)}}, \\ \hat{\vartheta}^{(k+1)} &= \hat{\vartheta}^{(k)} - \tilde{E}^{-1}(\partial\hat{\tau}_{\text{init}}/\partial\vartheta)\tilde{E}[\hat{\tau}_{\text{init}}(\vartheta) - \tilde{\beta}^{(k)\top}\hat{\xi}]|_{\vartheta=\vartheta^{(k)}}.\end{aligned}$$

If the initial estimator $\tilde{\vartheta}^{(1)}$ is consistent, then the finitely iterated estimator $\tilde{\vartheta}^{(k+1)}$ for $k \geq 1$ is consistent and asymptotically equivalent to $\tilde{\vartheta}_{\text{reg}}$. Particularly, if $\tilde{\vartheta}^{(1)} = \hat{\vartheta}_{\text{aug}}$, then $\tilde{\vartheta}^{(k+1)}$ is weakly locally efficient and doubly robust. Moreover, $\tilde{\vartheta}^{(k+1)}$ is asymptotically efficient in the class of estimators that are solutions to (13), including $\hat{\vartheta}_{\text{init}}$ and $\hat{\vartheta}_{\text{aug}}$, provided that the propensity score model is correctly specified.

For a structural mean model, we can remove the dependency of $\hat{G}_t$, $\hat{\beta}$, and $\tilde{\beta}$ on ϑ by using the following estimators. For the marginal structural mean model (4) and $\phi(t, V)$ free of θ, redefine $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ as solutions to

$$\sum_{t\in\mathcal{T}} \int \phi(t, v) y_t \,\mathrm{d}\hat{G}_t - \tilde{E}[\phi(t, V)\nu(t, V; \theta)] = 0,$$

$$\tilde{E}\left[\frac{\phi(T, V)Y}{\hat{\pi}(T, X)} - \tilde{\beta}^{\top}\hat{\xi} - \sum_{t\in\mathcal{T}} \phi(t, V)\nu(t, V; \theta)\right] = 0, \tag{19}$$

where $\eta(T, Y, X; \theta) = \phi(T, V)(Y - \nu(T, V; \theta))$ is replaced by $\eta^{\dagger}(T, Y, X) = \phi(T, V)Y$, free of θ. For the nested structural mean model (5) and $\varphi(t, V)$ free of θ_1, redefine $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ as solutions to

$$\sum_{t\in\mathcal{T}} \int \varphi(t, v) y_t \,\mathrm{d}\hat{G}_t - \tilde{E}[\varphi(t, V)c_1(t, V; \theta_1)] = 0,$$

$$\tilde{E}\left[\frac{\varphi(T, V)Y}{\hat{\pi}(T, X)} - \tilde{\beta}^{\top}\hat{\xi} - \sum_{t\in\mathcal{T}} \varphi(t, V)c_1(t, V; \theta_1)\right] = 0, \tag{20}$$

where $\eta(T, Y, X; \theta_1) = \varphi(T, V)(Y - c_1(T, V; \theta_1))$ is replaced by $\eta^{\dagger}(T, Y, X) = \varphi(T, V)Y$, free of θ_1. Theorems 3–4 remain valid with the modification of η to $\eta^{\dagger}$. If the propensity score model is correctly specified, then $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ are asymptotically equivalent to the optimal estimator in the class of estimators that are solutions to (19) or (20), with $\tilde{\beta}$ replaced by a $\dim(\hbar) \times d$ matrix b. Replace condition (15) by the condition that

$$E[\eta^{\dagger}(T, Y, X)|T = t, X] = c^{\top}\hbar^{*}(t, X) \text{ almost surely}$$

for some $\dim(\hbar) \times d$ matrix c. If this condition holds, then $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ are asymptotically equivalent to the optimal estimator in the class of estimators that are solutions to (8). Moreover, $\tilde{\theta}_{\text{reg}}$ remains consistent even if the propensity score model is misspecified. Consider the following choices similar to (17) and (18):

$$\begin{aligned}\hbar^{(1)}(t, X) &= \hat{E}(\eta^{\dagger}|T = t, X)\\ \text{or } &= [1\{t = j\}\hat{E}^{\top}(\eta^{\dagger}|T = j, X)]^{\top}_{j\in\mathcal{T}}.\end{aligned}$$

Then $\hat{\vartheta}_{\text{lik}}$ and $\tilde{\vartheta}_{\text{reg}}$ are weakly locally efficient and $\tilde{\theta}_{\text{reg}}$ is doubly robust.

3.6. Further Consideration

Different roles of models. The propensity score and outcome regression models are related differently to a structural model, which parameterizes $P(Y_t|V)$. The propensity score model parameterizes $P(T = t|X)$, and hence is unrelated to the structural model and can be estimated from the data (T_i, X_i) without Y_i $(i = 1, \ldots, n)$. In contrast, the outcome regression model parameterizes $P(Y|T = t, X) = P(Y_t|X)$ and hence is related to the structural model due to the law of iterated expectations $E[E(Y_t|X)|V] = E(Y_t|V)$. The outcome regression model and the structural model can be incompatible with each other in the sense that there exists no data-generating distribution satisfying both models.

