Biometrika (2008), **94**, 2, *pp*. 1–22

Printed in Great Britain
Advance Access publication on 31 July 2008

Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting

BY Z. TAN

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.
ztan@stat.rutgers.edu

SUMMARY

Consider the problem of estimating the mean of an outcome in the presence of missing data or estimating population average treatment effects in causal inference. A doubly robust estimator remains consistent if an outcome regression model or a propensity score model is correctly specified. We build on the nonparametric likelihood approach of Tan and propose new doubly robust estimators. These estimators have desirable properties in efficiency if the propensity score model is correctly specified, and in boundedness even if the inverse probability weights are highly variable. We compare new and existing estimators in a simulation study and find that the robustified likelihood estimators yield overall the smallest mean squared errors.

Some key words: Causal inference; Double robustness; Inverse weighting; Missing data; Nonparametric likelihood; Propensity score.

1. INTRODUCTION

Consider the problem of estimating the mean of an outcome in the presence of missing data under ignorability (Rubin, 1976). A related problem is to estimate population average treatment effects under no unmeasured confounding in causal inference (Neyman, 1923; Rubin, 1974). Such problems can be handled in two different ways. One approach is to model the mean of the outcome given covariates, called the outcome regression function, and derive an estimator based on the fitted values for observed and missing outcomes. The other approach is to model the probability of non-missingness given the covariates, called the propensity score (Rosenbaum & Rubin, 1983), and derive an estimator through inverse probability weighting of observed outcomes. Inverse-probability-weighted estimators are central to the semiparametric theory of estimation with missing data (e.g., Tsiatis, 2006; van der Laan & Robins, 2003).

The two approaches rely on different modelling assumptions and one does not necessarily dominate the other (Tan, 2007). A doubly robust approach makes use of both the outcome regression model and the propensity score model and derives an estimator that remains consistent if either of the two models is correctly specified. A prototypical doubly robust estimator is the augmented inverse-probability-weighted estimator of Robins et al. (1994). Recently, a number of alternative doubly robust estimators have been proposed. See Kang & Schafer (2007) and the related discussions. All existing doubly robust estimators are locally efficient: they attain the semiparametric variance bound, and hence asymptotically equivalent to each other, if both the propensity score model and the outcome regression model are correctly specified. Therefore, it is important to compare doubly robust estimators in their statistical properties if only one of the models is correctly specified or if both models are misspecified.

We review various doubly robust estimators and highlight statistical criteria underlying their construction. Some estimators are intrinsically efficient: if the propensity score model is correctly specified, then each of them is asymptotically efficient among a class of augmented inverse-probability-weighted estimators that use the same fitted outcome regression function (Tan, 2006, 2007). Some estimators are improved-locally efficient: if the propensity score model is correctly specified, then they are asymptotically at least as efficient as the augmented inverse-probability-weighted estimator that uses the true propensity score and an optimally fitted outcome regression function (Rubin & van der Laan, 2008; Tan, 2008). Some estimators are population-bounded or sample-bounded: they lie within the range of all possible values or that of observed values of the outcome (Robins et al., 2007). The properties of boundedness rule out estimates outside the population or sample range even when the inverse probability weights are highly variable.

We propose a robustification of the likelihood estimator of Tan (2006), named calibrated likelihood estimator, by calibrating the coefficients in a linear, extended propensity score model. The estimator is computationally convenient, involving two steps of maximizing concave functions. Moreover, the estimator is locally and intrinsically efficient and sample-bounded, and is further improved-locally efficient if the outcome regression function is suitably estimated. No existing doubly robust estimators achieve these four properties simultaneously.

We further derive a robustification of the likelihood estimator of Tan (2006), named augmented likelihood estimator, by incorporating an augmentation term. This estimator satisfies only a weaker form of boundedness than population and sample boundedness. We compare new and existing estimators in a simulation study and find that the calibrated and augmented likelihood estimators yield overall the smallest mean squared errors.

2. Missing data problems

2·1. *Setup*

Let X be a vector of covariates and Y be an outcome. The variables X are always observed, but Y may be missing. Let R be the non-missing indicator such that $R = 1$ or 0 if Y is observed or missing respectively. Throughout, assume that the missing data mechanism is ignorable, that is, R and Y are conditionally independent given X (Rubin, 1976).

Suppose that an independent and identically distributed sample of n units is available. The observed data consist of $(X_i, R_i, R_i Y_i)$, $i = 1, \ldots, n$. Our objective is to estimate the population mean $\mu = E(Y)$. Although this problem is simple to describe, it provides a basic setting for us to investigate methods for handling missing data.

2·2. *Models*

There are two different ways of postulating dimension-reduction assumptions to obtain consistent and asymptotically normal estimators of μ. One approach is to specify a parametric model for the outcome regression function $m(X) = E(Y \mid X)$ in the form

$$E(Y \mid X) = m(X;\alpha) = \Psi\{\alpha^{\mathrm{T}} g(X)\}, \tag{1}$$

where Ψ is an inverse link function, $g(x)$ is a vector of known functions including the constant 1, and α is a vector of unknown parameters. Let $\hat{\alpha}_{\mathrm{OLS}}$ be the maximum quasi-likelihood estimator of α or its variant. For concreteness, fix $\hat{\alpha}_{\mathrm{OLS}}$ as the estimator that solves the equation $0 = \tilde{E}\,[R\{Y - m(X;\alpha)\}g(X)]$, where $\tilde{E}$ denotes sample average. Let $\hat{\mu}_{\mathrm{OLS}} = \tilde{E}\{\hat{m}_{\mathrm{OLS}}(X)\}$, where $\hat{m}_{\mathrm{OLS}}(X) = m(X;\hat{\alpha}_{\mathrm{OLS}})$. Under regularity conditions, if model (1) is correctly specified, then $\hat{\mu}_{\mathrm{OLS}}$ is consistent and asymptotically normal, with asymptotic variance no greater than the semiparametric variance bound, provided that $E(Y^2) < \infty$.

The other approach is to specify a parametric model for the propensity score $\pi(X) = P(R = 1 \mid X)$ in the form

$$P(R = 1 \mid X) = \pi(X; \gamma) = \Pi\{\gamma^{\mathrm{T}} f(X)\}, \tag{2}$$

where Π is an inverse link function, $f(x)$ is a vector of known functions, and γ is a vector of unknown parameters. Let $\hat{\gamma}_{\mathrm{ML}}$ be the maximum likelihood estimator of γ and hence a solution to the equation $0 = \tilde{E}\,[\{R - \pi(X;\gamma)\}\varrho(X;\gamma)f(X)]$, where $\varrho(X;\gamma) = \Pi'\{\gamma^{\mathrm{T}} f(X)\}/[\pi(X;\gamma)\{1 - \pi(X;\gamma)\}]$ and Π' is the derivative of Π. Two non-augmented inverse-probability-weighted estimators are

$$\hat{\mu}_{\text{IPW}} = \tilde{E}\left\{\frac{RY}{\hat{\pi}_{\mathrm{ML}}(X)}\right\}, \quad \hat{\mu}_{\text{IPW,ratio}} = \tilde{E}\left\{\frac{RY}{\hat{\pi}_{\mathrm{ML}}(X)}\right\}\Big/\tilde{E}\left\{\frac{R}{\hat{\pi}_{\mathrm{ML}}(X)}\right\},$$

where $\hat{\pi}_{\mathrm{ML}}(X) = \pi(X; \hat{\gamma}_{\mathrm{ML}})$. Under regularity conditions, if model (2) is correctly specified, then $\hat{\mu}_{\text{IPW}}$ and $\hat{\mu}_{\text{IPW,ratio}}$ are consistent and asymptotically normal, with asymptotic variances no smaller than the semiparametric variance bound, provided that $E\{\pi^{-1}(X)\} < \infty$ and $E\{Y^2\pi^{-1}(X)\} < \infty$. See Tan (2007) for a comparison between the two approaches.

2·3. *Existing estimators*

The estimator $\hat{\mu}_{\text{OR}}$ is based on model (1) only, and $\hat{\mu}_{\text{IPW}}$ and $\hat{\mu}_{\text{IPW,ratio}}$ are based on model (2) only. Alternatively, a range of estimators have been proposed by using both model (1) and model (2) to gain efficiency and robustness. Many such estimators can be cast in the form

$$\hat{\mu}(\hat{\pi}, \hat{m}) = \tilde{E}\left[\frac{RY}{\hat{\pi}(X)} - \left\{\frac{R}{\hat{\pi}(X)} - 1\right\}\hat{m}(X)\right] = \tilde{E}\left[\hat{m}(X) + \frac{R}{\hat{\pi}(X)}\{Y - \hat{m}(X)\}\right],$$

where $\hat{\pi}(X)$ and $\hat{m}(X)$ are fitted values of $\pi(X)$ and $m(X)$ respectively. See Kang & Schafer (2007), Robins et al. (2007), and Tan (2006, 2007, 2008) for related discussions.

Consider the following estimators of μ, with the same choice $\hat{\pi}_{\mathrm{ML}}(X)$ for $\hat{\pi}(X)$ but different choices for $\hat{m}(X)$. Robins et al. (1994) proposed the estimator $\hat{\mu}_{\text{AIPW}} = \hat{\mu}(\hat{\pi}_{\mathrm{ML}}, \hat{m}_{\mathrm{OLS}})$. Scharfstein et al. (1999) suggested the estimator

$$\hat{\mu}_{\text{OLS,ext}} = \hat{\mu}\{\hat{\pi}_{\mathrm{ML}}, \hat{m}_{\text{ext}}(\hat{\pi}_{\mathrm{ML}})\} = \tilde{E}\{\hat{m}_{\text{ext}}(X; \hat{\pi}_{\mathrm{ML}})\},$$

where $\hat{m}_{\text{ext}}(X;\hat{\pi}) = m_{\text{ext}}\{X; \hat{\kappa}(\hat{\pi})\}$ and $\hat{\kappa}(\hat{\pi})$ is a solution to $0 = \tilde{E}[R\{Y - m_{\text{ext}}(X;\kappa)\}\{\hat{\pi}^{-1}(X), g^{\mathrm{T}}(X)\}^{\mathrm{T}}]$ for the extended outcome regression model $E(Y \mid X) = m_{\text{ext}}(X;\kappa) = \Psi\{\kappa_1\hat{\pi}^{-1}(X) + \kappa_2^{\mathrm{T}} g(X)\}$ with $\kappa = (\kappa_1, \kappa_2^{\mathrm{T}})^{\mathrm{T}}$. Kang & Schafer (2007) considered the estimator

$$\hat{\mu}_{\mathrm{WLS}} = \hat{\mu}\{\hat{\pi}_{\mathrm{ML}}, \hat{m}_{\mathrm{WLS}}(\hat{\pi}_{\mathrm{ML}})\} = \tilde{E}\{\hat{m}_{\mathrm{WLS}}(X; \hat{\pi}_{\mathrm{ML}})\},$$

where $\hat{m}_{\mathrm{WLS}}(X;\hat{\pi}) = m\{X; \hat{\alpha}_{\mathrm{WLS}}(\hat{\pi})\}$ and $\hat{\alpha}_{\mathrm{WLS}}(\hat{\pi})$ is a solution to $0 = \tilde{E}[R\hat{\pi}^{-1}(X)\{Y - m(X;\alpha)\}g(X)]$ and hence differs from $\hat{\alpha}_{\mathrm{OLS}}$ in using weight $\hat{\pi}^{-1}(X)$. Rubin & van der Laan (2008) proposed two related estimators

$$\hat{\mu}_{\mathrm{RV}} = \hat{\mu}\{\hat{\pi}_{\mathrm{ML}}, \hat{m}_{\mathrm{RV}}(\hat{\pi}_{\mathrm{ML}})\}, \quad \tilde{\mu}_{\mathrm{RV}} = \hat{\mu}\{\hat{\pi}_{\mathrm{ML}}, \tilde{m}_{\mathrm{RV}}(\hat{\pi}_{\mathrm{ML}})\},$$

where $\hat{m}_{\mathrm{RV}}(X;\hat{\pi}) = m\{X; \hat{\alpha}_{\mathrm{RV}}(\hat{\pi})\}$ and $\hat{\alpha}_{\mathrm{RV}}(\hat{\pi}) = \mathrm{argmin}_\alpha\, \tilde{E}([RY/\hat{\pi}(X) - \{R/\hat{\pi}(X) - 1\}m(X;\alpha)]^2)$ for the first estimator and $\tilde{m}_{\mathrm{RV}}(X;\hat{\pi}) = m\{X;\ \tilde{\alpha}_{\mathrm{RV}}(\hat{\pi})\}$ and $\tilde{\alpha}_{\mathrm{RV}}(\hat{\pi}) = \mathrm{argmin}_\alpha\, \tilde{E}[\{R/\hat{\pi}(X)\}\{R/\hat{\pi}(X) - 1\}\{Y - m(X;\alpha)\}^2]$ for the second estimator. The estimator $\tilde{\alpha}_{\mathrm{RV}}(\hat{\pi})$ is a weighted least-squares estimator using weight $\hat{\pi}^{-1}(X)\{\hat{\pi}^{-1}(X) - 1\}$. Our notation makes explicit the dependency of $\hat{m}_{\text{ext}}(\hat{\pi})$, $\hat{m}_{\mathrm{WLS}}(\hat{\pi})$, $\hat{m}_{\mathrm{RV}}(\hat{\pi})$, and $\tilde{m}_{\mathrm{RV}}(\hat{\pi})$ on $\hat{\pi}$.

The choice $\hat{\pi}_{\mathrm{ML}}(X)$ for $\hat{\pi}(X)$ is derived under model (2), independently of model (1). A more elaborate choice can be derived under an extended propensity score model with extra linear predictors depending on $\hat{m}(X)$. Consider the model

$$P(R=1 \mid X) = \pi_{\text{ext}}(X;\nu) = \Pi\left\{\nu_1^{\mathrm{T}} \frac{\hat{v}(X)}{\hat{\varrho}_{\text{ML}}(X)\hat{\pi}_{\text{ML}}(X)} + \nu_2^{\mathrm{T}} f(X)\right\}, \tag{3}$$

where $\nu = (\nu_1^{\mathrm{T}}, \nu_2^{\mathrm{T}})^{\mathrm{T}}$, $\hat{v}(X) = \{1, \hat{m}(X)\}^{\mathrm{T}}$, and $\hat{\varrho}_{\text{ML}}(X) = \varrho(X;\hat{\gamma}_{\text{ML}})$. Let $\hat{\nu}(\hat{m})$ be the maximum likelihood estimator of ν and write $\hat{\pi}_{\text{ext}}(X;\hat{m}) = \pi_{\text{ext}}\{X;\hat{\nu}(\hat{m})\}$. Substitution of $\hat{\pi}_{\text{ext}}(\hat{m}_{\text{OLS}})$ for $\hat{\pi}_{\text{ML}}$ in $\hat{\mu}_{\text{IPW}}$ yields the estimator of Rotnitzky & Robins (1995), $\hat{\mu}_{\text{IPW,ext}} = \hat{\mu}\{\hat{\pi}_{\text{ext}}(\hat{m}_{\text{OLS}}), 0\}$. For $\hat{m} = \hat{m}_{\text{OLS}}$ or $\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})$, substitution of $\hat{\pi}_{\text{ext}}(\hat{m})$ for $\hat{\pi}_{\text{ML}}$ in $\hat{\mu}(\hat{\pi}_{\text{ML}}, \hat{m})$, but not for that within $\hat{m}$, yields the estimators

$$\hat{\mu}_{\text{AIPW,ext}} = \hat{\mu}\{\hat{\pi}_{\text{ext}}(\hat{m}_{\text{OLS}}), \hat{m}_{\text{OLS}}\}, \quad \hat{\mu}_{\text{WLS,ext}} = \hat{\mu}[\hat{\pi}_{\text{ext}}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}, \hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})].$$

by Robins et al. in a 2008 technical report at Harvard University. In addition, they proposed

$$\hat{\mu}_{\text{WLS,ext2}} = \hat{\mu}(\hat{\pi}_{\text{ext}}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}, \hat{m}_{\text{WLS}}[\hat{\pi}_{\text{ext}}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}])$$

through a further iteration from $\hat{\mu}_{\text{WLS,ext}}$.

