A unified theory on empirical likelihood methods for missing data

Sixia Chen

Efficient estimation with missing data is an important practical problem with many application areas. Parameter estimation under nonresponse is considered when the parameter is defined as a solution to an estimating equation. Using a response probability model, a complete-response empirical likelihood method can be constructed and the non-parametric maximum likelihood estimator can be obtained by solving the weighted estimating equation where the weights are computed by maximizing the complete-response empirical likelihood subject to the constraints that incorporate the auxiliary information obtained from the full sample. Often the constraints are constructed from the working outcome regression model for the conditional distribution of the estimating function given the observation. The proposed method achieves the semi-parametric lower bound when we correctly specify the conditional expectation of the estimating function, regardless of whether the response probability is known or estimated. When the response probability is estimated nonparametrically, the resulting empirical likelihood method automatically achieves the semi-parametric lower bound without specifying the conditional distribution of the estimating function. Asymptotic theories are derived and simulation studies are also presented.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: Missing at random, Nonparametric estimation, Propensity score, Response mechanism.

1. INTRODUCTION

The empirical likelihood (EL) method, proposed by Owen (1988, 1990), has become a very powerful tool for nonparametric inference in statistics. It uses a likelihood-based approach without having to make a parametric distributional assumption about the data observation. Thus, the EL method often leads to efficient estimation and enables likelihood-ratio type inference. Qin and Lawless (1994) considered the situation when the parameter of interest is the solution to a system of estimating equations. Owen (2001) provide a comprehensive overview of the EL method.

In the case of missing data, however, the EL method is not directly applicable and some adjustment needs to be made. Qin (1993) addressed this problem using a biased sampling argument of Vardi (1985). Wang and Rao (2002) used regression-type imputation approaches to empirical likelihood inference. Wang and Chen (2009) used a nonparametric regression imputation approach to handle missing data in the empirical likelihood inference. The imputation approach uses some assumptions about the missing data given the observed data and usually assumes that the response mechanism is ignorable in the sense of Rubin (1976). Under an ignorable missing mechanism, the explicit modeling of the response model is avoided.

In this paper, we consider an alternative approach to handling missing data using a model for response probability. Use of a parametric response probability model in the empirical likelihood inference has been considered in Qin and Zhang (2007) and in Chen et al. (2008). Qin et al. (2009) and Tan (2011) considered using EL to model the complete likelihood, where the nonparametric likelihood function is computed for the whole sample including the units with missing data. The use of complete likelihood attains the full efficiency and also provides a nice theory of the limiting chi-squared distribution in the likelihood ratio test statistics. However, as a practical matter, the unit-level information for the complete likelihood is not always available and the complete likelihood cannot be computed. For example, in voluntary surveys, the individual values of auxiliary variables in the nonsampled part are not usually available. In this case, the approach of using the complete likelihood may not be applicable.

In the case in which the response mechanism is nonparametrically modeled, the literature is somewhat sparse. Cheng (1994) discussed some asymptotic properties of the mean estimator using the kernel regression method to estimate the conditional outcome regression model under an ignorable missing case. Recently, Kim and Yu (2011) extended the approach of Cheng (1994) to handle nonignorable nonresponse. Xue (2009) discussed an empirical likelihood method for linear models using the weights computed from a nonparametric model where the kernel regression method is used to estimate the response model. Da Silva and Opsomer (2009) considered another type of nonparametric response probability estimation using local polynomial regression. Hirano et al. (2003) and Cattaneo (2010) discussed semiparametric efficiency of the nonparametric response propensity estimators in the context of estimating average treatment effect in econometrics.