The propensity score and outcome regression models play different roles in the development of our methods, even though double robustness depends equally on both models. See Tan (2006, 2007) for related discussions. Our strategy for data analysis is to first build and check propensity score models, and then incorporate outcome regression models for variance and bias reduction. A referee pointed out that standard methods (e.g., ROC approaches) for checking model (3) as a binary-outcome regression model are not fully satisfactory. Our approach is to examine the balance of multiple covariate across treated and untreated groups after inverse weighting as in Tan (2006, Figure 1).

Homogeneous outcome regression models. A particular outcome regression model can be specified such that it is a submodel to and always compatible with the nested structural mean model (5) or (6). Consider the outcome regression model

$$E(Y_t|X) = \Psi[g_0(X;\alpha_0) + c_1(t, V;\theta_1)], \tag{21}$$

where Ψ is the identity or exponential link, $g_0(X;\alpha_0)$ and $g_1(t, X;\alpha_1) = c_1(t, V;\theta_1)$ are known functions such that $c_1(0, V;\theta_1) = 0$, and α_0 and $\alpha_1 = \theta_1$ are vectors of parameters. This model states that additive or multiplicative causal effects are homogeneous across subpopulations with fixed V but otherwise different X.

Consider the choices (17) and (18), or the simpler choices

$$\begin{aligned}\hbar^{(1)}(t, X) &= \varphi(t, V)\hat{E}(Y|T = 0, X)\\ \text{or } &= [1\{t = j\}\varphi^\top(j, V)]^\top_{j\in\mathcal{T}}\hat{E}(Y|T = 0, X).\end{aligned}$$

In the Supplementary Appendix, we show that $\hat{\vartheta}_{\text{lik}}$, $\hat{\vartheta}_{\text{reg}}$, and $\tilde{\vartheta}_{\text{reg}}$ are weakly locally efficient and doubly robust for $\vartheta = \theta_1$ in the nested structural mean model (5) or (6) when the homogeneous outcome regression model (21) is specified.

Binary treatments. The mapping (14) from $\hbar$ to h is many-to-one in that a given function h can be obtained from many different $\hbar$. In the case where $\mathcal{T} = \{0, 1\}$, we can exploit this relationship to derive an alternative regression estimator that is a more direct extension than $\tilde{\vartheta}_{\text{reg}}$ to the regression estimator in Tan (2006).

Write $\pi(X;\gamma) = \pi(1, X;\gamma)$. For $h(t, X;\gamma)$ such that $\sum_{t=0,1} h(t, X;\gamma) = 0$, write $h(X;\gamma) = h(1, X;\gamma)$ and let

$$\hbar_0(t, X;\gamma) = -\frac{1\{t = 0\}}{\pi(X;\gamma)}h(X;\gamma), \quad \hbar_1(t, X;\gamma) = \frac{1\{t = 1\}}{1 - \pi(X;\gamma)}h(X;\gamma),$$

which both give rise to $h(t, X;\gamma)$ under mapping (14). Let $\zeta_0(T, X;\gamma)$ and $\zeta_1(T, X;\gamma)$ be the ζ-variables based on $\hbar_0$ and $\hbar_1$. Define $\tilde{\tilde{\vartheta}}_{\text{reg}}$ and $\tilde{\tilde{\vartheta}}^{(\text{m})}_{\text{reg}}$ as solutions to $\tilde{E}[\tilde{\tilde{\tau}}_{\text{reg}}(\vartheta)] = 0$ and $\tilde{E}[\tilde{\tilde{\tau}}^{(\text{m})}_{\text{reg}}(\vartheta)] = 0$, where $\tilde{\tilde{\tau}}_{\text{reg}}(\vartheta) = \hat{\tau}_{\text{init}}(\vartheta) - \tilde{\tilde{\beta}}^\top(\vartheta)\hat{\xi}$ with

$$\tilde{\tilde{\beta}}(\vartheta) = \sum_{t=0,1} \tilde{E}^{-1}(\hat{\xi}\hat{\zeta}_t^\top)\tilde{E}[\hat{\xi}\hat{\tau}^\top_{\text{init}}(\vartheta)1\{T = t\}]$$

and $\tilde{\tilde{\tau}}^{(\text{m})}_{\text{reg}}(\vartheta)$ is the same as $\tilde{\tilde{\tau}}_{\text{reg}}(\vartheta)$ with $\hat{\xi}$ and $\hat{\zeta}_t$ replaced by $(\hat{\xi}^{(0)\top}, \hat{\xi}^{(1)\top})^\top$ and $(\hat{\zeta}_t^{(0)\top}, \hat{\zeta}_t^{(1)\top})^\top$. If the propensity score model is correctly specified, then $\tilde{\tilde{\vartheta}}_{\text{reg}}$ is asymptotically equivalent to $\tilde{\vartheta}_{\text{reg}}$

and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ is to $\tilde{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$. Replace condition (15) by the condition that

$$E[\eta(T, Y, X; \vartheta)|T = 0, X] = c_0^{\top} \hbar_0^*(0, X) \text{ and}$$
$$E[\eta(T, Y, X; \vartheta)|T = 1, X] = c_1^{\top} \hbar_1^*(1, X) \text{ almost surely}$$