The targeted maximum likelihood approach of van der Laan & Rubin (2006, Sections 6·2–6·3) is closely related to the estimators $\hat{\mu}_{\text{OLS,ext}}$ and $\hat{\mu}_{\text{IPW,ext}}$. With $\hat{m}_{\text{OLS}}$ and $\hat{\pi}_{\text{ML}}$ as initial fitted values, this approach leads to the estimators

$$\hat{\mu}_{\text{TML}} = \hat{\mu}\{\hat{\pi}_{\text{ML}}, \hat{m}_{\text{TML}}(\hat{\pi}_{\text{ML}})\} = \tilde{E}\{\hat{m}_{\text{TML}}(X;\hat{\pi}_{\text{ML}})\}, \quad \hat{\mu}_{\text{TIPW}} = \hat{\mu}\{\hat{\pi}_{\text{TML}}(\hat{m}_{\text{OLS}}), 0\},$$
$$\hat{\mu}_{\text{TAIPW}} = \hat{\mu}\{\hat{\pi}_{\text{TML}}(\hat{m}_{\text{OLS}}), \hat{m}_{\text{TML}}(\hat{\pi}_{\text{ML}})\},$$

where $\hat{m}_{\text{TML}}(X;\hat{\pi})$ is obtained by fitting $E(Y \mid X) = m_{\text{ext}}(X;\kappa)$ with κ_2 fixed at $\hat{\alpha}_{\text{OLS}}$, and $\hat{\pi}_{\text{TML}}(X;\hat{m})$ is obtained by fitting $P(R=1 \mid X) = \pi_{\text{ext}}(X;\nu)$ with ν_2 fixed at $\hat{\gamma}_{\text{ML}}$. The estimators $\hat{\mu}_{\text{IPW,ext}}$ and $\hat{\mu}_{\text{TIPW}}$ are similar to the two likelihood estimators of Tan (2006). The first estimator accommodates the variation of $\hat{\gamma}_{\text{ML}}$ whereas the second ignores that variation.

2·4. *Comparison*

Consider the following criteria for evaluating estimators of μ. Note that improved local efficiency implies local efficiency, and sample boundedness implies population boundedness.

(a) Double robustness: $\hat{\mu}$ remains consistent if either model (1) or model (2) is correctly specified.
(b) Local efficiency: $\hat{\mu}$ attains the semiparametric variance bound, i.e., it is asymptotically equivalent to the first order to $\tilde{E}[RY/\pi(X) - \{R/\pi(X) - 1\}m(X)]$ if both model (1) and model (2) are correctly specified.
(c) Improved local efficiency: $\hat{\mu}$ is asymptotically at least as efficient as $\tilde{E}[RY/\pi(X) - \{R/\pi(X) - 1\}m(X;\alpha)]$ for arbitrary α if model (2) is correctly specified.
(d) Intrinsic efficiency: $\hat{\mu}$ attains the minimum asymptotic variance among the class of estimators $\tilde{E}[RY/\hat{\pi}_{\text{ML}}(X) - b_1^{\mathrm{T}}\{R/\hat{\pi}_{\text{ML}}(X) - 1\}\hat{v}(X)]$ for arbitrary b_1 if model (2) is correctly specified, where $\hat{v}(X) = \{1, \hat{m}(X)\}^{\mathrm{T}}$ and $\hat{m}(X)$ is the fitted value of $m(X)$ used in $\hat{\mu}$. Therefore, $\hat{\mu}$ is asymptotically at least as efficient as $\hat{\mu}_{\text{IPW}}$, $\hat{\mu}_{\text{IPW,ratio}}$, and $\hat{\mu}(\hat{\pi}_{\text{ML}}, \hat{m})$.
(e) Population boundedness: $\hat{\mu}$ lies within the range of all possible values of Y, if model (1) or model (2) or both are misspecified.
(f) Sample boundedness: $\hat{\mu}$ lies within the range of $\{Y_i : R_i = 1,\ i = 1, \ldots, n\}$, if model (1) or model (2) or both are misspecified.

The upper half of Table 1 presents a comparison of various estimators in Section 2·3 in terms of the foregoing criteria. See Sections 3–4 for a discussion of the likelihood and regression estimators in the lower half of Table 1.

Table 1. *Theoretical comparison of estimators*

	$\hat\mu_{\mathrm{TAIPW}}$ $\hat\mu_{\mathrm{AIPW}}$	$\hat\mu_{\mathrm{TML}}$ $\hat\mu_{\mathrm{OLS,ext}}$	$\hat\mu_{\mathrm{WLS}}$	$\hat\mu_{\mathrm{RV}}$	$\tilde\mu_{\mathrm{RV}}$	$\hat\mu_{\mathrm{IPW,ext}}$	$\hat\mu_{\mathrm{AIPW,ext}}$ $\hat\mu_{\mathrm{WLS,ext}}$	$\hat\mu_{\mathrm{WLS,ext2}}$
DR	✓	✓	✓	×	✓	×	✓	✓
LE	✓	✓	✓	✓	✓	✓	✓	✓
IE	×	×	×	×	×	✓	✓	✓
ILE	×	×	×	✓	✓	×	×	×
PB	×	✓	✓	×	×	×	×	✓
SB	×	×	×	×	×	×	×	×
		$\hat\mu_{\mathrm{LIK,OLS}}$	$\hat\mu_{\mathrm{REG,OLS}}$	$\tilde\mu_{\mathrm{REG,OLS}}$	$\tilde\mu_{\mathrm{LIK2,OLS}}$	$\tilde\mu_{\mathrm{LIK2,WLS}}$	$\tilde\mu_{\mathrm{LIK2,RV}}$	
DR		×	×	✓	✓	✓	✓	
LE		✓	✓	✓	✓	✓	✓	
IE		✓	✓	✓	✓	✓	✓	
ILE		×	×	×	×	×	✓	
PB		✓	×	×	✓	✓	✓	
SB		✓	×	×	✓	✓	✓	

DR, LE, IE, ILE, PB, and SB correspond to criteria (a)–(f).

3. Proposed approach

3·1. *Summary*

We extend the nonparametric likelihood approach of Tan (2006). The main contribution is to obtain an estimator of μ that is doubly robust, locally and intrinsically efficient, and sample-bounded simultaneously. Moreover, our approach is flexible enough to allow different choices, such as $\hat m_{\mathrm{OLS}}$, $\hat m_{\mathrm{WLS}}(\hat\pi_{\mathrm{ML}})$, and $\tilde m_{\mathrm{RV}}(\hat\pi_{\mathrm{ML}})$, for the fitted value $\hat m$. If $\hat m = \tilde m_{\mathrm{RV}}(\hat\pi_{\mathrm{ML}})$, then the resulting estimator is further improved-locally efficient.

3·2. *Non-doubly-robust likelihood estimator*

We describe the likelihood estimator of Tan (2006) in the current setup of missing data. The nonparametric likelihood of (X_i, R_i, R_iY_i), $i = 1, \ldots, n$, is

$$L_1 \times L_2 = \left[\prod_{i=1}^{n} \pi(X_i;\gamma)^{R_i}\{1-\pi(X_i;\gamma)\}^{1-R_i}\right] \times \left[\prod_{i:R_i=1} G_1(\{X_i, Y_i\}) \prod_{i:R_i=0} G_0(\{X_i\})\right],$$

where G_1 is the joint distribution of (X, Y) and G_0 is the marginal distribution of X. Maximizing L_1 leads to the maximum likelihood estimator $\hat\gamma_{\mathrm{ML}}$. Recall that $\hat m(x)$ is a fitted value of $m(x)$ based on model (1) and $\hat\upsilon(x) = \{1, \hat m(x)\}^{\mathrm T}$. Let $\hat h = (\hat h_1^{\mathrm T}, \hat h_2^{\mathrm T})^{\mathrm T}$ where

$$\hat h_1(x) = \{1 - \hat\pi_{\mathrm{ML}}(x)\}\,\hat\upsilon(x), \quad \hat h_2(x) = \frac{\partial\pi}{\partial\gamma}(x;\hat\gamma_{\mathrm{ML}}).$$

We choose to ignore the fact that G_0 and the marginal distribution of X under G_1 are identical, and retain only the constraints $\int \hat h(x)\,\mathrm{d}G_1 = \int \hat h(x)\,\mathrm{d}G_0$, i.e.,

$$\int \{1-\hat\pi(x)\}\,\mathrm{d}G_1 = \int \{1-\hat\pi(x)\}\,\mathrm{d}G_0,$$
$$\int \{1-\hat\pi(x)\}\hat m(x)\,\mathrm{d}G_1 = \int \{1-\hat\pi(x)\}\hat m(x)\,\mathrm{d}G_0,$$
$$\int \frac{\partial\pi}{\partial\gamma}(x;\hat\gamma_{\mathrm{ML}})\,\mathrm{d}G_1 = \int \frac{\partial\pi}{\partial\gamma}(x;\hat\gamma_{\mathrm{ML}})\,\mathrm{d}G_0.$$

See Kong et al. (2003) for a related formulation. The first two constraints respectively ensure that the resulting estimator of μ is consistent under correctly specified model (2) and locally efficient, whereas the third constraint accounts for the variation of $\hat{\gamma}_{\mathrm{ML}}$ such that the resulting estimator is intrinsically efficient. Furthermore, we require that G_1 be a probability measure supported on $\{(X_i, Y_i) : R_i = 1,\ i = 1, \ldots, n\}$ and hence $\int \mathrm{d}G_1 = 1$, and G_0 be a nonnegative measure (not necessarily a probability) supported on $\{X_i : R_i = 0,\ i = 1, \ldots, n\}$. Maximizing L_2 subject to these constraints leads to the estimators

$$\hat{G}_1(\{X_i, Y_i\}) = \frac{n^{-1}}{\omega(X_i; \hat{\lambda})} \qquad \text{if } R_i = 1,$$

$$\hat{G}_0(\{X_i\}) = \frac{n^{-1}}{1 - \omega(X_i; \hat{\lambda})} \quad \text{if } R_i = 0,$$

where $\omega(X; \lambda) = \hat{\pi}_{\mathrm{ML}}(X) + \lambda^{\mathrm{T}} \hat{h}(X)$, $\hat{\lambda} = \mathrm{argmax}_\lambda\, \ell(\lambda)$, and $\ell(\lambda) = \tilde{E}[R \log\{\omega(X; \lambda)\} + (1 - R) \log\{1 - \omega(X; \lambda)\}]$. The function $\ell(\lambda)$ is finite and concave on the set $\{\lambda : \omega(X_i; \lambda) > 0 \text{ if } R_i = 1 \text{ and } \omega(X_i; \lambda) < 1 \text{ if } R_i = 0,\ i = 1, \ldots, n\}$. Moreover, $\ell(\lambda)$ is strictly concave and bounded from above, and hence has a unique maximum, if and only if the set

$$\{\lambda : \lambda^{\mathrm{T}} \hat{h}(X_i) \geq 0 \text{ if } R_i = 1 \text{ and } \lambda^{\mathrm{T}} \hat{h}(X_i) \leq 0 \text{ if } R_i = 0,\ i = 1, \ldots, n\} \text{ is empty.} \tag{4}$$

See the Appendix for a proof. From our experience, $\hat{\lambda}$ can be computed effectively by using a globally convergent optimization algorithm such as the R package `trust`.

Setting the gradient of $\ell(\lambda)$ to 0 shows that $\hat{\lambda}$ is a solution to

$$0 = \tilde{E}\left[\frac{R - \omega(X; \lambda)}{\omega(X; \lambda)\{1 - \omega(X; \lambda)\}} \hat{h}(X)\right]. \tag{5}$$

By construction, $\hat{\lambda}$ also satisfies

$$1 = \int \mathrm{d}\hat{G}_1 = \tilde{E}\left\{\frac{R}{\omega(X; \hat{\lambda})}\right\}. \tag{6}$$

The resulting estimator of μ is

$$\hat{\mu}_{\mathrm{LIK}} = \int y \,\mathrm{d}\hat{G}_1 = \tilde{E}\left\{\frac{RY}{\omega(X; \hat{\lambda})}\right\}.$$

The estimator $\hat{\mu}_{\mathrm{LIK}}$ is structurally similar to $\hat{\mu}_{\mathrm{IPW,ext}}$ based on the extended model (3). The value $\hat{\lambda}$ can be interpreted as the maximum likelihood estimator of λ under the linear, extended propensity score model $P(R = 1 \mid X) = \omega(X; \lambda)$. However, there are important differences between $\hat{\mu}_{\mathrm{LIK}}$ and $\hat{\mu}_{\mathrm{IPW,ext}}$. First, $\omega(X_i; \hat{\lambda})$ may not lie between 0 and 1 for all $i = 1, \ldots, n$. It is only required that $\omega(X_i; \hat{\lambda}) > 0$ if $R_i = 1$ and $\omega(X_i; \hat{\lambda}) < 1$ if $R_i = 0$. Moreover, equation (6) automatically holds, whereas $\tilde{E}\{R/\hat{\pi}_{\mathrm{ext}}(X)\} = 1$ does not. By (6), $\omega(X_i; \hat{\lambda})$ with $R_i = 1$ are bounded from below by n^{-1}, and $\hat{\mu}_{\mathrm{LIK}}$ is sample-bounded. In contrast, $\hat{\pi}_{\mathrm{ext}}(X_i)$ with $R_i = 1$ may be arbitrarily close to 0, and $\hat{\mu}_{\mathrm{IPW,ext}}$ is not sample-bounded.

Tan (2006, Theorem 4) obtained an asymptotic expansion of $\hat{\mu}_{\mathrm{LIK}}$, assuming that model (2) is correctly specified. Here, we provide a general asymptotic expansion of $\hat{\mu}_{\mathrm{LIK}}$, allowing for misspecification of model (1) and model (2). See Manski (1988) for related asymptotic theory in misspecified models. Under regularity conditions, $\hat{\lambda}$ converges to a constant λ^* in probability with the expansion

$$\hat{\lambda} - \lambda^* = \hat{B}^{-1}\, \tilde{E}\left[\frac{R-\omega(X;\lambda^*)}{\omega(X;\lambda^*)\{1-\omega(X;\lambda^*)\}}\hat{h}(X)\right] + o_p(n^{-1/2}),$$

where

$$\hat{B} = \tilde{E}\left[\frac{\{R-\omega(X;\lambda^*)\}^2}{\omega^2(X;\lambda^*)\{1-\omega(X;\lambda^*)\}^2}\hat{h}(X)\hat{h}^{\mathrm{T}}(X)\right].$$

Moreover, a Taylor expansion of $\hat{\mu}_{\mathrm{LIK}}$ about λ^* yields

$$\hat{\mu}_{\mathrm{LIK}} = \tilde{E}\left\{\frac{RY}{\omega(X;\lambda^*)}\right\} - \hat{C}^{\mathrm{T}}\hat{B}^{-1}\,\tilde{E}\left[\frac{R-\omega(X;\lambda^*)}{\omega(X;\lambda^*)\{1-\omega(X;\lambda^*)\}}\hat{h}(X)\right] + o_p(n^{-1/2}), \quad (7)$$

where $\hat{C} = \tilde{E}[\{RY/\omega^2(X;\lambda^*)\}\hat{h}(X)]$. If model (2) is correctly specified, then $\lambda^* = 0$ and hence the expansion reduces to $\hat{\mu}_{\mathrm{LIK}} = \hat{\mu}_{\mathrm{REG}} + o_p(n^{-1/2})$ with $\hat{\mu}_{\mathrm{REG}} = \tilde{E}(\hat{\eta}) - \hat{\beta}^{\mathrm{T}}\tilde{E}(\hat{\xi})$, where $\hat{\eta} = RY/\hat{\pi}_{\mathrm{ML}}(X)$, $\hat{\xi} = [\{R/\hat{\pi}_{\mathrm{ML}}(X)-1\}\hat{\upsilon}^{\mathrm{T}}(X), \{R-\hat{\pi}_{\mathrm{ML}}(X)\}\hat{\varrho}_{\mathrm{ML}}(X)f^{\mathrm{T}}(X)]^{\mathrm{T}}$, $\hat{B} = \tilde{E}(\hat{\xi}\hat{\xi}^{\mathrm{T}})$, $\hat{C} = \tilde{E}(\hat{\xi}\hat{\eta})$, and $\hat{\beta} = \hat{B}^{-1}\hat{C}$ is the least-squares estimator in the linear regression of $\hat{\eta}$ on $\hat{\xi}$. The estimator $\hat{\mu}_{\mathrm{REG}}$ is locally and intrinsically efficient (Robins et al., 1995), but not doubly robust. See Section 4·5 for a further discussion.