In this paper, we propose a unified approach of the EL method with missing data that avoids using the complete likelihood. Under the setup of estimating function in Qin and Lawless (1994), the proposed method can handle the situation regardless of whether the response probabilities are known or estimated, parametrically or even nonparametrically. When the response probabilities are estimated from a parametric model, the proposed method under the ignorable response mechanism is similar to the method of Qin and Zhang (2007). The proposed method is directly applicable to the problem of the propensity score weighting method. The propensity score weighting method can be found, for example, in Durrant and Skinner (2006), Kim and Kim (2007), and Chang and Kott (2008). We show that employing EL method using a suitable choice of control variables leads to efficient estimation in the sense that it achieves the lower bound of the asymptotic variance. Optimal choice of the control variable requires correct specification of the conditional distribution of the missing data given the observation. Under the nonparametric propensity score method, which will be discussed in Section 5, the lower bound of the asymptotic variance can be achieved without correctly specifying the conditional distribution.

In Section 2, we first review the existing methods of empirical likelihood under missing data and discuss a unified approach of the EL method. Asymptotic properties of the proposed estimator under known response probabilities are discussed in Section 3. The proposed EL estimator is discussed under estimated response probability in Section 4. Use of the nonparametric response model for the EL approach is discussed in Section 5. Results from the simulation study are reported in Section 6. Concluding remarks are made in Section 7.

2. BASIC SETUP

Consider a multivariate random variable $(\mathbf{X}, Y)$ with distribution function $F(\mathbf{x}, y)$ that is completely unspecified except that $\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\theta}_0)\} = 0$ for some $\boldsymbol{\theta}_0$. We are interested in estimating the parameter $\boldsymbol{\theta}_0$ from a random sample of the distribution. To avoid unnecessary details, we assume that the solution to $\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\theta})\} = 0$ is unique. For simplicity, we assume that the dimension of $\mathbf{U}$ is equal to the dimension of $\boldsymbol{\theta}$.

If $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$, are n independent realizations of the random variable $(\mathbf{X}, Y)$, a consistent estimator of $\boldsymbol{\theta}_0$ can be obtained by solving

$$(1) \qquad \sum_{i=1}^{n} \mathbf{U}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = 0.$$

In this paper, we consider the problem of estimating $\boldsymbol{\theta}_0$ when $\mathbf{x}$ is always observed and y is subject to missingness. Let $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. We consider an approach based on the empirical likelihood (EL) method.

To explain the idea, first note that the joint density of the observed data can be written as

$$(2) \quad p^{n_r}(1-p)^{n-n_r} \times \prod_{r_i=1} f(\mathbf{x}_i, y_i | r_i = 1) \prod_{r_i=0} f(\mathbf{x}_i | r_i = 0),$$

where n_r is the response sample size, $p = \mathrm{Pr}(r = 1)$, $f(\mathbf{x}, y | r)$ is the conditional density of $(\mathbf{X}, Y)$ given r, and $f(\mathbf{x}_i | r_i = 0) = \int f(\mathbf{x}_i, y_i | r_i = 0) dy_i$ is the marginal density of $\mathbf{X}$ among $r = 0$.

In the empirical likelihood approach, the distribution is assumed to have the support on the sample observation. Let $F_1(\mathbf{x}, y) = \mathrm{Pr}(\mathbf{X} \leq \mathbf{x}, Y \leq y | r = 1)$ and $F_0(\mathbf{x}, y) = \mathrm{Pr}(\mathbf{X} \leq \mathbf{x}, Y \leq y | r = 0)$. Under the empirical likelihood approach, we can express

$$(3) \qquad F_1(\mathbf{x}, y) = \sum_{r_i=1} \omega_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y),$$

where $\sum_{r_i=1} \omega_i = 1$, ω_i is the point mass assigned to $(\mathbf{x}_i, y_i)$ in the nonparametric distribution of $F_1(\mathbf{x}, y)$, and $I(B)$ is an indicator function for event B. To express $F_0(\mathbf{x}, y)$ using ω_i, note that we can write

$$f(\mathbf{x}_i, y_i | r_i = 0) = f(\mathbf{x}_i, y_i | r_i = 1) \times \frac{Odd(\mathbf{x}_i, y_i)}{\mathrm{E}\{Odd(\mathbf{x}_i, y_i) | r_i = 1\}},$$

where

$$Odd(\mathbf{x}, y) = \frac{\mathrm{Pr}(r = 0 \mid \mathbf{x}, y)}{\mathrm{Pr}(r = 1 \mid \mathbf{x}, y)}.$$