for some $\dim(\hbar) \times d$ matrices c_0 and c_1. If this condition holds, then $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ and $\tilde{\vartheta}_{\mathrm{reg}}$ are asymptotically efficient in the class of estimators that are solutions to (8). Moreover, $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ remain consistent even if the propensity score model is misspecified. The foregoing condition is generally weaker than (15) because c_0 and c_1 can differ from each other. For $h^{(1)}(t, X; \gamma)$ determined by taking $\hbar^{(1)}(t, X; \gamma)$ as (18), the resulting $\hbar_0^{(1)}(0, X; \gamma)$ or $\hbar_1^{(1)}(1, X; \gamma)$ contains more nonzero functions than, respectively, $\hat{E}(\eta|T = 0, X)$ or $\hat{E}(\eta|T = 1, X)$, the vector of nonzero functions in $\hbar^{(1)}(0, X; \gamma)$ or $\hbar^{(1)}(1, X; \gamma)$.

3.7. Comparison

Table 1 presents a comparison of various estimators in terms of three asymptotic properties: intrinsically efficient or optimal in using control variates as discussed in Corollary 1, weakly locally efficient, and doubly robust.

The estimator $\tilde{\vartheta}_{\mathrm{reg}}$ is structurally similar to $\tilde{\vartheta}_{\mathrm{aug}}$, except that the special regression coefficient $\tilde{\beta}$ is associated with $\hat{\xi}$ in $\tilde{\vartheta}_{\mathrm{reg}}$ instead of the identity matrix in $\tilde{\vartheta}_{\mathrm{aug}}$. The two estimators $\tilde{\vartheta}_{\mathrm{reg}}$ and $\tilde{\vartheta}_{\mathrm{aug}}$ are consistent if either the propensity score model or the outcome regression model is correct, and equally efficient if both the propensity score model and the outcome regression model are correct. However, $\tilde{\vartheta}_{\mathrm{reg}}$ is asymptotically at least as efficient as $\tilde{\vartheta}_{\mathrm{aug}}$ if the outcome regression model is misspecified but the propensity score model is correct, due to the fact that $\tilde{\vartheta}_{\mathrm{reg}}$, but not $\tilde{\vartheta}_{\mathrm{aug}}$, is intrinsically efficient. (Be aware that $\tilde{\vartheta}_{\mathrm{reg}}$ is not necessarily more efficient than $\tilde{\vartheta}_{\mathrm{aug}}$ or vice versa if the propensity score model is misspecified but the outcome regression model is correct.) As a further implication of intrinsic efficiency, $\tilde{\vartheta}_{\mathrm{reg}}$, but not $\tilde{\vartheta}_{\mathrm{aug}}$, is asymptotically guaranteed to gain efficiency over $\tilde{\vartheta}_{\mathrm{init}}$ as long as the propensity score model is correct.

TABLE 1: Theoretical comparison of estimators.

	Section 2.2		Section 3.4			Sections 3.1 and 3.2		
	$\hat{\vartheta}_{\mathrm{init}}$	$\hat{\vartheta}_{\mathrm{aug}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$	$\tilde{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$	$\hat{\vartheta}_{\mathrm{reg}}$	$\tilde{\vartheta}_{\mathrm{reg}}$	$\hat{\vartheta}_{\mathrm{lik}}$
Intrinsically efficient	×	×	×	×	×	✓	✓	✓
Weakly locally efficient	×	✓	✓	✓	✓	✓	✓	✓
Doubly robust	×	✓	×	✓	×	×	✓	×

$\hat{\vartheta}_{\mathrm{init}}$ and $\hat{\vartheta}_{\mathrm{aug}}$ are called IPW and AIPW (or *G* and AG) estimators for marginal (or nested) structural models; $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$, $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$, and $\tilde{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$ are defined by removing constraints C2 and control variates $\hat{s}$ in the construction of $\hat{\vartheta}_{\mathrm{lik}}$, $\hat{\vartheta}_{\mathrm{reg}}$, and $\tilde{\vartheta}_{\mathrm{reg}}$. For a binary treatment (Section 3.6), $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ satisfy the same properties as do $\tilde{\vartheta}_{\mathrm{reg}}$ and $\tilde{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$, respectively.

4. SIMULATION STUDY

Assume that T is binary with $P(T=1)=P(T=0)=1/2$ and that X_1 is continuous taking values in $(-5/2, 5/2)$ and X_2 is discrete taking values in $\{0, 1, 2, 3\}$ with

$$P(X_2 = x_2|T=1) \propto a_1(x_2)\exp(x_2/2),$$

$$P(X_1|X_2 = x_2, T=1) \text{ is the restriction of } \mathrm{N}[d(x_2/2-1/4),\ \sigma_1^2] \text{ to } (-5/2,\ 5/2),$$

$$P(X_2 = x_2|T=0) \propto a_0(x_2)\exp(-x_2/2),$$

$$P(X_1|X_2 = x_2, T=0) \text{ is the restriction of } \mathrm{N}[d(x_2/2-5/4), \sigma_1^2] \text{ to } (-5/2, 5/2),$$

where $a_1(x_2)$ or $a_0(x_2)$ is the probability of $\mathrm{N}[d(x_2/2-1/4), \sigma_1^2]$ or $\mathrm{N}[d(x_2/2-5/4), \sigma_0^2]$ in $(-5/2, 5/2)$. Moreover, assume that $P(Y_t|X)$ is normal with mean $E(Y_t|X)$ and variance δ^2, where $X=(X_1, X_2)$. It is easy under this design to assess the degree of overlap between the distributions of X within $\{T=1\}$ and $\{T=0\}$. For a single covariate distributed as $\mathrm{N}(d, \sigma_1^2)$ and $\mathrm{N}(-d, \sigma_0^2)$ within $\{T=1\}$ and $\{T=0\}$, values of d between 1/8 and 1/2 and of σ_1^2/σ_0^2 between 1/2 and 2 are typical in practice (Rubin, 1973).