3·3. *Doubly robust likelihood estimator*

The estimator $\hat{\mu}_{\mathrm{LIK}}$ is sample-bounded and locally and intrinsically efficient. If $\hat{m} = \hat{m}_{\mathrm{RV}}(\hat{\pi}_{\mathrm{ML}})$ or $\tilde{m}_{\mathrm{RV}}(\hat{\pi}_{\mathrm{ML}})$, then $\hat{\mu}_{\mathrm{LIK}}$ is further improved-locally efficient because it is asymptotically at least as efficient as $\hat{\mu}_{\mathrm{RV}}$ or $\tilde{\mu}_{\mathrm{RV}}$, which is improved-locally efficient. However, $\hat{\mu}_{\mathrm{LIK}}$ is not doubly robust. It may be inconsistent if model (1) is correctly specified but model (2) is misspecified. We propose a robustification of $\hat{\mu}_{\mathrm{LIK}}$ such that it satisfies double robustness in addition to sample boundedness and local and intrinsic efficiency.

We first discuss a simple version of our proposal. Consider the system of estimating equations

$$0 = \tilde{E}\left[\left\{\frac{R}{\omega(X;\lambda)} - 1\right\}\hat{\upsilon}(X)\right], \quad (8)$$

$$0 = \tilde{E}\left[\frac{R-\omega(X;\lambda)}{\omega(X;\lambda)\{1-\omega(X;\lambda)\}}\hat{h}_2(X)\right], \quad (9)$$

which are equivalent to (5) except that $(R-\omega)/\{\omega(1-\omega)\}$ is replaced by $(R/\omega - 1)/(1-\hat{\pi}_{\mathrm{ML}})$ in the equations associated with $\hat{h}_1 = (1-\hat{\pi}_{\mathrm{ML}})\hat{\upsilon}$. Let $\tilde{\lambda}$ be a solution to (8)–(9) subject to the constraint that $\omega(X_i;\lambda) > 0$ if $R_i = 1$ $(i = 1,\ldots,n)$ and let

$$\tilde{\mu}_{\mathrm{LIK}} = \tilde{E}\left\{\frac{RY}{\omega(X;\tilde{\lambda})}\right\}.$$

Note that $\hat{\upsilon}(X)$ includes the constant 1 and hence $\tilde{E}\{R/\omega(X;\tilde{\lambda})\} = 1$ by (8). Therefore, $\tilde{\mu}_{\mathrm{LIK}}$ is sample-bounded in a similar manner as $\hat{\mu}_{\mathrm{LIK}}$ is.

We derive asymptotic expansions for $\tilde{\lambda}$ and $\tilde{\mu}_{\mathrm{LIK}}$, allowing for misspecification of model (1) and model (2), in parallel to those for $\hat{\lambda}$ and $\hat{\mu}_{\mathrm{LIK}}$. Under regularity conditions, $\tilde{\lambda}$ converges to a constant $\lambda^\dagger$ in probability with the expansion

$$\tilde{\lambda} - \lambda^\dagger = \tilde{B}^{\mathrm{T}-1}\,\tilde{E}\left(\begin{bmatrix}\left\{\frac{R}{\omega(X;\lambda^\dagger)} - 1\right\}\hat{\upsilon}(X) \\ \frac{R-\omega(X;\lambda^\dagger)}{\omega(X;\lambda^\dagger)\{1-\omega(X;\lambda^\dagger)\}}\hat{h}_2(X)\end{bmatrix}\right) + o_p(n^{-1/2}),$$

where

$$\tilde{B} = \tilde{E}\left(\begin{bmatrix} \frac{R}{\omega^2(X;\lambda^\dagger)}\hat{h}_1(X)\hat{v}^{\mathrm{T}}(X) & \frac{\{R-\omega(X;\lambda^\dagger)\}^2}{\omega^2(X;\lambda^\dagger)\{1-\omega(X;\lambda^\dagger)\}^2}\hat{h}_1(X)\hat{h}_2^{\mathrm{T}}(X) \\ \frac{R}{\omega^2(X;\lambda^\dagger)}\hat{h}_2(X)\hat{v}^{\mathrm{T}}(X) & \frac{\{R-\omega(X;\lambda^\dagger)\}^2}{\omega^2(X;\lambda^\dagger)\{1-\omega(X;\lambda^\dagger)\}^2}\hat{h}_2(X)\hat{h}_2^{\mathrm{T}}(X) \end{bmatrix}\right).$$

Moreover, a Taylor expansion of $\tilde{\mu}_{\mathrm{LIK}}$ about $\lambda^\dagger$ yields

$$\tilde{\mu}_{\mathrm{LIK}} = \tilde{E}\left\{\frac{RY}{\omega(X;\lambda^\dagger)}\right\} - \hat{C}^{\mathrm{T}}\tilde{B}^{\mathrm{T}-1}\,\tilde{E}\left(\begin{bmatrix} \left\{\frac{R}{\omega(X;\lambda^\dagger)} - 1\right\}\hat{v}(X) \\ \frac{R-\omega(X;\lambda^\dagger)}{\omega(X;\lambda^\dagger)\{1-\omega(X;\lambda^\dagger)\}}\hat{h}_2(X) \end{bmatrix}\right) + o_p(n^{-1/2}). \quad (10)$$

If model (2) is correctly specified, then $\lambda^\dagger = 0$ and hence the expansion reduces to $\tilde{\mu}_{\mathrm{LIK}} = \tilde{\mu}_{\mathrm{REG}} + o_p(n^{-1/2})$ with $\tilde{\mu}_{\mathrm{REG}} = \tilde{E}(\hat{\eta}) - \tilde{\beta}^{\mathrm{T}}\tilde{E}(\hat{\xi})$, where $\hat{\zeta} = [R\hat{v}^{\mathrm{T}}(X)/\hat{\pi}_{\mathrm{ML}}(X), \{R - \hat{\pi}_{\mathrm{ML}}(X)\}\hat{\varrho}_{\mathrm{ML}}(X)f^{\mathrm{T}}(X)]^{\mathrm{T}}$, $\tilde{B} = \tilde{E}(\hat{\xi}\hat{\zeta}^{\mathrm{T}})$, and $\tilde{\beta} = \tilde{B}^{-1}\hat{C}$. In this case, $\hat{\mu}_{\mathrm{REG}}$ and $\tilde{\mu}_{\mathrm{REG}}$ are asymptotically equivalent to the first order and hence so are $\hat{\mu}_{\mathrm{LIK}}$ and $\tilde{\mu}_{\mathrm{LIK}}$. However, $\tilde{\mu}_{\mathrm{REG}}$ is akin to the doubly robust regression estimator of Tan (2006). These regression estimators, unlike $\hat{\mu}_{\mathrm{REG}}$, satisfies double robustness in addition to local and intrinsic efficiency.

The estimators $\hat{\mu}_{\mathrm{LIK}}$ and $\tilde{\mu}_{\mathrm{LIK}}$ are sample-bounded and locally and intrinsically efficient. However, $\tilde{\mu}_{\mathrm{LIK}}$, unlike $\hat{\mu}_{\mathrm{LIK}}$, is further doubly robust. This difference follows from the general asymptotic expansions (7) for $\hat{\mu}_{\mathrm{LIK}}$ and (10) for $\tilde{\mu}_{\mathrm{LIK}}$. The leading terms are structurally similar to respectively $\hat{\mu}_{\mathrm{REG}}$, which is not doubly robust, and $\tilde{\mu}_{\mathrm{REG}}$, which is doubly robust. Alternatively, $\tilde{\mu}_{\mathrm{LIK}}$ is doubly robust because

$$\tilde{E}\left\{\frac{R}{\omega(X;\tilde{\lambda})}\hat{m}(X)\right\} = \tilde{E}\{\hat{m}(X)\} \quad (11)$$

by (8) and hence $\tilde{\mu}_{\mathrm{LIK}}$ is identical to $\hat{\mu}\{\omega(\cdot;\tilde{\lambda}),\hat{m}\}$ in the typical form of doubly robust estimators. In contrast, $\tilde{E}\{R\hat{m}(X)/\omega(X;\hat{\lambda})\} = \tilde{E}\{\hat{m}(X)\}$ does not necessarily hold for $\hat{\mu}_{\mathrm{LIK}}$. We regard $\tilde{\lambda}$ as a calibration of the maximum likelihood estimator $\hat{\lambda}$ in the linear, extended propensity score model $P(R = 1 \mid X) = \omega(X;\lambda)$ such that equation (11) holds.

So far, we seem to fulfil the objective of deriving an estimator that is doubly robust, locally and intrinsically efficient, and sample-bounded. However, there remain subtle issues about the existence and computation of $\tilde{\lambda}$. First, it is difficult to characterize conditions under which there exists a solution to (8)–(9) subject to the constraint that $\omega(X_i;\lambda) > 0$ if $R_i = 1$ $(i = 1,\ldots,n)$. Moreover, algorithms for solving nonlinear equations such as (8)–(9) may fail to locate a solution, much less all possible solutions, if any exists. It presents a further challenge to accommodate the constraint on the domain of λ. Finally, if indeed there exists no solution or multiple solutions, it remains difficult to redefine $\tilde{\lambda}$ or select $\tilde{\lambda}$ among multiple solutions. These difficulties are applicable not only to (8)–(9), but to nonlinear estimating equations in general. See Small et al. (2000) for a survey that mainly deals with multiple solutions.

We now discuss a more effective version of our proposal to address the foregoing issues. Recall that $\hat{\lambda}$ is defined as a maximizer of $\ell(\lambda)$. Under condition (4), $\ell(\lambda)$ is strictly concave and bounded from above and hence $\hat{\lambda}$ exists and is unique. Consider the following two-step estimator.

(a) Compute $\hat{\lambda} = (\hat{\lambda}_1^{\mathrm{T}}, \hat{\lambda}_2^{\mathrm{T}})^{\mathrm{T}}$, partitioned according to $\hat{h} = (\hat{h}_1^{\mathrm{T}}, \hat{h}_2^{\mathrm{T}})^{\mathrm{T}}$.
(b) Compute $\tilde{\lambda}_{\mathrm{step2}} = (\tilde{\lambda}_{1,\mathrm{step2}}^{\mathrm{T}}, \hat{\lambda}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\tilde{\lambda}_{1,\mathrm{step2}} = \mathrm{argmax}_{\lambda_1}\ \kappa_1(\lambda_1)$ and

$$\kappa_1(\lambda_1) = \tilde{E}\left[R\frac{\log\{\omega(X;\lambda_1,\hat{\lambda}_2)\} - \log\{\omega(X;\hat{\lambda})\}}{1 - \hat{\pi}_{\mathrm{ML}}(X)} - \lambda_1^{\mathrm{T}}\hat{v}(X)\right].$$

The function $\kappa_1(\lambda_1)$ is finite and concave on the set $\{\lambda_1 : \omega(X_i;\lambda_1,\hat\lambda_2) > 0 \text{ if } R_i = 1,\, i = 1,\ldots,n\}$. Moreover, as shown in the Appendix, $\kappa_1(\lambda_1)$ is strictly concave and bounded from above, and hence has a unique maximum, if and only if the set

$$\{\lambda_1 : \lambda_1^{\mathrm{T}}\hat v(X_i) \geq 0 \text{ if } R_i = 1,\ i = 1,\ldots,n, \text{ and } \tilde E\{\lambda_1^{\mathrm{T}}\hat v(X)\} \leq 0\} \text{ is empty.} \tag{12}$$

Like $\hat\lambda$ in step (a), $\tilde\lambda_{1,\text{step2}}$ in step (b) can be computed effectively by using a globally convergent optimization algorithm such as the R package `trust`.

Setting the gradient of $\kappa_1(\lambda_1)$ to 0 shows that $\tilde\lambda_{1,\text{step2}}$ is a solution to

$$0 = \tilde E\left[\left\{\frac{R}{\omega(X;\lambda_1,\hat\lambda_2)} - 1\right\}\hat v(X)\right], \tag{13}$$

which is equivalent to (8) with λ_2 evaluated at $\hat\lambda_2$. In fact, we consider (13) as estimating equations and obtain $\kappa_1(\lambda_1)$ as an objective function by integrating the right side of (13). This construction is feasible because the matrix of the partial derivatives of the right side of (13) is symmetric and negative-semidefinite. In the degenerate case where $\hat h_2(X)$ is removed from $\hat h(X)$, then λ consists of λ_1 only and hence $\tilde\lambda$ and $\tilde\lambda_{\text{step2}}$ are identical.

The resulting estimator of μ is

$$\tilde\mu_{\text{LIK2}} = \tilde E\left\{\frac{RY}{\omega(X;\tilde\lambda_{\text{step2}})}\right\}.$$

The estimator $\tilde\mu_{\text{LIK2}}$, like $\tilde\mu_{\text{LIK}}$, is sample-bounded and doubly robust due to, respectively, $\tilde E\{R/\omega(X;\tilde\lambda_{\text{step2}})\} = 1$ and $E\{R\hat m(X)/\omega(X;\tilde\lambda_{\text{step2}})\} = E\{\hat m(X)\}$ by (13). Furthermore, $\tilde\mu_{\text{LIK2}}$ is asymptotically equivalent to the first order to $\hat\mu_{\text{LIK}}$ and $\tilde\mu_{\text{LIK}}$ if model (2) is correctly specified, and hence is locally and intrinsically efficient. See the Appendix for an asymptotic expansion of $\tilde\mu_{\text{LIK2}}$, allowing for misspecification of model (1) and model (2).

The foregoing development allows a general choice of the fitted value $\hat m(X)$. The estimator $\tilde\mu_{\text{LIK2}}$ is doubly robust, locally and intrinsically efficient, and sample-bounded. Nevertheless, different choices of $\hat m(X)$ lead to specific versions of $\tilde\mu_{\text{LIK2}}$ that differ beyond the four properties. Denote by $\tilde\mu_{\text{LIK2,OLS}}$, $\tilde\mu_{\text{LIK2,WLS}}$, and $\tilde\mu_{\text{LIK2,RV}}$ the versions of $\tilde\mu_{\text{LIK2}}$ corresponding to $\hat m = \hat m_{\text{OLS}}$, $\hat m_{\text{WLS}}(\hat\pi_{\text{ML}})$, and $\tilde m_{\text{RV}}(\hat\pi_{\text{ML}})$, and similarly denote those of $\hat\mu_{\text{LIK}}$, $\hat\mu_{\text{REG}}$, and $\tilde\mu_{\text{REG}}$. The estimator $\tilde\mu_{\text{LIK2,RV}}$, unlike $\tilde\mu_{\text{LIK2,OLS}}$ and $\tilde\mu_{\text{LIK2,WLS}}$, is further improved-locally efficient. See Table 1 for a comparison of these estimators among other estimators.

4. Extensions and comparisons

4·1. *Specification of* $\hat v(X)$

The vector $\hat v(X)$ is so far fixed as $\{1,\hat m(X)\}^{\mathrm{T}}$. However, it can be replaced throughout by a general vector of known functions of X including the constant 1 as in Tan (2006). With this extension, $\hat\mu_{\text{LIK}}$ and $\tilde\mu_{\text{LIK2}}$ still have asymptotic expansions in the current forms. The two estimators are sample-bounded and intrinsically efficient. Furthermore, if

$$\hat m(X) = b_1^{\mathrm{T}}\hat v(X) \text{ for some vector } b_1, \tag{14}$$

then $\hat\mu_{\text{LIK}}$ is locally efficient, and $\tilde\mu_{\text{LIK2}}$ is doubly robust and locally efficient. Condition (14) automatically holds for $\hat v(X) = \{1,\hat m(X)\}^{\mathrm{T}}$ with $b_1 = (0,1)^{\mathrm{T}}$.