Thus, we can express $F_0(\mathbf{x}, y) = \mathrm{Pr}(\mathbf{X} \leq \mathbf{x}, Y \leq y | r = 0)$ by

$$(4) \qquad F_0(\mathbf{x}, y) = \frac{\sum_{r_i=1} \omega_i O_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y)}{\sum_{r_i=1} \omega_i O_i},$$

where $O_i = Odd(\mathbf{x}_i, y_i)$. Note that $F_0(\mathbf{x}, y)$ is completely determined by two factors: ω_i and O_i. The factor ω_i is determined by the distribution $F_1(\mathbf{x}, y)$ and the factor O_i is determined by the response mechanism. If $Odd(\mathbf{x}, y)$ is a known function of $(\mathbf{x}, y)$, then we have only to determine ω_i.

From (4), the joint distribution of $(\mathbf{x}, y)$ can be written as

$$F_w(\mathbf{x}, y) = p \times \sum_{r_i=1} \omega_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y)$$

$$+ (1-p) \times \left\{ \frac{\sum_{r_i=1} \omega_i O_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y)}{\sum_{r_i=1} \omega_i O_i} \right\}$$

$$= p \times \left\{ \sum_{r_i=1} \omega_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y) \right.$$

$$\left. + (1/p - 1) \frac{\sum_{r_i=1} \omega_i O_i I(\mathbf{x}_i \leq \mathbf{x}, y_i \leq y)}{\sum_{r_i=1} \omega_i O_i} \right\}.$$

Note that (3) implies

$$\sum_{r_i=1} \omega_i(O_i+1) = E\left\{\frac{1}{\pi(\mathbf{X},Y)}\Big|r=1\right\}$$

$$= \int \frac{1}{\pi(\mathbf{x},y)} f(\mathbf{x},y|r=1)dxdy$$

$$= \int \frac{1}{\pi(\mathbf{x},y)}\frac{\pi(\mathbf{x},y)f(\mathbf{x},y)}{p}d\mathbf{x}dy = 1/p.$$

Thus, we have $\sum_{r_i=1}\omega_i O_i = 1/p - 1$ and

$$F_w(\mathbf{x},y) = \frac{\sum_{r_i=1}\omega_i(1+O_i)I(\mathbf{x}_i\leq\mathbf{x},y_i\leq y)}{\sum_{r_i=1}\omega_i(O_i+1)}.$$

We propose maximizing the pseudo-likelihood $\prod_{r_i=1}f(\mathbf{x}_i,y_i|r_i=1)$ in (2) in constructing the empirical likelihood. The proposed empirical likelihood approach can be formulated as maximizing

$$(5) \qquad l_e(\theta) = \sum_{r_i=1}\log(\omega_i),$$

subject to

$$(6) \qquad \sum_{r_i=1}\omega_i = 1, \quad \sum_{r_i=1}\omega_i(1+O_i)\mathbf{U}(\mathbf{x}_i,y_i;\boldsymbol{\theta}) = 0.$$

Note that, in constraint (6), the observed values of $\mathbf{x}_i$ with $r_i = 0$ are not used. To incorporate the partial information, we can impose

$$(7) \qquad \frac{\sum_{r_i=1}\omega_i(1+O_i)\mathbf{h}(\mathbf{x}_i;\boldsymbol{\theta})}{\sum_{r_i=1}\omega_i(1+O_i)} = n^{-1}\sum_{i=1}^{n}\mathbf{h}(\mathbf{x}_i;\boldsymbol{\theta})$$

as an additional constraint for some $\mathbf{h}(\mathbf{x}_i;\boldsymbol{\theta})$. The choice of $\mathbf{h}(\mathbf{x};\boldsymbol{\theta})$ will be discussed later.