Consider the following simulation settings: (i) $d=1/4$ or $1/2$, (ii) $(\sigma_0^2, \sigma_1^2) = (1,1), (4/3, 2/3)$, or $(2/3, 4/3)$, (iii) $E(Y_0|X) = \mu_{0,\mathrm{LIN}}(X)$, $\mu_{0,\mathrm{EXP}}(X)$, $\mu_{0,\mathrm{LIN}}(X)$, or $\mu_{0,\mathrm{LIN}}(X)$, and $E(Y_1|X) = 1 + X_2 + \mu_{0,\mathrm{LIN}}(X)$, $1 + X_2 + \mu_{0,\mathrm{EXP}}(X)$, $\mu_{1,\mathrm{LIN}}(X)$, or $1 + X_2 + \mu_{0,\mathrm{EXP}}(X)$, respectively, where $\mu_{0,\mathrm{LIN}}(X) = 1 + X_1 + X_2 + 0.6X_2(1+X_1)$, $\mu_{0,\mathrm{EXP}} = \exp(X_1) + X_2 + 0.6X_2\exp(X_1)$, and $\mu_{1,\mathrm{LIN}}(X) = 1 + .5X_1 + X_2 + 0.6X_2(1+.5X_1)$, and (iv) $\delta^2 = 1/2$.

4.1. Marginal and Nested Structural Mean Models

For subpopulation causal inference, consider the model with the identity link

$$Y_t|X_2 \sim \mathrm{factor}(X_2) + t + t : X_2, \tag{22}$$

where factor(X_2) denotes X_2 as a factor and $t : X_2$ denotes the product tX_2, and "+" denotes a linear combination of predictors. See McCullagh & Nelder (1989, Section 3.4) for the standard notation for specification of linear predictors in generalized linear models. Model (22) can be treated as marginal structural model (4) or nested structural model (5) with $V = X_2$, where $c_1(t, V) \sim t + t : X_2$ and $c_0(V) \sim \mathrm{factor}(X_2)$ spans all functions of X_2. Technically, this model is correct under the first two specifications of $E(Y_t|X)$ with $\theta_1 = (1,1)^\top$, but not under the last two. However, in the latter case, a pseudo-true value of θ_1 can be defined by the best approximation to $E(Y_t|X_2)$, and estimators are evaluated against this value (e.g., Neugebauer & van der Laan, 2003).

Two possible propensity score models are logit regression with linear predictors

$$T|X \sim X_1 + X_1^2 + X_2 + X_2^2 + X_1 : X_2 \quad \text{and} \quad T|X \sim X_1 + X_2.$$

The first model is correct under all three specifications of (σ_0^2, σ_1^2), whereas the second model is correct under only the first specification: $\sigma_0^2 = \sigma_1^2$. An outcome regression model is linear regression with linear predictors

$$Y|T, X \sim T + X_1 + X_2 + X_1 : X_2 + T : X_1 + T : X_2 + T : X_1 : X_2.$$

This model is correct under the first and third specifications of $E(Y_t|X)$, but not under the other two. Define the outcome regression (OR) estimator, $\hat{\vartheta}_{\mathrm{OR}}$, by the least-squares fit of $\hat{E}(Y_t|X_1, X_2)$ against $\mathrm{factor}(X_2) + t + t : X_2$. We compare $\hat{\vartheta}_{\mathrm{OR}}$ and the eight estimators in Table 1, except $\tilde{\vartheta}_{\mathrm{reg}}$

TABLE 2: Numerical comparison of estimators (marginal structural model).