Consider the case where model (1) is linear with identity link Ψ. Then $g(X)$ is an alternative choice of $\hat v(X)$ satisfying (14). For this choice, intrinsic efficiency implies improved local efficiency and hence $\hat{\mu}_{\text{LIK}}$ and $\tilde{\mu}_{\text{LIK2}}$ are improved-locally efficient. This result can also be seen from the following relationship. Suppose that $\hat{h}_2(X)$ is removed from $\hat{h}(X)$ throughout. Then $\hat{\mu}_{\text{REG}}$ and $\tilde{\mu}_{\text{REG}}$ are identical to $\hat{\mu}_{\text{RV}}$ and $\tilde{\mu}_{\text{RV}}$ respectively, which are improved-locally efficient (Tan, 2008). The estimators $\hat{\mu}_{\text{LIK}}$ and $\tilde{\mu}_{\text{LIK2}}$ have increased asymptotic variances, but are still asymptotically equivalent to the first order to $\hat{\mu}_{\text{REG}}$ and $\tilde{\mu}_{\text{REG}}$ if model (2) is correctly specified. Therefore, the original estimators $\hat{\mu}_{\text{LIK}}$ and $\tilde{\mu}_{\text{LIK2}}$ are improved-locally efficient.

4·2. *Estimation of $E(X)$ and G_1*

The estimators $\hat{\mu}_{\text{LIK}}$ and $\tilde{\mu}_{\text{LIK2}}$ for $\mu = E(Y)$ can be used for estimating $E(X)$ with Y replaced by X, and similarly for estimating the expectations of functions of X. The resulting estimators have similar properties to those of $\hat{\mu}_{\text{LIK}}$ and $\tilde{\mu}_{\text{LIK2}}$.

Suppose that X is contained in $\hat{v}(X)$ by specification. If model (2) is correctly specified, then $\tilde{E}\{RX/\omega(X;\hat{\lambda})\}$ is asymptotically at least as efficient as $\tilde{E}[RX/\hat{\pi}_{\text{ML}}(X) - \{R/\hat{\pi}_{\text{ML}}(X) - 1\}X] = \tilde{E}(X)$ by intrinsic efficiency, and hence asymptotically equivalent to the first order to $\tilde{E}(X)$. The estimator $\tilde{E}\{RX/\omega(X;\tilde{\lambda}_{\text{step2}})\}$, in contrast with $\tilde{E}\{RX/\omega(X;\hat{\lambda})\}$, is identical to $\tilde{E}(X)$ by (13), whether or not model (2) is correctly specified.

Estimation of $E(Y)$, $E(X)$, and the expectations of functions of (X, Y) is unified in estimation of G_1 from the distributional perspective of Tan (2006). Let $\tilde{G}_{1,\text{step2}}$ be the probability measure supported on $\{(X_i, Y_i) : R_i = 1,\ i = 1, \ldots, n\}$ such that if $R_i = 1$ then

$$\tilde{G}_{1,\text{step2}}(\{X_i, Y_i\}) = \frac{n^{-1}}{\omega(X_i;\tilde{\lambda}_{\text{step2}})}.$$

Then $\hat{G}_1$ and $\tilde{G}_{1,\text{step2}}$ are both estimators of G_1, supported on the completely observed data. However, $\tilde{G}_{1,\text{step2}}$ satisfies $\int \hat{v}(x)\,\mathrm{d}\tilde{G}_{1,\text{step2}} = \tilde{E}\{\hat{v}(X)\}$, i.e., the weighted average of $\hat{v}(X)$ under $\tilde{G}_{1,\text{step2}}$ is exactly matched to the overall sample average of $\hat{v}(X)$.

We compare our approach with the empirical likelihood approach of Qin & Zhang (2003). Their approach is to maximize $\prod_{i:R_i=1} G_1(\{X_i, Y_i\})$ subject to the constraints that G_1 is a probability measure supported on $\{(X_i, Y_i) : R_i = 1,\ i = 1, \ldots, n\}$ and $\int \hat{a}(x)\,\mathrm{d}G_1 = \tilde{E}\{\hat{a}(X)\}$, where $\hat{a}(x) = \{\hat{\pi}_{\text{ML}}(x), \hat{m}(x)\}^{\text{T}}$. The maximization leads to the estimator that if $R_i = 1$ then

$$\hat{G}_{\text{QZ}}(\{X_i, Y_i\}) = \frac{n_1^{-1}}{1 + \hat{\lambda}_{\text{QZ}}^{\text{T}}[\hat{a}(X_i) - \tilde{E}\{\hat{a}(X)\}]},$$

where $n_1 = \sum_{i=1}^n R_i$, $\hat{\lambda}_{\text{QZ}} = \text{argmax}_{\lambda_1}\ \ell_{\text{QZ}}(\lambda_1)$ and $\ell_{\text{QZ}}(\lambda_1) = \tilde{E}\{R\log(1 + \lambda_1^{\text{T}}[\hat{a}(X_i) - \tilde{E}\{\hat{a}(X)\}])\}$. The estimator $\hat{\mu}_{\text{QZ}} = \int y\,\mathrm{d}\hat{G}_{\text{QZ}}$ is sample-bounded due to $\int \mathrm{d}\hat{G}_{\text{QZ}} = 1$, and doubly robust and locally efficient due to $\int \hat{m}(x)\,\mathrm{d}\hat{G}_{\text{QZ}} = \tilde{E}\{\hat{m}(X)\}$. However, $\hat{\mu}_{\text{QZ}}$ is not intrinsically or improved-locally efficient, even in the special case where $\pi(X)$ is known and substituted for $\hat{\pi}_{\text{ML}}(X)$ and $\hat{m}_{\text{RV}}(\hat{\pi}_{\text{ML}})$ or $\tilde{m}_{\text{RV}}(\hat{\pi}_{\text{ML}})$ is used for $\hat{m}$.

4·3. *Augmentation of $\hat{\mu}_{LIK}$*

The estimator $\tilde{\mu}_{\text{LIK2}}$ is derived as a robustification of $\hat{\mu}_{\text{LIK}}$ to realize double robustness and retain sample boundedness and local and intrinsic efficiency. Our method is to calibrate the estimation of λ. An alternative method for robustification is to augment $\hat{\mu}_{\text{LIK}}$ with the additional term $\tilde{E}[\{R/\omega(X;\hat{\lambda}) - 1\}\hat{m}(X)]$, in a similar manner to augmenting $\hat{\mu}_{\text{IPW,ext}}$ to $\hat{\mu}_{\text{AIPW,ext}}$ by Robins et al. in their 2008 technical report. The resulting estimator is doubly robust and locally and intrinsically efficient, but not sample-bounded.

Recall that $\hat{\lambda} = \hat{\lambda}(\hat{m})$ depends on $\hat{m}$ and write $\hat{\omega}(X; \hat{m}) = \omega\{X; \hat{\lambda}(\hat{m})\}$. Substitution of $\hat{\omega}(\hat{m})$ for $\hat{\pi}_{\text{ext}}(\hat{m})$ in various estimators in Section 2·3 leads to

$$\hat{\mu}_{\text{AIPW,lik}} = \hat{\mu}\{\hat{\omega}(\hat{m}_{\text{OLS}}), \hat{m}_{\text{OLS}}\}, \quad \hat{\mu}_{\text{WLS,lik}} = \hat{\mu}[\hat{\omega}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}, \hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})],$$
$$\hat{\mu}_{\text{WLS,lik2}} = \hat{\mu}(\hat{\omega}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}, \hat{m}_{\text{WLS}}[\hat{\omega}\{\hat{m}_{\text{WLS}}(\hat{\pi}_{\text{ML}})\}]), \quad \tilde{\mu}_{\text{RV,lik}} = \hat{\mu}[\hat{\omega}\{\tilde{m}_{\text{RV}}(\hat{\pi}_{\text{ML}})\}, \tilde{m}_{\text{RV}}(\hat{\pi}_{\text{ML}})].$$

These estimators are similar to their counterparts in Section 2·3 in terms of the six properties in Table 1. The estimator $\hat{\mu}\{\hat{\omega}(\hat{m}), \hat{m}\}$ is not population-bounded or sample-bounded, whereas $\hat{\mu}_{\text{WLS,lik2}}$ is population-bounded. Nevertheless, $\hat{\mu}\{\hat{\omega}(\hat{m}), \hat{m}\}$ is bounded in the absolute value by $\Delta = \max\{|\ \hat{m}(X_i)\ | : i = 1, \ldots, n\} + \max\{|\ Y_i - \hat{m}(X_i)\ | : R_i = 1,\ i = 1, \ldots, n\}$, due to normalization (6). In contrast, $\hat{\mu}\{\hat{\pi}_{\text{ext}}(\hat{m}), \hat{m}\}$ may lie outside this range, because such a normalization does not hold for $\hat{\pi}_{\text{ext}}(X)$ as discussed in Section 3·2.

Kang & Schafer (2007) and Robins et al. (2007) considered a modification of $\hat{\mu}(\hat{\pi}, \hat{m})$ by deliberately normalizing the weights, that is,

$$\begin{aligned}\hat{\mu}_{\text{ratio}}(\hat{\pi}, \hat{m}) &= \tilde{E}^{-1}\left\{\frac{R}{\hat{\pi}(X)}\right\} \tilde{E}\left(\frac{RY}{\hat{\pi}(X)} - \frac{R}{\hat{\pi}(X)}[\hat{m}(X) - \tilde{E}\{\hat{m}(X)\}]\right)\\ &= \tilde{E}\{\hat{m}(X)\} + \tilde{E}^{-1}\left\{\frac{R}{\hat{\pi}(X)}\right\} \tilde{E}\left[\frac{R}{\hat{\pi}(X)}\{Y - \hat{m}(X)\}\right].\end{aligned}$$

The estimator $\hat{\mu}_{\text{ratio}}\{\hat{\pi}_{\text{ext}}(\hat{m}), \hat{m}\}$ is bounded in the absolute value by Δ. Moreover, it is similar to $\hat{\mu}\{\hat{\pi}_{\text{ext}}(\hat{m}), \hat{m}\}$ and $\hat{\mu}\{\hat{\omega}(\hat{m}), \hat{m}\}$ in terms of the six properties in Table 1. These estimators, two based on $\hat{\pi}_{\text{ext}}$ and one based on $\hat{\omega}$, are asymptotically equivalent to each other if model (2) is correctly specified, but may differ in various ways otherwise.

4·4. *Bounded robustification of $\hat{\mu}_{IPW,ext}$*

The estimator $\hat{\mu}_{\text{AIPW,ext}}$ is doubly robust but not sample-bounded. An alternative robustification of $\hat{\mu}_{\text{IPW,ext}}$ can be derived such that it is doubly robust and sample-bounded in a similar manner as $\tilde{\mu}_{\text{LIK2}}$ is derived from $\hat{\mu}_{\text{LIK}}$. Our method is to calibrate estimation of ν in the extended model (3). For simplicity, fix $\Pi(z) = \text{expit}(z)$, i.e., $\{1 + \exp(-z)\}^{-1}$. Then $\varrho(X; \gamma) \equiv 1$ free of γ, and $\pi_{\text{ext}}(X; \nu)$ reduces to $\Pi\{\nu_1^{\text{T}} \hat{v}(X)/\hat{\pi}_{\text{ML}}(X) + \nu_2^{\text{T}} f(X)\}$.

Recall that $\hat{\nu} = (\hat{\nu}_1^{\text{T}}, \hat{\nu}_2^{\text{T}})^{\text{T}}$ is the maximum likelihood estimator of ν and hence a solution to

$$\begin{aligned}0 &= \tilde{E}\big[\{R - \pi_{\text{ext}}(X; \nu)\}\, f(X)\big],\\ 0 &= \tilde{E}\left[\{R - \pi_{\text{ext}}(X; \nu)\}\, \frac{\hat{v}(X)}{\hat{\pi}_{\text{ML}}(X)}\right].\end{aligned} \tag{15}$$

Let $\tilde{\nu}_{\text{step2}} = (\tilde{\nu}_{1,\text{step2}}^{\text{T}}, \hat{\nu}_2^{\text{T}})^{\text{T}}$, $\tilde{\nu}_{1,\text{step2}} = \text{argmax}_{\nu_1}\ \mathcal{J}_1(\nu_1)$, and

$$\mathcal{J}_1(\nu_1) = \tilde{E}\left[-R\hat{\pi}_{\text{ML}}(X) \exp\left\{-\nu_1^{\text{T}} \frac{\hat{v}(X)}{\hat{\pi}_{\text{ML}}(X)} - \hat{\nu}_2^{\text{T}} f(X)\right\} - (1 - R)\nu_1^{\text{T}} \hat{v}(X)\right]$$

by integrating the right side of (17) below. The function $\mathcal{J}_1(\nu_1)$, unlike $\ell(\lambda)$ and $\kappa_1(\lambda_1)$, is finite and concave everywhere. Moreover, $\mathcal{J}_1(\nu_1)$ is strictly concave and bounded from above, and hence has a unique maximum, if and only if the set

$$\{\nu_1 : \nu_1^{\text{T}} \hat{v}(X_i) \geq 0 \text{ if } R_i = 1,\ i = 1, \ldots, n, \text{ and } \tilde{E}\{(1 - R)\nu_1^{\text{T}} \hat{v}(X)\} \leq 0\} \text{ is empty.} \tag{16}$$

See the Appendix for a proof. The existence condition (16) for $\tilde{\nu}_{1,\text{step2}}$ is more demanding than (12) for $\tilde{\lambda}_{1,\text{step2}}$ in that (16) implies (12), but not necessarily vice versa. Setting the gradient of

$\mathcal{J}_1(\nu_1)$ to 0 shows that $\tilde{\nu}_{1,\text{step2}}$ is a solution to

$$0 = \tilde{E}\left[\left\{\frac{R}{\pi_{\text{ext}}(X;\nu_1,\hat{\nu}_2)} - 1\right\}\hat{v}(X)\right], \tag{17}$$

which is equivalent to (15) with $(R - \pi_{\text{ext}})$ replaced by $(R/\pi_{\text{ext}} - 1)\hat{\pi}_{\text{ML}}$ and ν_2 evaluated at $\hat{\nu}_2$. The resulting estimator of μ is $\tilde{\mu}_{\text{IPW,ext2}} = \tilde{E}\{RY/\pi_{\text{ext}}(X;\tilde{\nu}_{\text{step2}})\}$. This estimator, like $\tilde{\mu}_{\text{LIK2}}$, is doubly robust, locally and intrinsically efficient, and sample-bounded.

We compare $\tilde{\mu}_{\text{IPW,ext2}}$ with the bounded, doubly robust estimator of Robins et al. (2007, Section 4·1·2). Consider the extended propensity score model $\pi_{\text{ext,RSLR}}(X;\chi,\gamma) = \Pi(\chi[\hat{m}(X) - \tilde{E}\{\hat{m}(X)\}] + \gamma^{\mathrm{T}} f(X))$. Let $\hat{\chi} = \hat{\chi}(\hat{m})$ be a solution to

$$0 = \tilde{E}\left(\frac{R}{\pi_{\text{ext,RSLR}}(X;\chi,\hat{\gamma}_{\text{ML}})}[\hat{m}(X) - \tilde{E}\{\hat{m}(X)\}]\right),$$

and write $\hat{\pi}_{\text{ext,RSLR}}(X;\hat{m}) = \pi_{\text{ext,RSLR}}\{X;\hat{\chi}(\hat{m}),\hat{\gamma}_{\text{ML}}\}$. The estimator $\hat{\mu}_{\text{IPW,ext,RLSR}} = \hat{\mu}_{\text{ratio}}\{\hat{\pi}_{\text{ext,RSLR}}(\hat{m}), 0\}$ is sample-bounded. Moreover, it is identical to $\hat{\mu}_{\text{ratio}}\{\hat{\pi}_{\text{ext,RSLR}}(\hat{m}), \hat{m}\}$ by the construction of $\hat{\chi}$ and hence is doubly robust and locally efficient. However, it is not intrinsically or improved-locally efficient, even in the case where $\hat{\gamma}_{\text{ML}}$ is replaced by the true value and $\hat{m}(X) - \tilde{E}\{\hat{m}(X)\}$ in $\pi_{\text{ext,RSLR}}(X;\chi,\gamma)$ is replaced by $[\hat{m}(X) - \tilde{E}\{\hat{m}(X)\}]/\pi(X)$.