There are several other approaches using the empirical likelihood with missing data. Qin et al. (2002) considered using empirical likelihood for nonignorable nonresponse. Wang and Rao (2002) proposed empirical likelihood-based inference under imputation for missing response data. Qin and Zhang (2007) proposed an empirical likelihood method for estimating the mean response under ignorable missing data where the response probability $\pi_i = \Pr(r_i = 1|\mathbf{X}_i)$ is parametrically modeled by $\pi_i = \pi_i(\boldsymbol{\phi}_0)$ for some $\boldsymbol{\phi}_0$. Specifically, they proposed maximizing

$$l = \sum_{r_i=1}\log\left\{\pi_i(\hat{\boldsymbol{\phi}})p_i/\hat{\nu}\right\},$$

subject to
(8)

$$\sum_{r_i=1}p_i = 1, \ \sum_{r_i=1}p_i\pi_i(\hat{\boldsymbol{\phi}}) = \hat{\nu}, \ \sum_{r_i=1}p_i\mathbf{h}(\mathbf{x}_i) = n^{-1}\sum_{i=1}^{n}\mathbf{h}(\mathbf{x}_i),$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimator of $\boldsymbol{\phi}_0$ in the response probability, $\mathbf{h}(\mathbf{x}_i)$ is an arbitrary variable, and $\hat{\nu} = n^{-1}\sum_{i=1}^{n}\pi_i(\hat{\boldsymbol{\phi}})$. Once the estimated probability $\hat{p}_i$ is computed by the above maximization procedure, the population mean can be estimated by $\hat{\boldsymbol{\theta}} = \sum_{r_i=1}\hat{p}_iy_i$. Chen et al. (2008) built two empirical likelihoods for response and nonresponse variables separately and formulated two estimating equations based on these two empirical likelihoods. In the context of the current setup, their proposed method can be described as maximizing $l = \sum_{r_i=1}\log(p_i) + \sum_{r_j=0}\log(q_j)$, subject to $\sum_{r_i=1}p_i = 1, p_i \geq 0, \sum_{r_j=0}q_j = 1, q_j \geq 0$, and

$$(9) \quad \sum_{r_i=1}p_i\frac{\mathbf{h}(\mathbf{x}_i;\boldsymbol{\theta})-\boldsymbol{\mu}}{\pi_i(\hat{\boldsymbol{\phi}})} = 0, \quad \sum_{r_j=0}q_j\frac{\mathbf{h}(\mathbf{x}_j;\boldsymbol{\theta})-\boldsymbol{\mu}}{1-\pi_j(\hat{\boldsymbol{\phi}})} = 0,$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimator. Qin et al. (2009) considered maximizing the complete likelihood $l_c = \sum_{i=1}^{n}\log(\omega_i)$ subject to

$$(10) \qquad \sum_{i=1}^{n}\omega_i = 1, \quad \sum_{i=1}^{n}\omega_i\frac{r_i}{\pi_i(\hat{\boldsymbol{\phi}})}\mathbf{U}_i(\boldsymbol{\theta}) = 0,$$

and

$$(11) \qquad \sum_{i=1}^{n}\omega_i\left\{\frac{r_i}{\pi_i(\hat{\boldsymbol{\phi}})} - 1\right\}\mathbf{h}_i(\boldsymbol{\theta}) = 0,$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimator. The computation requires that the individual values of $\mathbf{x}_i$ for $r_i = 0$ be available, which is not always possible, as discussed in Section 1. For example, in survey sampling problem, we only observe $(\mathbf{x}_i, y_i)$ for $r_i = 1$ and the aggregate information $\bar{\mathbf{x}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ is available. In this case, the method of Qin et al. (2009) is not applicable.