	$\hat\vartheta_{\rm OR}$	$\hat\vartheta_{\rm init}$	$\hat\vartheta_{\rm aug}$	$\hat\vartheta^{\rm (m)}_{\rm reg}$	$\tilde{\tilde\vartheta}^{\rm (m)}_{\rm reg}$	$\hat\vartheta^{\rm (m)}_{\rm lik}$	$\hat\vartheta_{\rm reg}$	$\tilde{\tilde\vartheta}_{\rm reg}$	$\hat\vartheta_{\rm lik}$
Quadratic propensity score model (C-PS)									
LINP (C-OR)	0.000046	−0.0077	−0.00021	0.0088	0.00056	00050	0.016	0.00087	0.0064
	0.0343	0.104	0.0397	0.0446	0.0412	0.0409	0.0496	0.0483	0.0446
	···	0.104	0.0381	0.0401	0.0372	0.0401	0.0416	0.0428	0.0416
EXPP (M-OR)	−0.22	0.0024	0.0093	0.012	0.015	–0.00051	0.019	0.0033	0.010
	0.169	0.213	0.164	0.0566	0.0613	0.0551	0.0493	0.0527	0.0454
	···	0.209	0.157	0.0490	0.0548	0.0490	0.0419	0.0478	0.0419
LINN (C-OR)	0.00034	−0.0062	0.000082	0.0068	0.00085	0.0011	0.012	0.00012	0.0077
	0.0433	0.120	0.0475	0.0541	0.0486	0.0489	0.0605	0.0549	0.0539
	···	0.118	0.0459	0.0479	0.0452	0.0479	0.0491	0.0499	0.0491
EXPN (M-OR)	−0.085	−0.0045	0.0011	0.015	0.0075	0.0058	0.015	0.0011	0.0095
	0.0826	0.154	0.122	0.0952	0.0840	0.0845	0.109	0.0864	0.0954
	···	0.150	0.115	0.0871	0.0802	0.0871	0.0898	0.0830	0.0898
Linear propensity score model (M-PS)									
LINP (C-OR)	0.000046	−0.051	−0.00026	0.027	0.00075	0.00052	0.029	0.00093	0.012
	0.0343	0.271	0.0372	0.0512	0.0426	0.0418	0.0519	0.0446	0.0467
	···	0.272	0.0362	0.0459	0.0392	0.0459	0.0461	0.0410	0.0461
EXPP (M-OR)	−0.22	−0.35	−0.30	0.045	0.022	−0.0042	0.025	0.019	−0.012
	0.169	0.563	0.238	0.0568	0.0513	0.0507	0.0537	0.0490	0.0482
	···	0.563	0.240	0.0501	0.0473	0.0501	0.0490	0.0451	0.0490
LINN (C-OR)	0.00034	−0.058	0.000026	0.013	0.0010	0.0010	0.017	0.0012	0.012
	0.0433	0.261	0.0456	0.0614	0.0499	0.0497	0.0624	0.0517	0.0549
	···	0.262	0.0443	0.0535	0.0469	0.0535	0.0534	0.0484	0.0534
EXPN (M-OR)	−0.085	−0.036	−0.022	0.046	−0.00018	0.0013	0.051	−0.0011	0.019
	0.0826	0.283	0.0821	0.103	0.0823	0.0842	0.106	0.0834	0.0944
	···	0.286	0.0847	0.0949	0.0806	0.0944	0.0938	0.0816	0.0938

C-PS (or M-PS) indicates correct (or misspecified) propensity score model, C-OR (or M-OR) indicates correct (or misspecified outcome regression model, and LINP, EXPP, LINN, and EXPN correspond to the four specifications of $E(Y_t|X)$. The results are based on 5,000 Monte Carlo samples each of size 1,000. Each cell gives the bias (upper) and the standard deviation (middle) of the point estimates, and $\sqrt{\text{mean of the variance estimates}}$ (lower).

and $\tilde\vartheta^{\rm (m)}_{\rm reg}$ are replaced by $\tilde{\tilde\vartheta}_{\rm reg}$ and $\tilde{\tilde\vartheta}^{\rm (m)}_{\rm reg}$ in Section 3.7 to take advantage of the fact that T is binary. Choice (18) is used for $\hbar^{(1)}(t, X)$.

Table 2 summarizes the estimates of the slope in θ_1 for $d = 1/2$ and $(\sigma_0^2, \sigma_1^2) = (4/3, 2/3)$ with model (22) treated as a marginal structural model. Similar comparisons are obtained for other values of d and (σ_0^2, σ_1^2) or with model (22) treated as a nested structural model. Given the choices $\phi_{\rm conv}$ and $\varphi_{\rm conv}$ in Section 2.2, the estimators of θ_1 in $\hat\vartheta_{\rm init}$ and $\hat\vartheta_{\rm aug}$ (i.e., IPW and AIPW estimators) with (22) as a marginal structural model are identical to $\hat\vartheta_{\rm init}$ and $\hat\vartheta_{\rm aug}$ (i.e., G and AG estimators) with (22) as a nested structural model.

The estimator $\hat\vartheta_{\rm OR}$ performs well if the response model used is correct but yields serious biases otherwise. The estimator $\hat\vartheta_{\rm init}$ has considerable variances in all cases and has negligible biases if the propensity score model used is correct but not otherwise.

TABLE 3: Numerical comparison of estimators (nested structural AFT model).

	$\hat{\vartheta}_{\mathrm{OR}}$	$\hat{\vartheta}_{\mathrm{init}}$	$\hat{\vartheta}_{\mathrm{aug}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$	$\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$	$\hat{\vartheta}_{\mathrm{reg}}$	$\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$	$\hat{\vartheta}_{\mathrm{lik}}$
Quadratic propensity score model (C-PS)									
LINP (C-OR)	0.00049	−0.019	−0.000069	0.00021	0.00021	0.00024	0.00038	−0.00063	0.00053
	0.0364	0.188	0.0479	0.0466	0.0513	0.0469	0.0476	0.0510	0.0491
	···	0.196	0.0459	0.0439	0.0494	0.0439	0.0436	0.0491	0.0436
EXPP (M-OR)	−0.13	−0.013	0.0046	0.0063	0.0054	0.0053	0.0021	0.0035	0.0021
	0.0655	0.239	0.102	0.0651	0.0658	0.0660	0.0576	0.0646	0.0591
	···	0.248	0.0972	0.0594	0.0629	0.0594	0.0533	0.0632	0.0533
Linear propensity score model (M-PS)									
LINP (C-OR)	0.00049	0.078	−0.00011	0.00015	0.00078	0.00013	0.00010	0.015	0.00043
	0.0364	0.410	0.0427	0.0426	0.0442	0.0426	0.0462	0.0521	0.0486
	···	0.417	0.0420	0.0424	0.0478	0.0424	0.0439	0.0602	0.0439
EXPP (M-OR)	−0.13	0.074	−0.00021	0.025	0.032	0.019	−0.024	–0.015	−0.016
	0.0655	0.410	0.0669	0.0653	0.0697	0.0677	0.0615	0.0660	0.0716
	···	0.417	0.0668	0.0650	0.0759	0.0650	0.0588	0.0832	0.0588