4·5. *Regression estimators*

The estimators $\hat{\mu}_{\text{REG}}$ and $\tilde{\mu}_{\text{REG}}$ are called regression estimators (Tan, 2006, 2007), with connection to survey sampling (e.g., Cochran, 1977) and Monte Carlo integration (e.g., Hammersley & Handscomb, 1964). The idea is to exploit the fact that if model (2) is correctly specified, then $\hat{\eta}$ has mean μ and $\hat{\xi}$ has mean 0 asymptotically. The estimator $\hat{\mu}_{\text{REG}}$ attains the minimum asymptotic variance among the class of estimators $\tilde{E}(\hat{\eta}) - b^{\mathrm{T}}\tilde{E}(\hat{\xi})$ for arbitrary b. Moreover, $\tilde{\mu}_{\text{REG}}$ is asymptotically equivalent to the first order to $\hat{\mu}_{\text{REG}}$ because both $\tilde{\beta}$ and $\hat{\beta}$ converge $\beta = E^{-1}(\xi\xi^{\mathrm{T}})E(\xi\eta)$ in probability. Note that $\tilde{E}(\hat{\xi}_2) = 0$ and hence $\tilde{E}(\hat{\eta}) - b^{\mathrm{T}}\tilde{E}(\hat{\xi})$ reduces to $\tilde{E}(\hat{\eta}) - b_1^{\mathrm{T}}\tilde{E}(\hat{\xi}_1)$, where $b = (b_1^{\mathrm{T}}, b_2^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\xi} = (\hat{\xi}_1^{\mathrm{T}}, \hat{\xi}_2^{\mathrm{T}})^{\mathrm{T}}$ according to $\hat{h} = (\hat{h}_1^{\mathrm{T}}, \hat{h}_2^{\mathrm{T}})^{\mathrm{T}}$.

The estimators $\hat{\mu}_{\text{REG}}$ and $\tilde{\mu}_{\text{REG}}$ are no longer asymptotically equivalent if model (2) is misspecified. In fact, $\tilde{\mu}_{\text{REG}}$ is doubly robust whereas $\hat{\mu}_{\text{REG}}$ is not. The estimator $\tilde{\mu}_{\text{REG}}$ is akin to the doubly robust regression estimator of Tan (2006), in which $\hat{\eta}$ is defined as $\{R\hat{v}^{\mathrm{T}}(X)/\hat{\pi}_{\text{ML}}(X), R\hat{\varrho}_{\text{ML}}(X)f^{\mathrm{T}}(X)\}^{\mathrm{T}}$. A benefit of using this version of $\hat{\eta}$ is that the resulting matrix $\tilde{B}$ is symmetric and negative-semidefinite. Moreover, if $\{\lambda : \lambda^{\mathrm{T}}h(X_i) = 0 \text{ if } R_i = 1,\ i = 1,\ldots,n\}$ is empty, then $\tilde{B}$ is negative-definite. This symmetrization tends to stabilize the inversion of $\tilde{B}$ in $\tilde{\beta} = \tilde{B}^{-1}\hat{C}$ and hence improve the finite-sample behavior of $\tilde{\mu}_{\text{REG}}$.

A similar symmetrization can be applied to estimating equations (8)–(9). Consider the following estimating equations in place of (9)

$$0 = \tilde{E}\left[\left\{\frac{R}{\omega(X;\lambda)} - 1\right\}\frac{\hat{h}_2(X)}{1 - \hat{\pi}_{\text{ML}}(X)}\right]. \tag{18}$$

The matrix of the partial derivatives of the right sides of (8) and (18) is symmetric and negative-semidefinite. If $\{\lambda : \lambda^{\mathrm{T}}\hat{h}(X_i) = 0 \text{ if } R_i = 1,\ i = 1,\ldots,n\}$ is empty, then the matrix is negative-definite. In fact, (8) and (18) are jointly equivalent to setting to 0 the gradient of $\kappa(\lambda) = \tilde{E}([R\log\{\omega(X;\lambda)\} - \lambda^{\mathrm{T}}\hat{h}(X)]/\{1 - \hat{\pi}_{\text{ML}}(X)\})$, similarly as (13) is obtained from $\kappa_1(\lambda_1)$. The function $\kappa(\lambda)$ has similar properties of concavity and boundedness to those of $\kappa_1(\lambda_1)$. Therefore, it is numerically convenient to redefine $\tilde{\lambda}$ as a maximizer to $\kappa(\lambda)$ or equivalently a solution to

(8) and (18) subject to the constraint that $\omega(X_i;\lambda) > 0$ if $R_i = 1$ ($i = 1, \ldots, n$). The resulting estimator $\tilde{\mu}_{\text{LIK}}$ is comparable to $\tilde{\mu}_{\text{LIK2}}$ in terms of the six properties in Table 1.

A limitation of the modified estimator $\tilde{\mu}_{\text{LIK}}$ as compared with $\tilde{\mu}_{\text{LIK2}}$ is that it is difficult to generalize $\tilde{\mu}_{\text{LIK}}$ while retaining the structure of $\tilde{\lambda}$ to the setup of causal inference with non-binary, discrete treatments. See Section 5·4 for a further discussion.

5. Causal inference

5·1. *Setup*

We now turn to causal inference in the framework of potential outcomes (Neyman, 1923; Rubin, 1974). Let X be a vector of covariates and Y be an outcome as before. Let T be a treatment variable taking values in $\mathcal{T} = \{0, 1, \ldots, J-1\}$ with $J \geq 2$, where 0 denotes the null treatment or placebo. For each $t \in \mathcal{T}$, let Y_t be the potential outcome that would be observed under treatment t. We make the consistency assumption that $Y = Y_t$ if $T = t$, and the no-confounding assumption that for each $t \in \mathcal{T}$, R_t and Y_t are conditionally independent given X, where $R_t = 1\{T = t\}$. Throughout, $1\{\cdot\}$ denotes the indicator function.

The observed data consist of independent and identically distributed (X_i, T_i, Y_i), $i = 1, \ldots, n$. Our objective is to estimate the population mean $\mu_t = E(Y_t)$ for $t \in \mathcal{T}$. The difference $\mu_t - \mu_0$ is called the average causal effect of treatment t. To a certain extent, this problem can be handled as J separate problems of estimating μ_t from the data $(X_i, R_{t,i}, R_{t,i}Y_{t,i})$, $i = 1, \ldots, n$, as in Sections 2–4. However, the estimators of μ_t obtained in this way are not jointly intrinsically efficient and hence those of $\mu_t - \mu_0$ may be inefficient even marginally.

5·2. *Models and existing estimators*

Consider a parametric model for $m(t, X) = E(Y \mid T = t, X)$ in the form

$$E(Y \mid T = t, X) = m(t, X; \alpha) \quad (t \in \mathcal{T}), \tag{19}$$

where $m(t, x; \alpha)$ is a known function and α is a vector of unknown parameters. To focus on main ideas, assume that $m(t, X; \alpha) = \Psi\{\alpha_t^{\mathrm{T}} g(X)\}$, where α_t is a vector of unknown parameters and $\alpha = (\alpha_0^{\mathrm{T}}, \ldots, \alpha_{J-1}^{\mathrm{T}})^{\mathrm{T}}$. This specification of (19) is separable in the sense that $m(t, X; \alpha)$ depends on α only through α_t. By abuse of notation, treat $m(t, X; \alpha)$ as $m(t, X; \alpha_t)$. Let $\hat{\alpha}_{t,\text{OLS}}$ be a solution to $0 = \tilde{E}[R_t\{Y - m(t, X; \alpha_t)\}g(X)]$ and write $\hat{m}_{\text{OLS}}(t, X) = m(t, X; \hat{\alpha}_{t,\text{OLS}})$.

Consider a parametric model for $\pi(t, X) = P(T = t \mid X)$ in the form

$$P(T = t \mid X) = \pi(t, X; \gamma) \quad (t \in \mathcal{T}), \tag{20}$$

where $\pi(t, x; \gamma)$ is a known function and γ is a vector of unknown parameters. Let $\hat{\gamma}_{\text{ML}}$ be the maximum likelihood estimator of γ and write $\hat{\pi}_{\text{ML}}(t, X) = \pi(t, X; \hat{\gamma}_{\text{ML}})$. A convenient specification of (20) is the multinomial logit model

$$\pi(t, X; \gamma) = \frac{\exp\{\gamma_t^{\mathrm{T}} f(X)\}}{\sum_{j \in \mathcal{T}} \exp\{\gamma_j^{\mathrm{T}} f(X)\}}, \tag{21}$$

where $\gamma = (\gamma_0^{\mathrm{T}}, \gamma_1^{\mathrm{T}}, \ldots, \gamma_{J-1}^{\mathrm{T}})^{\mathrm{T}}$ with $\gamma_0 = 0$. In this case, the score equations for $\hat{\gamma}_{\text{ML}}$ are $0 = \tilde{E}[\{R_t - \pi(t, X; \gamma)\}f(X)]$ for $t = 1, \ldots, J-1$.

To estimate μ_t, the estimators in Section 2·3 can be adopted. Replace $\hat{\mu}(\hat{\pi}, \hat{m})$ by

$$\hat{\mu}_t(\hat{\pi}, \hat{m}) = \tilde{E}\left[\frac{R_t Y}{\hat{\pi}(t, X)} - \left\{\frac{R_t}{\hat{\pi}(t, X)} - 1\right\}\hat{m}(t, X)\right],$$

where $\hat{\pi}(t, X)$ and $\hat{m}(t, X)$ are estimators of $\pi(t, X)$ and $m(t, X)$ respectively. Various choices of the two estimators are available. The estimator $\hat{m}_{\text{OLS}}(t, X)$ is a simple choice of $\hat{m}(t, X)$, and $\hat{\pi}_{\text{ML}}(t, X)$ is a simple choice of $\hat{\pi}(t, X)$. Moreover, there are iterative choices of $\hat{m}(t, X)$ and $\hat{\pi}(t, X)$. Let $\hat{m}_{\text{ext}}(t, X; \hat{\pi}) = m_{\text{ext}}\{t, X; \hat{\kappa}_t(\hat{\pi})\}$, $\hat{m}_{\text{WLS}}(t, X; \hat{\pi}) = m\{t, X; \hat{\alpha}_{t,\text{WLS}}(\hat{\pi})\}$, and $\tilde{m}_{\text{RV}}(t, X; \hat{\pi}) = m\{t, X; \tilde{\alpha}_{t,\text{RV}}(\hat{\pi})\}$, where $\hat{\kappa}_t(\hat{\pi})$, $\hat{\alpha}_{t,\text{WLS}}(\hat{\pi})$, and $\tilde{\alpha}_{t,\text{RV}}(\hat{\pi})$ are obtained by substituting R_t, $\hat{\pi}(t, X)$, and $m(t, X; \alpha_t)$ for R, $\hat{\pi}(X)$, and $m(X; \alpha)$ throughout in $\hat{\kappa}(\hat{\pi})$, $\hat{\alpha}_{\text{WLS}}(\hat{\pi})$, and $\tilde{\alpha}_{\text{RV}}(\hat{\pi})$. Construction of an extension to $\hat{\pi}_{\text{ext}}(\hat{m})$ seems difficult for a general specification of model (20) with $J > 2$. Nevertheless, the task is straightforward if the multinomial logit specification (21) is used. Consider the model

$$P(T = t \mid X) = \pi_{\text{ext}}(t, X; \nu) = \frac{1}{C(X; \nu)} \exp\left\{ \sum_{j \in \mathcal{T}} \nu_{1t,j}^{\mathrm{T}} \frac{\hat{v}(j, X)}{\hat{\pi}_{\text{ML}}(j, X)} + \nu_{2t}^{\mathrm{T}} f(X) \right\}, \quad (22)$$

where $\nu = (\nu_1^{\mathrm{T}}, \nu_2^{\mathrm{T}})^{\mathrm{T}}$, ν_1 is the vector of $\nu_{1t,j}$ for $t, j \in \mathcal{T}$ with $\nu_{10,j} = 0$ for $j \in \mathcal{T}$ and $\nu_{1t,0} = \nu_{11,0}$ for $t \neq 0$, ν_2 is the vector of ν_{2t} for $t \in \mathcal{T}$ with $\nu_{20} = 0$, $\hat{v}(j, X) = \{1, \hat{m}(j, X)\}^{\mathrm{T}}$, and $C(X; \nu)$ is determined by $\sum_{t \in \mathcal{T}} \pi_{\text{ext}}(t, X; \nu) \equiv 1$. Let $\hat{\nu}(\hat{m})$ be the maximum likelihood estimator of ν and write $\hat{\pi}_{\text{ext}}(t, X; \hat{m}) = \pi_{\text{ext}}\{t, X; \hat{\nu}(\hat{m})\}$. The foregoing choices of $\hat{m}(t, X)$ and $\hat{\pi}(t, X)$ can be employed in similar combinations to those of $\hat{m}(X)$ and $\hat{\pi}(X)$ in Section 2·3. Label the resulting estimators of μ_t accordingly.

For each $t \in \mathcal{T}$, the marginal behavior of $\hat{\mu}_t$ can be evaluated by the criteria in Section 2·4. However, consider the following criteria for the joint behavior of $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{J-1})$. We say that a vector-valued estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if the asymptotic variance matrix of $\hat{\theta}_1$ is smaller than that of $\hat{\theta}_2$ in the order on positive-definite matrices.

(a) Joint double robustness: $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{J-1})$ remains consistent if either model (19) or model (20) is correctly specified.
(b) Joint local efficiency: $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{J-1})$ attains the semiparametric variance bound if both model (19) and model (20) are correctly specified.
(c) Joint improved local efficiency: $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{J-1})$ is at least as efficient as $\{\hat{\mu}_0(\alpha_0), \hat{\mu}_1(\alpha_1), \ldots, \hat{\mu}_{J-1}(\alpha_{J-1})\}$ if model (20) is correctly specified, where $\hat{\mu}_t(\alpha_t) = \tilde{E}[R_t Y/\pi(t, X) - \{R_t/\pi(t, X) - 1\} m(t, X; \alpha_t)]$ for α_t a vector of arbitrary constants ($t \in \mathcal{T}$).
(d) Joint intrinsic efficiency: $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{J-1})$ is at least as efficient as $\{\hat{\mu}_0(b_0), \hat{\mu}_1(b_1), \ldots, \hat{\mu}_{J-1}(b_{J-1})\}$ if model (20) is correctly specified, where $\hat{\mu}_t(b_t) = \tilde{E}[R_t Y/\hat{\pi}_{\text{ML}}(t, X) - b_t^{\mathrm{T}}\{R_t/\hat{\pi}_{\text{ML}}(t, X) - 1\}\hat{v}(t, X)]$ for b_t a vector of arbitrary constants ($t \in \mathcal{T}$).
(e) Joint population boundedness: $\hat{\mu}_t$ is population-bounded for each $t \in \mathcal{T}$.
(f) Joint sample boundedness: $\hat{\mu}_t$ is sample-bounded for each $t \in \mathcal{T}$.

Joint double robustness, local efficiency, or population or sample boundedness is equivalent to the fact that $\hat{\mu}_t$ satisfies the corresponding property for each $t \in \mathcal{T}$. However, joint intrinsic or improved local efficiency is respectively more stringent than the fact that for each $t \in \mathcal{T}$, $\hat{\mu}_t$ satisfies intrinsic or improved local efficiency.

The comparison in Table 1 remains applicable except for one correction, if the estimators are replaced by the joint estimators of $(\mu_0, \mu_1, \ldots, \mu_{J-1})$ and the properties are replaced by those on the joint behavior. See Sections 5·3–5·4 for a description of the likelihood and regression estimators. The correction is that none of the joint estimators satisfies joint improved local efficiency, although Table 1 is still valid regarding whether or not the estimators of μ_t satisfy improved local efficiency marginally. See Tan (2008, Section 3) for a further discussion.