In Section 3, some asymptotic properties of the proposed EL estimator described in (5)–(7) are developed for the case when $\pi_i = \Pr(r_i = 1 \mid \mathbf{x}_i, y_i)$ is a known function of $(\mathbf{x}_i, y_i)$. In particular, we show that the optimal choice of $\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})$ that minimizes the asymptotic variance of the resulting EL estimator of $\boldsymbol{\theta}$ is

$$\mathbf{h}^*(\mathbf{x}_i; \boldsymbol{\theta}) = \tilde{\mathbf{U}}(\mathbf{x}_i; \boldsymbol{\theta}) \equiv E\left\{\mathbf{U}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) \mid \mathbf{x}_i\right\}.$$

In Section 4, we consider the case where $\pi_i = \Pr(r_i = 1 \mid \mathbf{x}_i, y_i)$ is a parametric model of the form $\Pr(r = 1 \mid \mathbf{x}, y) = \pi(\mathbf{x}; \boldsymbol{\phi}_0)$ for some $\boldsymbol{\phi}_0$. By plugging estimator $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}_0$ into the empirical likelihood procedure, we can find the empirical likelihood estimator. The asymptotical properties of this estimator are discussed in Section 4. If a parametric form of π is unknown, we can use a nonparametric model for π. Asymptotical properties of the EL estimator using a nonparametric estimator of π are discussed in Section 5.

3. ESTIMATION WITH KNOWN RESPONSE PROBABILITY

In this section, we assume that the true response probability $\pi = \Pr(r = 1|\mathbf{X}, Y)$ is known, which is often the case with survey sampling where π_i denotes the first-order inclusion probability and the response indicator r represents the sampling indicator. The regularity conditions of this section can be found in Appendix A.1. Our proposed estimator introduced in Section 2, (5)–(7), can be described as maximizing

$$(12) \qquad l = \sum_{r_i=1} \log(\omega_i),$$

subject to

$$(13)$$

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \pi_i^{-1} \left\{ \mathbf{h}_i(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}) \right\} = \mathbf{0},$$

$$\sum_{r_i=1} \omega_i \pi_i^{-1} \mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0}.$$

The choice of $\mathbf{h}_i(\boldsymbol{\theta})$ depends on two factors: One is the statistical efficiency and the other is the availability of $n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})$ in the sample. In some situations, individual values of $\mathbf{x}_i$ are not available in the sample, and $\bar{\mathbf{x}}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ is the only available information. In this case, $\mathbf{h}_i(\boldsymbol{\theta}) = \mathbf{x}_i$ can be used in (13), even though it may not be the optimal choice for minimizing the variance. The EL estimator of $\boldsymbol{\theta}$ is then obtained by $\hat{\boldsymbol{\theta}}_{h1} = \sum_{r_i=1} w_i^* \pi_i^{-1} y_i / \sum_{r_i=1} w_i^* \pi_i^{-1}$ where $w_i^* = n_r^{-1}\{1 + \hat{\boldsymbol{\lambda}} \pi_i^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_n)\}^{-1}$, $\bar{\mathbf{x}}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i$, and $\hat{\boldsymbol{\lambda}}$ is constructed to satisfy $\sum_{r_i=1} w_i^* \pi_i^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_n) = \mathbf{0}$.

The following theorem presents some asymptotic properties of the EL estimator $\hat{\boldsymbol{\theta}}_{h1}$. The proof is presented in Appendix A.2.

Theorem 3.1. *Let $\hat{\boldsymbol{\theta}}_{h1}$ be the solution to the maximization above. Then, under the regularity conditions (C1)–(C5) in Appendix A.1, we have*

$$(14)$$

$$\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0$$

$$= \boldsymbol{\tau} \frac{1}{n} \sum_{i=1}^n \left\{ \frac{r_i}{\pi_i} \mathbf{U}_i(\boldsymbol{\theta}_0) - \mathbf{B}\left(\frac{r_i}{\pi_i} - 1\right)\tilde{\mathbf{h}}_i(\boldsymbol{\theta}_0) \right\} + \mathbf{o}_p(n^{-1/2}),$$

where $\tilde{\mathbf{h}}_i(\boldsymbol{\theta}_0) = \mathbf{h}_i(\boldsymbol{\theta}_0) - \boldsymbol{\mu}_h$, $\boldsymbol{\mu}_h = \mathrm{E}(\mathbf{h})$, $\mathbf{B} = \mathrm{E}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi)\{\mathrm{E}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi)\}^{-1}$, and $\boldsymbol{\tau} = -\{\mathrm{E}(\partial\mathbf{U}/\partial\boldsymbol{\theta})\}^{-1}$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Hence, we have