See the footnote of Table 2.

The estimators $\hat{\vartheta}_{\mathrm{aug}}$ and $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ have small biases if either the propensity score model or the response model is correct. If the propensity score model is correct, then both estimators have similar variances when the response model is correct, but $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ compared with $\hat{\vartheta}_{\mathrm{aug}}$ has variances reduced by a factor of 1–9 when the response model is wrong. The estimator $\hat{\vartheta}_{\mathrm{reg}}$ has similar variances as $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$, but has larger biases than $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ especially if the propensity score model is wrong. The estimators $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$ and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ are similar to $\hat{\vartheta}_{\mathrm{reg}}$ and $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$, respectively, but sometimes have larger biases and variances.

The estimators $\hat{\vartheta}_{\mathrm{lik}}$ and $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$ have similar mean squared errors as, respectively, $\hat{\vartheta}_{\mathrm{reg}}$ and $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$, or $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$, if the propensity score model is correct. Otherwise, $\hat{\vartheta}_{\mathrm{lik}}$ and $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$ appear to still have small mean squared errors whether or not the response model is correct, although they may not be consistent in theory. The estimators $\hat{\vartheta}_{\mathrm{lik}}$ and $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$ and marginalized $\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$ and $\tilde{\tilde{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ yield overall the smallest mean squared errors.

4.2. Nested Structural AFT Model

Suppose that values of Y are censored at $C = 10$ in the setting described at the beginning of Section 4. Only the first two specifications of $P(Y_t|X)$ are considered. Model (22) is treated as nested structural AFT model (7). The outcome regression model

$$Y|T, X \sim T + X_1 + X_2 + X_1 : X_2 + T : X_2$$

is treated as AFT model (2) and fit by the method of Buckley & James (1979).

Table 3 summarizes the estimates of the slope in θ_1 for $d = 1/2$ and $(\sigma_0^2, \sigma_1^2) = (4/3, 2/3)$. Choice (9) is used for the transformation function κ. Comparisons of the nine estimators are qualitatively similar to those in Section 4.1. The estimator $\hat{\vartheta}_{\mathrm{OR}}$ has serious biases if the response model is wrong. The estimator $\hat{\vartheta}_{\mathrm{init}}$ has considerable variances in all cases and serious biases if the propensity score model is wrong. The estimators $\hat{\vartheta}_{\mathrm{aug}}$, $\hat{\vartheta}_{\mathrm{reg}}$, $\tilde{\tilde{\vartheta}}_{\mathrm{reg}}$, and $\hat{\vartheta}_{\mathrm{lik}}$ have small biases if either the propensity score model or the response model is correct. If the propensity score model is correct, then these estimators have similar variances when the response model is correct, but $\hat{\vartheta}_{\mathrm{reg}}$,

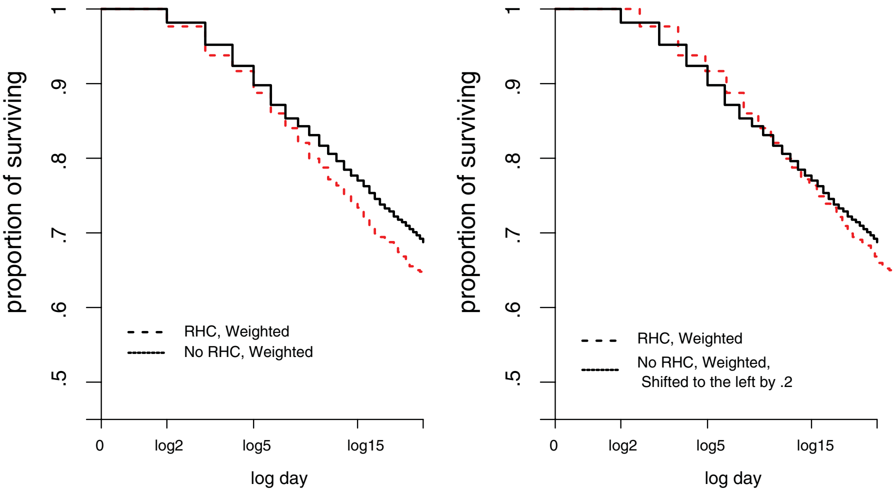

FIGURE 1: Thirty-day survival curves. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$\tilde{\tilde{\vartheta}}_{\text{reg}}$, and $\hat{\vartheta}_{\text{lik}}$ compared with $\hat{\vartheta}_{\text{aug}}$ have variances reduced by a factor of 1–3 when the response model is wrong.