Note that $(\hat\mu_{t,\mathrm{IPW,ext}})_{t\in\mathcal{T}}$ satisfies joint intrinsic efficiency because $\hat\upsilon(j,X)/\hat\pi_{\mathrm{ML}}(j,X)$, $j\in\mathcal{T}$, are simultaneously included as extra linear predictors for $\log\{\pi(t,X)/\pi(0,X)\}$ for each $t\neq 0$ in model (22). For fixed $j\neq 0$, if model (22) were specified such that $\log\{\pi(t,X)/\pi(0,X)\} = \nu_{2t}^{\mathrm{T}} f(X)$ if $t\neq 0$ or j, or $\nu_{1j,j}^{\mathrm{T}}\hat\upsilon(j,X)/\hat\pi_{\mathrm{ML}}(j,X)+\nu_{2j}^{\mathrm{T}} f(X)$ if $t=j$, then $\hat\mu_{j,\mathrm{IPW,ext}}$ would satisfy intrinsic efficiency marginally, but $(\hat\mu_{t,\mathrm{IPW,ext}})_{t\in\mathcal{T}}$ would not satisfy joint intrinsic efficiency. See Tan (2007, Section 3) for a related discussion.

5·3. *Non-doubly-robust likelihood estimator*

We present the likelihood estimator of Tan (2006) in the setup of causal inference, with the extension to accommodate discrete, binary or non-binary, treatments. See a 2007 Rutgers University technical report by Tan for a further extension to deal with marginal and nested structural models. The nonparametric likelihood of (X_i,T_i,Y_i), $i=1,\ldots,n$, is

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

where G_t is the joint distribution of (X,Y_t), $t\in\mathcal{T}$. Maximizing L_1 leads to the maximum likelihood estimator $\hat\gamma_{\mathrm{ML}}$. Recall that $\hat m(t,x)$ is an estimator of $m(t,x)$ based on model (19) and $\upsilon(t,x)=\{1,\hat m(t,x)\}^{\mathrm{T}}$. Let $\hat h=(\hat h_1^{\mathrm{T}},\hat h_2^{\mathrm{T}})^{\mathrm{T}}$ and $\hat h_1=(\hat h_{10}^{\mathrm{T}},\hat h_{11}^{\mathrm{T}},\ldots,\hat h_{1,J-1}^{\mathrm{T}})^{\mathrm{T}}$ where

$$\hat h_{1j}(t,x)=[1\{t=j\}-\hat\pi_{\mathrm{ML}}(t,x)]\hat\upsilon(j,x)\quad (j\in\mathcal{T}),\quad \hat h_2(t,x)=\frac{\partial\pi}{\partial\gamma}(t,x;\hat\gamma_{\mathrm{ML}}).$$

By construction, $\sum_{t\in\mathcal{T}}\hat h(t,x)\equiv 0$ because $\sum_{t\in\mathcal{T}}\hat\pi_{\mathrm{ML}}(t,x)\equiv 1$. We choose to ignore the fact that G_t, $t\in\mathcal{T}$, induce the same marginal distribution of X, and retain only the constraints $\sum_{t\in\mathcal{T}}\int\hat h(t,x)\,\mathrm{d}G_t=0$, i.e.,

$$0=\sum_{t\in\mathcal{T}}\int[1\{t=j\}-\hat\pi_{\mathrm{ML}}(t,x)]\,\mathrm{d}G_t\quad (j\in\mathcal{T}),$$

$$0=\sum_{t\in\mathcal{T}}\int[1\{t=j\}-\hat\pi_{\mathrm{ML}}(t,x)]\hat m(j,x)\,\mathrm{d}G_t\quad (j\in\mathcal{T}),$$

$$0=\sum_{t\in\mathcal{T}}\int\frac{\partial\pi}{\partial\gamma}(t,x;\hat\gamma_{\mathrm{ML}})\,\mathrm{d}G_t.$$

Furthermore, we require that G_t be a probability measure supported on $\{(X_i,Y_i):T_i=t,\ i=1,\ldots,n\}$ and hence $\int\mathrm{d}G_t=1$, $t\in\mathcal{T}$. Maximizing L_2 subject to these constraints leads to the estimators that if $T_i=t$ then

$$\hat G_t(\{X_i,Y_i\})=\frac{n^{-1}}{\omega(t,X_i;\hat\lambda)},$$

where $\omega(t,X;\lambda)=\hat\pi_{\mathrm{ML}}(t,X)+\lambda^{\mathrm{T}}\hat h(t,X)$, $\hat\lambda=\mathrm{argmax}_\lambda\,\ell(\lambda)$, and $\ell(\lambda)=\tilde E[\log\{\omega(T,X;\lambda)\}]$. The function $\ell(\lambda)$ is finite and concave on the set $\{\lambda:\omega(T_i,X_i;\lambda)>0,\ i=1\ldots,n\}$. Moreover, $\ell(\lambda)$ is strictly concave and bounded from above, and hence has a unique maximum, if and only if $\{\lambda:\lambda^{\mathrm{T}}\hat h(T_i,X_i)\geq 0,\ i=1\ldots,n\}$ is empty. This proposition follows in a similar manner as that concerning $\ell(\lambda)$ and condition (4) in Section 3·2.

The estimators $\hat G_t$, $t\in\mathcal{T}$, are similar to $\hat G_1$ in Section 3·2. If $J=2$, $\hat\pi_{\mathrm{ML}}(1,X)$ is identified as $\hat\pi_{\mathrm{ML}}(X)$, $\hat h_{10}$ is removed in $\hat h$, and the constraint $\int\mathrm{d}G_0=1$ is cancelled, then $\hat G_1$ reduces to exactly $\hat G_1$ in Section 3·2. For causal inference, $\hat G_t$, $t\in\mathcal{T}$, are equally of interest and constrained as probability measures. In contrast, only $\hat{G}_1$, but not $\hat{G}_0$, is of interest and constrained as a probability measure in the missing data setup.

Setting the gradient of $\ell(\lambda)$ to 0 shows that $\hat{\lambda}$ is a solution to

$$0 = \tilde{E}\left\{\frac{\hat{h}(T,X)}{\omega(T,X;\lambda)}\right\}, \tag{23}$$

or equivalently $0 = \sum_{t\in\mathcal{T}} \int \hat{h}(t,x)\,\mathrm{d}\hat{G}_t$. The resulting estimator of μ_t is

$$\hat{\mu}_{t,\mathrm{LIK}} = \int y_t\,\mathrm{d}\hat{G}_t = \tilde{E}\left\{\frac{R_t Y}{\omega(T,X;\hat{\lambda})}\right\}.$$

We derive the following asymptotic expansions for $\hat{\lambda}$ and $\hat{\mu}_{t,\mathrm{LIK}}$, allowing for misspecification of model (19) and model (20), similarly as in Section 3·2. Under regularity conditions, $\hat{\lambda}$ converges to a constant λ^* with the expansion $\hat{\lambda}-\lambda^* = \hat{B}^{-1}\tilde{E}\{\hat{h}(T,X)/\omega(T,X;\lambda^*)\} + o_p(n^{-1/2})$. Moreover, $\hat{\mu}_{t,\mathrm{LIK}}$ has the expansion

$$\hat{\mu}_{t,\mathrm{LIK}} = \tilde{E}\left\{\frac{R_t Y}{\omega(t,X;\lambda^*)}\right\} - \hat{C}_t^{\mathrm{T}}\hat{B}^{-1}\tilde{E}\left\{\frac{\hat{h}(T,X)}{\omega(T,X;\lambda^*)}\right\} + o_p(n^{-1/2}),$$

where $\hat{B} = \tilde{E}\{h(T,X)\hat{h}^{\mathrm{T}}(T,X)/\omega^2(T,X;\lambda^*)\}$ and $\hat{C}_t = \tilde{E}\{R_t Y/\omega^2(T,X;\lambda^*)\}$. If model (20) is correctly specified, then $\lambda^*=0$ and hence $\hat{\mu}_{t,\mathrm{LIK}}$ is asymptotically equivalent to the first order to $\hat{\mu}_{t,\mathrm{REG}} = \tilde{E}(\hat{\eta}_t) - \hat{C}_t^{\mathrm{T}}\hat{B}^{-1}\tilde{E}(\hat{\xi})$, where $\hat{\eta}_t = R_t Y/\hat{\pi}_{\mathrm{ML}}(T,X)$, $\hat{\xi} = \hat{h}(T,X)/\hat{\pi}_{\mathrm{ML}}(T,X)$, $\hat{B} = \tilde{E}(\hat{\xi}\hat{\xi}^{\mathrm{T}})$, and $\hat{C}_t = \tilde{E}(\hat{\xi}\hat{\eta}_t)$.

5·4. *Doubly robust likelihood estimator*

The estimator $\hat{\mu}_{t,\mathrm{LIK}}$ is sample-bounded and locally and intrinsically efficient marginally. Moreover, $(\hat{\mu}_{0,\mathrm{LIK}}, \hat{\mu}_{1,\mathrm{LIK}}, \ldots, \hat{\mu}_{J-1,\mathrm{LIK}})$ satisfies joint intrinsic efficiency. However, $\hat{\mu}_{t,\mathrm{LIK}}$ is not doubly robust. We propose a robustfication of $\hat{\mu}_{t,\mathrm{LIK}}$ such that the resulting estimator of μ_t satisfies double robustness in addition to sample boundedness and local and intrinsic efficiency, and the joint estimator satisfies joint intrinsic efficiency.

For our derivation, rewrite $\hat{h}(t,x)$ as

$$\hat{h}(t,x) = \hat{\hbar}(t,x) - \hat{\pi}_{\mathrm{ML}}(t,x)\sum_{j\in\mathcal{T}}\hat{\hbar}(j,x), \tag{24}$$

where $\hat{\hbar} = (\hat{\hbar}_1^{\mathrm{T}}, \hat{\hbar}_2^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\hbar}_2$ is defined the same as $\hat{h}_2$, but $\hat{\hbar}_1$ is defined as $\hat{h}_1$ with $\hat{h}_{1j}(t,x)$ replaced by $\hat{\hbar}_{1j}(t,x) = 1\{t=j\}\upsilon(j,x)$, $j\in\mathcal{T}$. Instead of (23), consider the system of estimating equations

$$0 = \tilde{E}\left\{\frac{\hat{\hbar}(T,X)}{\omega(T,X;\lambda)} - \sum_{t\in\mathcal{T}}\hat{\hbar}(t,X)\right\}, \tag{25}$$

i.e., $0 = \tilde{E}[\{R_t/\omega(T,X;\lambda)-1\}\hat{\upsilon}(t,X)]$, $t\in\mathcal{T}$, and $0 = \tilde{E}\{\hat{\hbar}_2(T,X)/\omega(T,X;\lambda)\}$. In retrospect, the vector of estimating functions $\hat{\hbar}(T,X)/\omega(T,X;\lambda) - \sum_{t\in\mathcal{T}}\hat{\hbar}(t,X)$ in (25) equals $\hat{h}(T,X)/\omega(T,X;\lambda)$ in (23) left-multiplied by the matrix $I - \sum_{t\in\mathcal{T}}\hat{\hbar}(T,X)\lambda^{\mathrm{T}}$, where I is the appropriate identity matrix. Let $\tilde{\lambda}$ be a solution to (25) subject to the constraint that $\omega(T_i,X_i;\lambda)>0$ $(i=1,\ldots,n)$ and let $\tilde{\mu}_{t,\mathrm{LIK}} = \tilde{E}\{R_t Y/\omega(T,X;\tilde{\lambda})\}$.

We derive the following asymptotic expansions for $\tilde\lambda$ and $\tilde\mu_{t,\text{LIK}}$, allowing for misspecification of model (19) and model (20), similarly as in Section 3·3. Under regularity conditions, $\tilde\lambda$ converges to a constant $\lambda^\dagger$ with the expansion $\tilde\lambda - \lambda^\dagger = \hat B^{\text{T}-1}\tilde E\{\hat\hbar(T,X)/\omega(T,X;\lambda^\dagger) - \sum_{t\in\mathcal{T}}\hat\hbar(T,X)\} + o_p(n^{-1/2})$. Moreover, $\tilde\mu_{t,\text{LIK}}$ has the expansion

$$\tilde\mu_{t,\text{LIK}} = \tilde E\left\{\frac{R_tY}{\omega(t,X;\lambda^\dagger)}\right\} - \hat C_t^{\text{T}}\tilde B^{\text{T}-1}\tilde E\left\{\frac{\hat\hbar(T,X)}{\omega(T,X;\lambda^\dagger)} - \sum_{t\in\mathcal{T}}\hat\hbar(T,X)\right\} + o_p(n^{-1/2}),$$

where $\tilde B = \tilde E\{h(T,X)\hat\hbar^{\text{T}}(T,X)/\omega^2(T,X;\lambda^\dagger)\}$. If model (20) is correctly specified, then $\lambda^\dagger = 0$ and hence $\tilde\mu_{t,\text{LIK}}$ is asymptotically equivalent to the first order to $\tilde\mu_{t,\text{REG}} = \tilde E(\hat\eta_t) - \hat C_t^{\text{T}}\tilde B^{\text{T}-1}\tilde E(\hat\xi)$, where $\hat\zeta = \hat\hbar(T,X)/\hat\pi_{\text{ML}}(T,X)$ and $\tilde B = \tilde E(\hat\xi\hat\zeta^{\text{T}})$. The estimators $\hat\mu_{t,\text{REG}}$ and $\tilde\mu_{t,\text{REG}}$ are similar to $\hat\mu_{\text{REG}}$ and $\tilde\mu_{\text{REG}}$ respectively. Both estimators are locally and intrinsically efficient, but $\tilde\mu_{t,\text{REG}}$ is doubly robust whereas $\hat\mu_{t,\text{REG}}$ is not.

The estimator $\tilde\mu_{t,\text{LIK}}$ is similar to $\tilde\mu_{\text{LIK}}$, satisfying double robustness, local and intrinsic efficiency, and sample boundedness but suffering from subtle limitations. As discussed in Section 3·3, it is difficult to study the existence of $\tilde\lambda$ in theory and to compute $\tilde\lambda$ effectively in practice. Alternatively, consider the following two-step estimator. Rewrite $\hat h_1$ as $(\hat h_{1t}^{\text{T}}, \hat h_{1(t)}^{\text{T}})^{\text{T}}$, where $\hat h_{1(t)}$ consists of the elements of $\hat h_1$ except $\hat h_{1t}$.

(a) Compute $\hat\lambda = (\hat\lambda_{1t}^{\text{T}}, \hat\lambda_{1(t)}^{\text{T}}, \hat\lambda_2^{\text{T}})^{\text{T}}$, partitioned according to $\hat h = (\hat h_{1t}^{\text{T}}, \hat h_{1(t)}^{\text{T}}, \hat h_2^{\text{T}})^{\text{T}}$.

(b) Compute $\tilde\lambda_{\text{step2}}^{(t)} = (\tilde\lambda_{1t,\text{step2}}^{\text{T}}, \hat\lambda_{1(t)}^{\text{T}}, \hat\lambda_2^{\text{T}})^{\text{T}}$, where $\tilde\lambda_{1t,\text{step2}} = \text{argmax}_{\lambda_{1t}}\ \kappa_1(\lambda_{1t})$ and

$$\kappa_1(\lambda_{1t}) = \tilde E\left[R_t\frac{\log\{\omega(t,X;\lambda_{1t},\hat\lambda_{1(t)},\hat\lambda_2)\} - \log\{\omega(t,X;\hat\lambda)\}}{1-\hat\pi_{\text{ML}}(t,X)} - \lambda_{1t}^{\text{T}}\hat v(t,X)\right].$$

The function $\kappa_1(\lambda_{1t})$ is finite and concave on the set $\{\lambda_{1t} : \omega(t,X_i;\lambda_{1t},\lambda_{1(t)},\hat\lambda_2) > 0 \text{ if } T_i = t,\ i = 1,\ldots,n\}$. Moreover, $\kappa_1(\lambda_{1t})$ is strictly concave and bounded from above, and hence has a unique maximum, if and only if $\{\lambda_{1t} : \lambda_{1t}^{\text{T}}\hat v(t,X_i) \geq 0 \text{ if } T_i = t,\ i = 1,\ldots,n, \text{ and } \tilde E\{\lambda_{1t}^{\text{T}}\hat v(t,X)\} \leq 0\}$ is empty. This proposition follows in a similar manner as that concerning $\kappa_1(\lambda_1)$ and condition (12) in Section 3·3.