$$(15) \qquad \sqrt{n}\left(\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0\right) \to^d N(0, \mathbf{V}_{h1}),$$

where $\to^d$ denotes convergence in distribution, $\mathbf{V}_{h1} = \boldsymbol{\tau}\boldsymbol{\Omega}_{h1}\boldsymbol{\tau}^\top$ and

$$\boldsymbol{\Omega}_{h1} = \mathrm{V}\left\{ \frac{r}{\pi}(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}}) + \mathbf{B}\tilde{\mathbf{h}} \right\}$$

$$(16) \qquad = \mathrm{E}\left\{ \left(\frac{1}{\pi} - 1\right)(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})^{\otimes 2} \right\} + \mathrm{V}(\mathbf{U}),$$

and $A^{\otimes 2} = AA^\top$.

Because $\boldsymbol{\Omega}_{h1} = \mathrm{E}\{(\pi^{-1} - 1)(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})^{\otimes 2}\} + \mathrm{V}(\mathbf{U})$, we always have $\mathrm{V}(\hat{\boldsymbol{\theta}}_{h1}) \geq \mathrm{V}(\hat{\boldsymbol{\theta}}_n)$, where $\hat{\boldsymbol{\theta}}_n$ is the solution to (1). According to the above theorem, we can get the consistent variance estimator by using $\hat{\mathbf{V}}_{h1} = \hat{\boldsymbol{\tau}}\hat{\boldsymbol{\Omega}}_{h1}\hat{\boldsymbol{\tau}}^\top$, where $\hat{\boldsymbol{\tau}} = -\{n^{-1}\sum_{i=1}^n r_i\pi_i^{-1}(\partial\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h1})/\partial\boldsymbol{\theta})\}^{-1}$ and $\hat{\boldsymbol{\Omega}}_{h1} = (n-1)^{-1} \times \sum_{i=1}^n (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^{\otimes 2}$, where $\boldsymbol{\eta}_i = r_i\pi_i^{-1}\{\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h1}) - \hat{\mathbf{B}}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h1})\} + \hat{\mathbf{B}}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h1})$, $\hat{\mathbf{B}} = \hat{\mathrm{E}}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi)\{\hat{\mathrm{E}}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi)\}^{-1}$, $\hat{\mathrm{E}}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi) = n^{-1}\sum_{i=1}^n r_i\pi_i^{-2}\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h1})\tilde{\mathbf{h}}_i^\top(\hat{\boldsymbol{\theta}}_{h1})$, and $\hat{\mathrm{E}}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi) = n^{-1}\sum_{i=1}^n r_i\pi_i^{-2}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h1})\tilde{\mathbf{h}}_i^\top(\hat{\boldsymbol{\theta}}_{h1})$, with $\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h1}) = \mathbf{h}_i(\hat{\boldsymbol{\theta}}_{h1}) - \hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\mu}}_h = n^{-1}\sum_{i=1}^n h_i(\hat{\boldsymbol{\theta}}_{h1})$.

For the special case of $\theta = \mathrm{E}(Y)$ and $h = x$, after some algebra, we have

$$\hat{\theta}_{h1} = \hat{\bar{y}}_d - \hat{B}_1\hat{B}_2^{-1}(\hat{\bar{x}}_d - \bar{x}_n) + o_p(n^{-1/2}),$$

where $(\hat{\bar{y}}_d, \hat{\bar{x}}_d) = (\sum_{r_i=1}\pi_i^{-1})^{-1}(\sum_{r_i=1}\pi_i^{-1}y_i, \sum_{r_i=1}\pi_i^{-1}x_i)$, $\hat{B}_1 = n^{-1}\sum_{r_i=1}\pi_i^{-2}(x_i - \hat{\bar{x}}_d)(y_i - \hat{\bar{y}}_d)$, and $\hat{B}_2 = n^{-1}\sum_{r_i=1}\pi_i^{-2}(x_i - \hat{\bar{x}}_d)^2$, which is close to the optimal estimator within the linear class. The resulting estimator is asymptotically equivalent to the optimal EL estimator considered in Kim (2009).