5. DATA ANALYSIS

RHC is a medical procedure performed daily in hospitals since the 1970s. Connors et al.'s (1996) observational study was influential, raising the concern that RHC may not benefit patients and may in fact cause harm. The study included 5,735 critically ill patients, who were admitted to the intensive care units of five medical centers. Data were collected on treatment status T ($= 1$ if RHC was used within 24 h of admission and 0 otherwise), health outcome Y (log of survival time up to 30 days), and a comprehensive list of 75 covariates X (specified by a panel of seven specialists in critical care). The survival time was censored at 30 days.

Connors et al. (1996) used propensity score matching to identify 1,008 pairs of patients managed with and without RHC, and estimated the odds ratio of surviving 30 days based on these matched patients. They also performed regression analysis in the 5,735 patients using Cox proportional hazards models to estimate the relative hazard of death by 30 days overall and in prespecified patient groups. It seems difficult to integrate the results from the two analyses, because the estimates are based on different sets of patients. We analyze the data using nested structural AFT models and illustrate the values of the new methods.

First, we fit a propensity score model, taken as the final model in a sequence of logit regression models built and checked in Tan (2006). The estimated joint distribution of (X, Y_0) or (X, Y_1) is supported on the untreated or the treated group, respectively. Two copies of the marginal distribution of each covariate can be extracted from the joint distributions. See the weighted histograms in Tan (2006, Figure 2). The marginal distribution of each potential outcome can also be extracted. Figure 1(left) replicates the weighted survival curves from Tan (2006, Figure 2) in the log scale of time. The RHC curve is always lower than the no-RHC curve, indicating a harmful effect of RHC. The RHC and no-RHC survival curves appear to differ approximately by a shift in the log scale of time. Figure 1(right) presents the weighted survival curves with the no-RHC curve shifted by -0.2.

As motivated by the foregoing exploratory analysis, consider a nested structural AFT model in the overall population: $P(Y_1) = P(Y_0 + \theta_1)$. The AFT regression model (2) includes the main

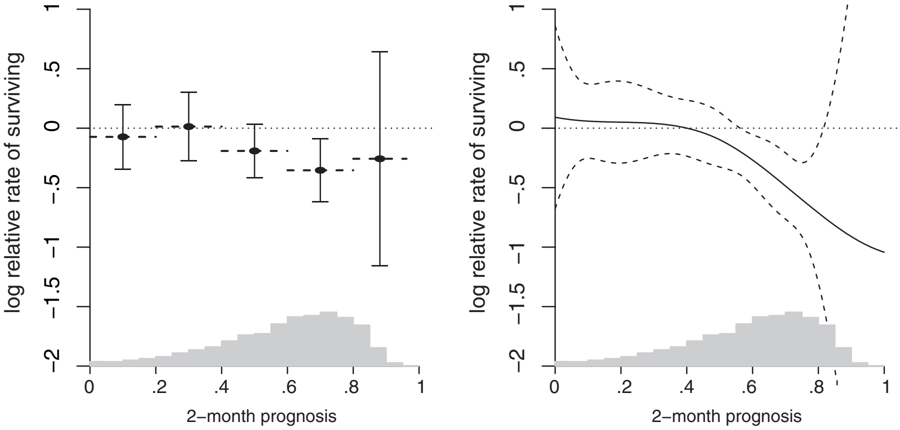

FIGURE 2: Log relative rate of surviving related to prognosis.

TABLE 4: Overall log relative rate of surviving.

	$\hat{\vartheta}_{\mathrm{OR}}$	$\hat{\vartheta}_{\mathrm{init}}$	$\hat{\vartheta}_{\mathrm{aug}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{reg}}$	$\tilde{\hat{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$	$\hat{\vartheta}^{(\mathrm{m})}_{\mathrm{lik}}$	$\hat{\vartheta}_{\mathrm{reg}}$	$\tilde{\hat{\vartheta}}_{\mathrm{reg}}$	$\hat{\vartheta}_{\mathrm{lik}}$
Null $\hbar^{(1)}$	···	–0.217	···	–0.181	–0.181	–0.181	–0.180	–0.180	–0.180
	···	0.248	···	0.0666	0.0665	0.0666	0.0659	0.0659	0.0659
Fitted $\hbar^{(1)}$	–0.273	···	–0.204	–0.203	–0.200	–0.203	–0.198	–0.193	–0.196
	0.0569	···	0.0655	0.0649	0.0645	0.0649	0.0651	0.0652	0.0651

Each cell gives point estimate (upper) and standard error (lower).

effects of T and X, and is estimated by the method of Buckley & James (1979). Table 4 summarizes the estimates of θ_1 for the following choices: (i) $\hbar^{(1)}(t, X)$ is null with no functions included and (ii) $\hbar^{(1)}(t, X) = [1\{t = 0\}\varphi^{\top}(0, V), 1\{t = 1\}\varphi^{\top}(1, V)]^{\top} \hat{E}[\kappa(Y_0)|X]$. All the estimates except $\hat{\vartheta}_{\mathrm{init}}$ (i.e., the G estimate) have similar standard errors and lead to rejection of $\theta_1 = 0$ at 1% significance level. The estimate $\hat{\vartheta}_{\mathrm{OR}}$ differs by approximately 1 standard error from the remaining estimates. This difference suggests that the AFT regression model with main effects only might be misspecified, given our model checking on the propensity score model. Remarkably, the regression and likelihood estimates with the null $\hbar^{(1)}$ have similar standard errors to that of $\hat{\vartheta}_{\mathrm{aug}}$. But the regression and likelihood estimates with the fitted $\hbar^{(1)}$ yields little variance reduction over those with the null $\hbar^{(1)}$, which indicates that the proportion of variability in survival time explained by the AFT regression model is small.