Setting the gradient of $\kappa_1(\lambda_{1t})$ to 0 shows that $\tilde\lambda_{1t,\text{step2}}$ is a solution to

$$0 = \tilde E\left[\left\{\frac{R_t}{\omega(t,X;\lambda_{1t},\hat\lambda_{1(t)},\hat\lambda_2)} - 1\right\}\hat v(t,X)\right]. \tag{26}$$

The resulting estimator of μ_t is

$$\tilde\mu_{t,\text{LIK2}} = \tilde E\left\{\frac{R_tY}{\omega(T,X;\tilde\lambda_{\text{step2}}^{(t)})}\right\}.$$

The estimator $\tilde\mu_{t,\text{LIK2}}$, like $\tilde\mu_{\text{LIK2}}$, is sample-bounded and doubly robust due to equation (26). Furthermore, $\tilde\mu_{t,\text{LIK2}}$ is asymptotically equivalent to the first order to $\hat\mu_{t,\text{LIK}}$ and $\tilde\mu_{t,\text{LIK}}$ if model (20) is correctly specified. Therefore, $\tilde\mu_{t,\text{LIK2}}$ satisfies local and intrinsic efficiency and $(\tilde\mu_{0,\text{LIK2}}, \tilde\mu_{1,\text{LIK2}}, \ldots, \tilde\mu_{J-1,\text{LIK2}})$ satisfies joint intrinsic efficiency. The foregoing results are valid for a general choice of $\hat m(t,X)$. For the choice $\hat m(t,X) = \tilde m_{\text{RV}}(t,X)$, the resulting estimator $\tilde\mu_{t,\text{LIK2}}$ satisfies improved local efficiency marginally, although $(\tilde\mu_{0,\text{LIK2}}, \tilde\mu_{1,\text{LIK2}}, \ldots, \tilde\mu_{J-1,\text{LIK2}})$ does not satisfy joint improved local efficiency.

In the case of $J = 2$, we relate $\tilde{\mu}_{t,\mathrm{REG}}$ to the doubly robust regression estimator of Tan (2006) and then derive a robustification of $\hat{\mu}_{t,\mathrm{LIK}}$ such that $\tilde{\mu}_{t,\mathrm{LIK}}$ and the resulting estimator are similarly related. First, the regression estimator of μ_t in Tan (2006) is $\tilde{E}(\hat{\eta}_t) - \hat{C}_t^{\mathrm{T}} \tilde{B}_t^{\mathrm{T}-1} \tilde{E}(\hat{\xi})$, where $\tilde{B}_t = \tilde{E}(\hat{\xi}\hat{\zeta}_t^{\mathrm{T}})$, and $\hat{\zeta}_0$ or $\hat{\zeta}_1$ is defined as $\hat{\zeta}$ with $\hat{h}(t,x)$ replaced by $\hat{\hbar}^{(0)}(t,x) = -1\{t=0\}/\hat{\pi}_{\mathrm{ML}}(0,x)\hat{h}(0,x)$ or $\hat{\hbar}^{(1)}(t,x) = 1\{t=1\}/\{1-\hat{\pi}_{\mathrm{ML}}(1,x)\}\hat{h}(1,x)$. The functions $\hat{\hbar}^{(0)}$ and $\hat{\hbar}^{(1)}$, like $\hat{\hbar}$, are mapped to $\hat{h}$ by (24). A benefit of using $\hat{\zeta}_t$ is that $\tilde{B}_t$ is symmetric and negative-semidefinite. If $\{\lambda : \lambda^{\mathrm{T}}\hat{h}(t,X_i) = 0 \text{ if } T_i = t,\ i = 1,\ldots,n\}$ is empty, then $\tilde{B}_t$ is negative-definite. Second, substitution of $\hat{\hbar}^{(1)}$ for $\hat{\hbar}$ in (25) yields $0 = \tilde{E}[\{R_1/\omega(1,X;\lambda) - 1\}\hat{h}(1,X)/\{1-\hat{\pi}_{\mathrm{ML}}(1,X)\}]$, which is equivalent to setting to 0 the gradient of $\kappa^{(1)}(\lambda) = \tilde{E}([R_1 \log\{\omega(1,X;\lambda)\} - \lambda^{\mathrm{T}}\hat{h}(1,X)]/\{1-\hat{\pi}_{\mathrm{ML}}(1,X)\})$. This system of estimating equations is similar to (8) and (18) and $\kappa^{(1)}(\lambda)$ is similar to $\kappa(\lambda)$ in Section 4·5. Therefore, it is numerically convenient to redefine $\tilde{\lambda}$ as a maximizer to $\kappa^{(1)}(\lambda)$. The modified estimator $\tilde{\mu}_{1,\mathrm{LIK}}$ provides a one-step alternative to $\tilde{\mu}_{1,\mathrm{LIK2}}$, which involves two steps of maximization. Substitution of $\hat{\hbar}^{(0)}$ for $\hat{\hbar}$ in (25) leads to similar results. However, this modification of $\tilde{\mu}_{t,\mathrm{LIK}}$ is not feasible for $J > 2$. In general, there exists no function like $\hat{\hbar}^{(0)}$ and $\hat{\hbar}^{(1)}$ that is mapped to $\hat{h}$ by (24) and of the form $1\{t=j\}\phi(x)$ for fixed $j \in \mathcal{T}$ and $\phi(x)$ a vector of functions of x.

6. SIMULATION STUDY

To compare estimators, we conduct a simulation study with the same design as in Kang & Schafer (2007). Let $X = (X_1, X_2, X_3, X_4)^{\mathrm{T}}$, $Y = 210 + 27{\cdot}4X_1 + 13{\cdot}7X_2 + 13{\cdot}7X_3 + 13{\cdot}7X_4 + \epsilon$, and $T = 1\{U \leq \text{expit}(-X_1 + 0{\cdot}5X_2 - 0{\cdot}25X_3 - 0{\cdot}1X_4)\}$, where $(X_1, X_2, X_3, X_4, \epsilon, U)$ are mutually independent, $(X_1, X_2, X_3, X_4, \epsilon)$ are marginally normally distributed with mean 0 and variance 1, and U is uniformly distributed on $(0, 1)$. Let $W = (W_1, W_2, W_3, W_4)^{\mathrm{T}}$, $W_1 = \exp(0{\cdot}5X_1)$, $W_2 = X_2/\{1+\exp(X_1)\} + 10$, $W_3 = (0{\cdot}04X_1X_3 + 0{\cdot}6)^3$, and $W_4 = (X_2 + X_4 + 20)^2$. Consider the following models: (a) $E(Y \mid T=t, X) = \alpha_{0t} + \alpha_{1t}^{\mathrm{T}}X$ for $t = 0,1$; (b) $E(Y \mid T=t, X) = \alpha_{0t} + \alpha_{1t}^{\mathrm{T}}W$ for $t = 0,1$; (c) $P(T=1 \mid X) = \text{expit}(\gamma_0 + \gamma_1^{\mathrm{T}}X)$; (d) $P(T=1 \mid X) = \text{expit}(\gamma_0 + \gamma_1^{\mathrm{T}}W)$. Models (a) and (c) are correctly specified, whereas (b) and (d) are misspecified.

We first investigate 22 estimators of μ_1 in the missing data setup. The observed data consist of realizations of (X, T, TY). The 22 estimators are labelled as follows:

(1–3) $\hat{\mu}_{\mathrm{LIK,OLS}}$, $\hat{\mu}_{\mathrm{REG,OLS}}$, $\tilde{\mu}_{\mathrm{REG,OLS}}$ (Sections 3·2 and 4·5);
(4–7) $\hat{\mu}_{\mathrm{AIPW}}$ (ratio), $\hat{\mu}_{\mathrm{OLS,ext}}$, $\hat{\mu}_{\mathrm{WLS}}$, $\tilde{\mu}_{\mathrm{RV}}$ (Section 2·3);
(8–12) $\hat{\mu}_{\mathrm{IPW,ext}}$ (ratio), $\tilde{\mu}_{\mathrm{IPW,ext2}}$, $\hat{\mu}_{\mathrm{AIPW,ext}}$ (ratio), $\hat{\mu}_{\mathrm{WLS,ext}}$ (ratio), $\hat{\mu}_{\mathrm{WLS,ext2}}$ (Sections 2·3 and 4·4);
(13–15) $\hat{\mu}_{\mathrm{TIPW}}$ (ratio), $\hat{\mu}_{\mathrm{TML}}$, $\hat{\mu}_{\mathrm{TAIPW}}$ (ratio) (Section 2·3);
(16–22) $\hat{\mu}_{\mathrm{AIPW,lik}}$, $\tilde{\mu}_{\mathrm{LIK2,OLS}}$, $\hat{\mu}_{\mathrm{WLS,lik}}$, $\hat{\mu}_{\mathrm{WLS,lik2}}$, $\tilde{\mu}_{\mathrm{LIK2,WLS}}$, $\tilde{\mu}_{\mathrm{RV,lik}}$, $\tilde{\mu}_{\mathrm{LIK2,RV}}$ (Sections 3·3 and 4·3).

The estimator $\tilde{\mu}_{\mathrm{REG,OLS}}$ is taken as the doubly robust regression estimator of Tan (2006). The six estimators marked by ratio in brackets are defined in the form of $\hat{\mu}_{\mathrm{ratio}}(\hat{\pi}, \hat{m})$, instead of $\hat{\mu}(\hat{\pi}, \hat{m})$.

Figure 1 presents the boxplots of 13 estimators from 5000 Monte Carlo samples of size $n = 1000$. The realizations of each estimator are censored within the range of y-axis, and the number of realizations that lie outside the range are indicated next to the lower and upper limits of y-axis. The 13 estimators perform differently mainly in the cases where the propensity score model is correct but the outcome regression model is misspecified and where both models are misspecified. The upper half of Table 2 presents the ratios of mean squared errors of the 13 estimators again estimator 22 in these two cases for $n = 200$ and 1000.

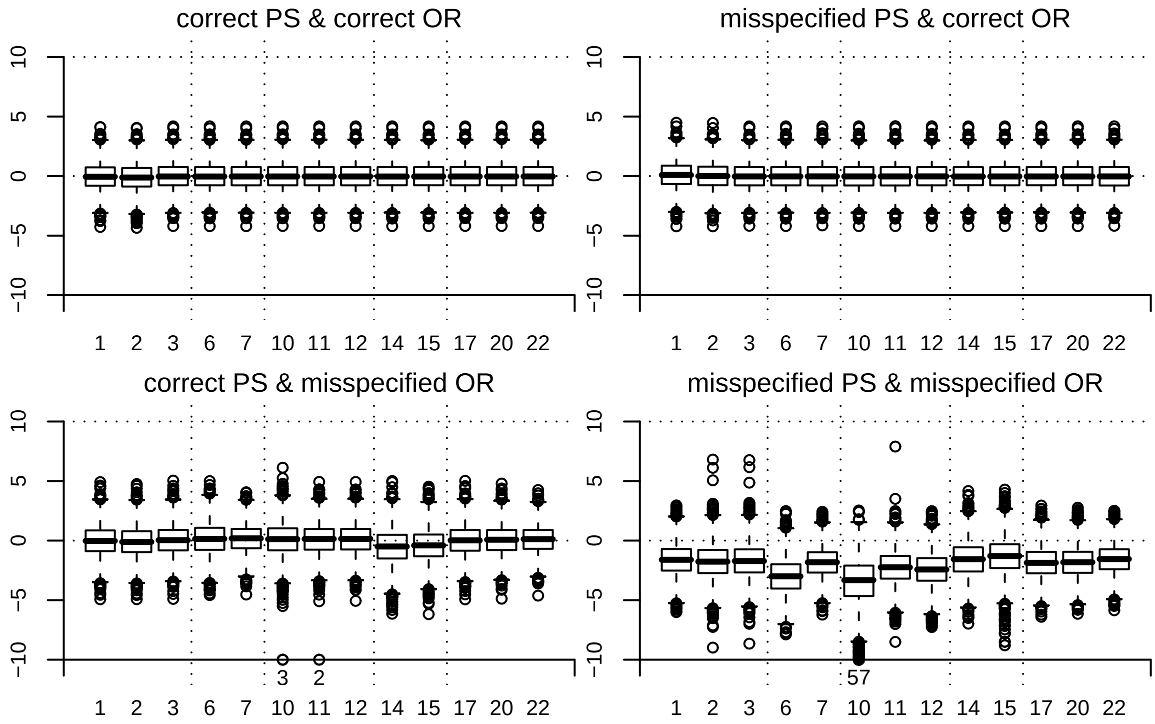

Fig. 1. Boxplots of estimators of $\mu_1 - 210$ ($n = 1000$)

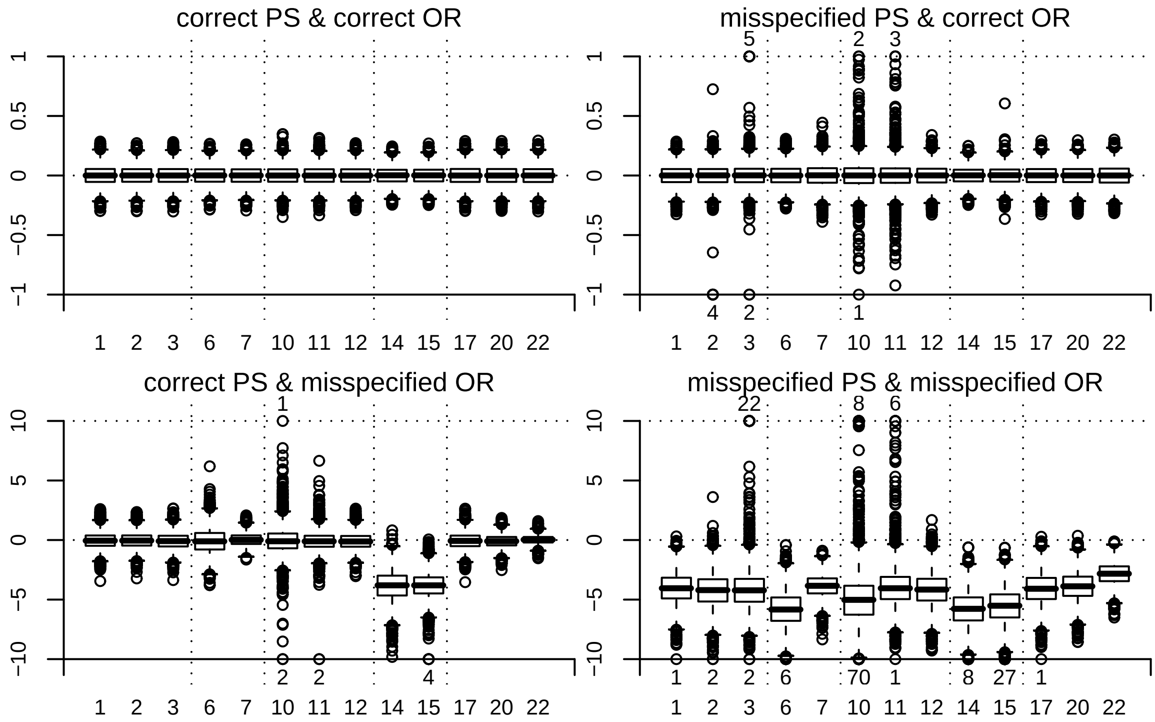

Fig. 2. Boxplots of estimators of $\mu_1 - \mu_0$ ($n = 1000$)

Table 2. *Ratios of mean squared errors*

Estimator	1	2	3	6	7	10	11	12	14	15	17	20	22
						Estimators of μ_1 in missing data setup							
C-PS&M-OR	1·23	1·28	1·20	1·32	1·07	1·39	1·25	1·24	1·51	1·35	1·24	1·17	1·00
	1·21	1·21	1·20	1·36	1·07	1·75	1·33	1·20	1·80	1·49	1·21	1·13	1·00
M-PS&M-OR	1·32	1·50	1·34	1·62	1·12	1·87	1·31	1·30	1·32	1·12	1·27	1·20	1·00
	1·10	1·31	1·27	2·82	1·24	4·86	1·78	1·99	1·21	1·04	1·31	1·27	1·00
						Estimators of $\mu_1 - \mu_0$ in causal inference setup							
C-PS&M-OR	1·97	1·79	2·76	4·50	1·85	5·00	3·27	3·16	19·8	18·1	2·65	1·87	1·00
	3·80	3·58	4·05	9·08	2·33	11·7	5·42	4·17	134	130	4·07	2·58	1·00
M-PS&M-OR	1·77	1·77	2·51	2·75	1·46	2·39	1·87	1·82	2·91	2·42	1·88	1·74	1·00
	2·02	2·22	73·3	3·99	1·77	3·42	2·08	2·16	3·97	3·72	2·07	1·87	1·00

C-PS (or M-PS): correct (or misspecified) propensity score model; M-OR: misspecified outcome regression model; Each cell gives the ratios of mean squared errors for $n = 200$ (upper) and $n = 1000$ (lower).

Among the estimators not shown, estimator 16 performs similarly as 17, estimators 18–19 similarly as 20, and estimator 21 similarly as 22. Estimators 4–5 perform overall poorly as in Kang & Schafer (2007, Tables 5 and 8). Estimator 8 yields outlying values in all the four cases whether the outcome regression model and the propensity score model are correct or misspecified. Estimators 9 and 13 improve upon estimator 8 when the propensity score model is correct, but still perform poorly when the propensity score model is misspecified.

The robustified likelihood estimators 16–22 provide the best performances for all the settings under study. Among these seven estimators, estimators 21–22 perform noticeably better than estimators 16–20 due to smaller variances when the propensity score model is correct but the outcome regression model is misspecified and due to smaller biases when both models are misspecified. The variance reduction in the first case reflects the result that estimators 21–22, but not estimators 16–20, are improved-locally efficient.

Estimators 1–3 have mean squared errors in the range of those of estimators 16–20 for all the settings. However, estimator 1 is not doubly robust and hence the fact that it is nearly unbiased when the outcome regression model is correct but the propensity score model is misspecified is not theoretically guaranteed. Estimators 2–3 yield outlying values when both models are misspecified, possibly because they are not bounded.

Estimators 6 and 10–12 have mean squared errors higher than the range of those of estimators 16–20 when the propensity score model is correct but the outcome regression model is misspecified and when both models are misspecified. The differences between estimator 16 and estimator 10 using $\hat{m}_{\text{OLS}}$ and between estimators 18–19 and estimators 6 and 11–12 using $\hat{m}_{\text{WLS}}$ indicate the advantage of using the extended propensity score $\hat{\omega}$ over $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{ext}}$.

Estimator 7 has mean squared errors slightly smaller than those of estimators 16–20 but still greater than those of estimators 21–22 when the propensity score model is correct but the outcome regression model is misspecified and when both models are misspecified. This comparison agrees with the facts that estimators 7 and 21–22 are improved-locally efficient using $\tilde{m}_{\text{RV}}$, but estimators 21–22 are further intrinsically efficient and bounded.

Estimators 14–15 improve upon related estimators 4–5, but still perform overall worse than estimators 16–22. Particularly, estimators 14–15 have considerable biases when the propensity score model is correctly specified but the outcome regression model is misspecified.

For the causal inference setup, Figure 2 presents the boxplots of 13 estimators of $\mu_1 - \mu_0$ for $n = 1000$ and the lower half of Table 2 presents the ratios of mean squared errors for $n = 200$ and 1000. The relative performances of the estimators are overall similar to those in the missing data setup. However, there are interesting new patterns. The reduction in mean squared errors by using estimators 21–22 over other estimators becomes more substantial than in the missing data setup when the propensity score model is correct but the outcome regression model is misspecified and when both models are misspecified. Estimators 2–3 yield an increased number of outlying values when the propensity score model is misspecified. Estimators 10–11 yield an increased number of outlying values except when both models are correct. Estimators 14–15 have increased biases when the outcome regression model is misspecified.

ACKNOWLEDGEMENT

The author thanks the Editor, an Associate Editor, and two referees for helpful comments. This research was supported by the U.S. National Science Foundation.

APPENDIX 1

Technical details

Condition (4) for $\ell(\lambda)$. The claim holds by the following results. Let $\Lambda_1 = \{\lambda : \lambda^{\mathrm{T}} \hat{h}(X_i) = 0,\, i = 1, \ldots, n\}$ and $\Lambda_2 = \{\lambda : \lambda^{\mathrm{T}} \hat{h}(X_i) \geq 0$ if $R_i = 1$ and $\lambda^{\mathrm{T}} \hat{h}(X_i) \leq 0$ if $R_i = 0,\, i = 1, \ldots, n\}$. First, if Λ_1 is empty, then $\ell(\lambda)$ is strictly concave. Otherwise, there exists some χ such that $\chi^{\mathrm{T}} (\partial^2 \ell / \partial\lambda\partial\lambda^{\mathrm{T}}) \chi = -\tilde{E}\{\omega^{-2} R (\chi^{\mathrm{T}} \hat{h})^2 + (1-\omega)^{-2} (1-R)(\chi^{\mathrm{T}} \hat{h})^2\} = 0$. Then $\chi^{\mathrm{T}} \hat{h}(X_i) = 0$ for $i = 1, \ldots, n$, a contradiction. Second, if Λ_2 is empty, then $\ell(\lambda)$ is bounded from above. Otherwise, there exists a sequence of pairs (c_k, χ_k), where $c_k > 0$ and χ_k is a unit vector, such that $\ell(c_k \chi_k) \to \infty$ as $k \to \infty$. Then $c_k \to \infty$. By compactness of the unit ball, there exists a unit vector χ_0 such that $\chi_k \to \chi_0$ as $k \to \infty$. For $i = 1, \ldots, n$ with $R_i = 1$, letting $k \to \infty$ in $\chi_k^{\mathrm{T}} \hat{h}(X_i) > -\hat{\pi}_{\mathrm{ML}}(X_i)/c_k$ yields $\chi_0^{\mathrm{T}} \hat{h}(X_i) \geq 0$. Similarly, for $i = 1, \ldots, n$ with $R_i = 0$, letting $k \to \infty$ in $\chi_k^{\mathrm{T}} \hat{h}(X_i) < \{1 - \hat{\pi}_{\mathrm{ML}}(X_i)\}/c_k$ yields $\chi_0^{\mathrm{T}} \hat{h}(X_i) \leq 0$. Third, if there exists some $\chi \in \Lambda_1$, then $\ell(\lambda + c\chi)$ is linear in c and hence $\ell(\lambda)$ is not strictly concave. If there exists some $\chi \in \Lambda_2$ but $\chi \notin \Lambda_1$, then $\ell(\lambda + c\chi) \to \infty$ as $c \to \infty$ and hence is unbounded.

Condition (12) for $\kappa_1(\lambda_1)$. The claim holds by the following results. Let $\Lambda_1 = \{\lambda_1 : \lambda_1^{\mathrm{T}} \hat{v}(X_i) = 0$ if $R_i = 1,\, i = 1, \ldots, n\}$ and $\Lambda_2 = \{\lambda_1 : \lambda_1^{\mathrm{T}} \hat{v}(X_i) \geq 0$ if $R_i = 1,\, i = 1, \ldots, n,$ and $\tilde{E}\{\lambda_1^{\mathrm{T}} \hat{v}(X)\} \leq 0\}$. First, if Λ_1 is empty, then $\kappa_1(\lambda_1)$ is strictly concave. Otherwise, there exists some χ such that $\chi^{\mathrm{T}} (\partial^2 \kappa_1 / \partial\lambda_1 \partial\lambda_1^{\mathrm{T}}) \chi = -\tilde{E}\{\omega^{-2} R (1 - \hat{\pi}_{\mathrm{ML}}) (\chi^{\mathrm{T}} \hat{v})^2\} = 0$. Then $\chi^{\mathrm{T}} \hat{v}(X_i) = 0$ for $i = 1, \ldots, n$ with $R_i = 1$, a contradiction. Second, if Λ_2 is empty, then $\kappa_1(\lambda_1)$ is bounded from above. Otherwise, there exists a sequence of pairs (c_k, χ_k), where $c_k > 0$ and χ_k is a unit vector, such that $\kappa_1(c_k \chi_k) \to \infty$ as $k \to \infty$. Then $c_k \to \infty$. By compactness of the unit ball, there exists a unit vector χ_0 such that $\chi_k \to \chi_0$ as $k \to \infty$. For $i = 1, \ldots, n$ with $R_i = 1$, letting $k \to \infty$ in $\chi_k^{\mathrm{T}} \hat{h}_1(X_i) > -\{\hat{\pi}_{\mathrm{ML}}(X_i) + \hat{\lambda}_2^{\mathrm{T}} \hat{h}_2(X_i)\}/c_k$ yields $\chi_0^{\mathrm{T}} \hat{v}(X_i) \geq 0$. Moreover, $\tilde{E}\{\chi_0^{\mathrm{T}} \hat{v}(X)\} \leq 0$. Otherwise $\kappa_1(c_k \chi_k) = c_k \tilde{E}(R[\log\{\omega(X; c_k\chi_k, \hat{\lambda}_2)\} - \log\{\omega(X; \hat{\lambda})\}]/(1 - \hat{\pi}_{\mathrm{ML}})/c_k - \chi_k^{\mathrm{T}} \hat{v}) \to -\infty$ as $k \to \infty$. Third, if there exists some $\chi \in \Lambda_1$, then $\kappa_1(\lambda_1 + c\chi)$ is linear in c and hence $\kappa_1(\lambda_1)$ is not strictly concave. If there exists some $\chi \in \Lambda_2$ but $\chi \notin \Lambda_1$, then $\kappa_1(\lambda_1 + c\chi) \to \infty$ as $c \to \infty$ and hence is unbounded.

Asymptotic expansion of $\tilde{\mu}_{LIK2}$. Let $\tilde{\chi} = \tilde{\lambda}_{1,\text{step2}} - \hat{\lambda}_1$. Then $\omega(X; \tilde{\lambda}_{\text{step2}}) = \omega(X; \hat{\lambda}_1 + \tilde{\chi}, \hat{\lambda}_2)$. Under regularity conditions, $\tilde{\chi}$ converges to a constant $\chi^\dagger$ in probability with the expansion $\tilde{\chi} - \chi^\dagger = \tilde{B}_1^{\mathrm{T}-1} \tilde{E}[\{R/\omega(X; \hat{\lambda}_1 + \chi^\dagger, \hat{\lambda}_2) - 1\} \hat{v}(X)] + o_p(n^{-1/2})$, where $\tilde{B}_1 = \tilde{E}[\{R/\omega^2(X; \hat{\lambda}_1 + \chi^\dagger, \hat{\lambda}_2)\} \hat{h}_1(X) \hat{v}^{\mathrm{T}}(X)]$. Moreover, a Taylor expansion of $\tilde{\mu}_{\mathrm{LIK2}}$ about $\chi^\dagger$ yields

$$\tilde{\mu}_{\mathrm{LIK2}} = \tilde{E}\left\{\frac{RY}{\omega(X; \hat{\lambda}_1 + \chi^\dagger, \hat{\lambda}_2)}\right\} - \hat{C}_1^{\mathrm{T}} \tilde{B}_1^{\mathrm{T}-1}\, \tilde{E}\left[\left\{\frac{R}{\omega(X; \hat{\lambda}_1 + \chi^\dagger, \hat{\lambda}_2)} - 1\right\} \hat{v}(X)\right] + o_p(n^{-1/2}),$$

where $\hat{C}_1 = \tilde{E}[\{RY/\omega^2(X; \hat{\lambda}_1 + \chi^\dagger, \hat{\lambda}_2)\} \hat{h}_1(X)]$. If model (2) is correctly specified, then $\chi^\dagger = 0$, and $\tilde{E}[\{R/\omega(X; \hat{\lambda}) - 1\} \hat{v}(X)] = o_p(n^{-1/2})$ by the discussion in Section 4·2. The foregoing expansion reduces to $\tilde{\mu}_{\mathrm{LIK2}} = \hat{\mu}_{\mathrm{LIK}} + o_p(n^{-1/2})$.

Condition (16) for $\mathcal{J}_1(\nu_1)$. The proof is similar to that for $\kappa_1(\nu_1)$ and condition (12). Let $\Lambda_1 = \{\nu_1 : \nu_1^{\mathrm{T}}\hat{v}(X_i) = 0$ if $R_i = 1,\ i = 1, \ldots, n\}$ and $\Lambda_2 = \{\nu_1 : \nu_1^{\mathrm{T}}\hat{v}(X_i) \geq 0$ if $R_i = 1,\ i = 1, \ldots, n,$ and $\tilde{E}\{(1-R)\nu_1^{\mathrm{T}}\hat{v}(X)\} \leq 0\}$. We only show that if Λ_2 is empty, then $\mathcal{J}_1(\nu_1)$ is bounded from above. Otherwise, there exists a sequence of pairs (c_k, χ_k), where $c_k > 0$ and χ_k is a unit vector, such that $\mathcal{J}_1(c_k\chi_k) \to \infty$ as $k \to \infty$. Then $c_k \to \infty$. By compactness of the unit ball, there exists a unit vector χ_0 such that $\chi_k \to \chi_0$ as $k \to \infty$. Then $\chi_0^{\mathrm{T}}\hat{v}(X_i) \geq 0$ for $i = 1, \ldots, n$ with $R_i = 1$. Otherwise $\chi_k^{\mathrm{T}}\hat{v}(X_i) < 0$ for all sufficiently large k and $\mathcal{J}_1(c_k\chi_k) \to -\infty$ as $k \to \infty$. Moreover, $\tilde{E}\{(1-R)\chi_0^{\mathrm{T}}\hat{v}(X)\} \leq 0$. Otherwise $\mathcal{J}_1(c_k\chi_k) \leq -c_k\tilde{E}\{(1-R)\chi_k^{\mathrm{T}}\hat{v}(X)\}$, which goes to $-\infty$ as $k \to \infty$.

REFERENCES

COCHRAN, W. G. (1977). *Sampling Techniques*. New York: Wiley, 3rd ed.

HAMMERSLEY, J. M. & HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. London: Methuen.

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

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. & TAN, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. R. Statist. Soc. B* **65**, 585–618.

MANSKI, C. F. (1988). *Analog Estimation Methods in Econometrics*. New York: Chapman & Hall.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statist. Sci.* **5**, 465–480.

QIN, J. & ZHANG, B. (2003). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc. B* **69**, 101–122.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc* **89**, 846–866.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc* **90**, 106–121.

ROBINS, J. M., SUED, M., LEI-GOMEZ, Q. & ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statist. Sci.* **22**, 544–559.

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

ROTNITZKY, A. & ROBINS, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–820.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

RUBIN, D. B. & VAN DER LAAN, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostat.* **4**, Article 5.

SCHARFSTEIN, D. O., ROTNIZKY, A. & ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *J. Amer. Statist. Assoc* **94**, 1096–1146.

SMALL, C. G., WANG, J. & YANG, Z. (2000). Eliminating multiple root problems in estimation (with discussion). *Statist. Sci.* **4**, 313–341.

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc* **101**, 1619–1637.

TAN, Z. (2007). Comment: Understanding OR, PS, and DR. *Statist. Sci.* **22**, 560–568.

TAN, Z. (2008). Comment: Improved local efficiency and double robustness. *Int. J. Biostat.* **4**, Article 10.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

VAN DER LAAN, M. J. & ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

VAN DER LAAN, M. J. & RUBIN, D. B. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2**, Article 11.

[*Received January* 2008. *Revised March* 2009]