Connors et al. (1996) examined the association of RHC with survival time in patient groups classified by the predicted probability of being alive at 2 months on study entry. Consider a nested structural AFT model: $P(Y_1|V) = P(Y_0 + \theta_1^{\top} c_1(V)|V)$, where V is the 2-month prognosis variable, and $c_1(V)$ is specified as (i) a five-level factor for V in the intervals divided by 0.2, 0.4, 0.6, 0.8, in contrast to the quintiles in Connors et al. (1996) or (ii) a vector of basis functions for cubic splines on [0, 1] with 5 degrees of freedom including the intercept. Figure 2 shows the estimates of $\theta_1^{\top} c_1(V)$ based on $\tilde{\hat{\vartheta}}^{(\mathrm{m})}_{\mathrm{reg}}$ for the two specifications of $c_1(V)$. The AFT regression model (2) includes the main effects of T and X and the interactions between T and $c_1(V)$. The causal effect of RHC in accelerating the time to death by 30 days appears nonzero at 95% confidence level only in patients with a predicted probability of 2-month survival at study entry between 0.6 and 0.8.

6. CONCLUSION

We consider Robins's marginal and nested structural models and propose likelihood and regression estimators in the cross-sectional setting. These estimators are, in theory, asymptotically more efficient than AIPW and AG estimators if the propensity score model is correct but the outcome regression model is misspecified. In a simulation study with various settings, they yield overall the smallest mean squared errors and sometimes have substantially smaller variances than AIPW and AG estimators. On the other hand, there are potential limitations. The likelihood estimator is generally not doubly robust. The tilde regression estimator, although doubly robust, may yield outlying values when the fitted propensity scores are close to 0 for some treated subjects. It is important to develop appropriate methods to address these issues.

ACKNOWLEDGEMENTS

The author thanks the editor, an associate editor, and two referees for constructive comments. A previous version of this work was completed at Johns Hopkins University and titled "Nonparametric likelihood and further development of inverse weighting and G-estimation for marginal and nested structural models." The research was supported by the U.S. National Science Foundation.

BIBLIOGRAPHY

H. Bang & J. M. Robins (2005). Doubly robust estimating in missing data and causal inference models. *Biometrics*, 61, 962–972.

J. Buckley & I. James (1979). Linear regression with censored data. *Biometrika*, 66, 429–436.

A. F. Connors, Jr., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, Jr., D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, Jr, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn & W. A. Knaus. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, 276, 889–897.

S. Goetgeluk, S. Vansteelandt & E. Goetghebeur (2009). Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B*, 70, 1049–1066.

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

L. P. Hansen (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

J. D. Y. Kang & J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22, 523–539.

E. L. Kaplan & P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

J. Keifer & J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887–906.

A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae & Z. Tan (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society, Series B*, 65, 585–618.

P. McCullagh & J. Nelder (1989). "*Generalized Linear Models*," 2nd ed., Chapman & Hall, New York.

S. A. Murphy & A. W. van der Vaart (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association*, 95, 449–485.

R. Neugebauer & M. J. van der Laan (2003). Locally efficient estimation of nonparametric causal effects on mean outcomes in longitudinal studies. U.C. Berkeley Division of Biostatistics Working Paper 134.

W. K. Newey & R. J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219–255.

J. Neyman (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Translated in *Statistical Science*, 1990, 5, 465–480.

A. Owen (2001). "*Empirical Likelihood*," Chapman & Hall, New York.

J. Qin & J. Lawless (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300–326.

J. Qin & B. Zhang (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society, Series B*, 69, 101–122.

J. M. Robins (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in "*Health Service Research Methodology: A Focus on AIDS*," L. Sechrest, H. Freeman & A. Mulley, editors, U.S. Public Health Service, Washington, DC, pp 113–159.

J. M. Robins (1999). Marginal structural models versus structural nested models as tools for causal inference, in "*Statistical Models in Epidemiology: The Environment and Clinical Trials*," E. M. Halloran & D. Berry, editors, Springer, New York, pp 95–134.

J. M. Robins & Y. Ritov (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine*, 16, 285–319.

J. M. Robins, A. Rotnizky & L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

P. R. Rosenbaum & D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

D. B. Rubin (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29, 185–203.

D. B. Rubin (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

D. O. Scharfstein, A. Rotnitzky & J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096–1146.

Z. Tan (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association*, 99, 1027–1036.

Z. Tan (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101, 1619–1637.

Z. Tan (2007). Comment: Understanding OR, PS, and DR. *Statistical Science*, 22, 560–568.

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

Y. Vardi (1985). Empirical distributions in selection bias models. *Annals of Statistics*, 25, 178–203.

H. White (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Received 1 February 2010
Accepted 22 May 2010