Remark 3.1. The EL estimator of Chen et al. (2008) satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_c),$$

where $\mathbf{V}_c = \boldsymbol{\tau}\boldsymbol{\Omega}_{hc}\boldsymbol{\tau}^\top$, and $\boldsymbol{\Omega}_{hc} = \mathrm{V}\{r\pi^{-1}\mathbf{U} - \mathbf{B}^*(r - \pi)\pi^{-1}(1 - \pi)^{-1}\tilde{\mathbf{h}}\}$ with $\mathbf{B}^* = \mathrm{E}(\pi^{-1}\mathbf{U}\tilde{\mathbf{h}}^\top)[\mathrm{E}\{\pi^{-1}(1 - \pi)^{-1}\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top\}]^{-1}$. Thus, the estimator of Chen et al. (2008) achieves the minimum variance when $\mathbf{h}/(1 - \pi) \propto \mathrm{E}(\mathbf{U} \mid \mathbf{x})$ while the asymptotic variance of the proposed EL estimator is minimized when $\mathbf{h} \propto \mathrm{E}(\mathbf{U} \mid \mathbf{x})$. The Qin-Zhang-Leung (QZL) estimator $\hat{\theta}_{QZL}$ defined in (10) and (11) satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{hq} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_q),$$

where $\mathbf{V}_q = \boldsymbol{\tau}\boldsymbol{\Omega}_{hq}\boldsymbol{\tau}^\top$, $\boldsymbol{\Omega}_{hq} = \mathrm{V}\left\{r\pi^{-1}\mathbf{U} - (r - \pi)\pi^{-1}\mathbf{B}_q\mathbf{h}\right\}$ and $\mathbf{B}_q = \mathrm{E}\{(\pi^{-1} - 1)\mathbf{U}\mathbf{h}^\top\}[\mathrm{E}\{(\pi^{-1} - 1)\mathbf{h}\mathbf{h}^\top\}]^{-1}$. Note that the choice of $\mathbf{B} = \mathbf{B}_q$ minimizes the variance of $r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathbf{B}\mathbf{h}$ and the QZL estimator is optimal in the sense that it minimizes the variance among its class. This is because the QZL estimator uses the complete likelihood $\sum_{i=1}^n \log(\omega_i)$ while our proposed estimator uses only pseudo-likelihood. If $\mathbf{h} \propto \mathrm{E}(\mathbf{U}|\mathbf{X})$, then all the estimators, excluding the estimator of Chen et al. (2008), achieve the same asymptotic variance. A numerical comparison is also made through a simulation study in Section 6.

In the following corollary, we find an optimal constraint that minimizes the asymptotic variance in (15). The proof is presented in Appendix A.3.

Corollary 3.1. *Under the setup of Theorem 3.1, the asymptotic variance of $\hat{\boldsymbol{\theta}}_{h1}$ is minimized when $\mathbf{h} \propto \mathbf{h}^* = \mathrm{E}(\mathbf{U}|\mathbf{X})$. The asymptotic variance satisfies*

$$(17) \quad \mathbf{V}_{h1} \geq \boldsymbol{\tau} \left\{ \mathrm{E}\left(\frac{\mathbf{U}\mathbf{U}^\top}{\pi}\right) - \mathrm{E}\left(\frac{1-\pi}{\pi}\mathbf{h}^*\mathbf{U}^\top\right) \right\} \boldsymbol{\tau}^\top.$$

The lower bound in (17) is the same as the semiparametric lower bound for the asymptotic variance discussed in Robins et al. (1994) and Chen et al. (2008).

Remark 3.2. To compute the solution to the constrained optimization problem of maximizing (12) subject to (13), the following two-step algorithm can be used. In the first step, the optimal weight that maximizes (12) subject to $\sum_{r_i=1} \omega_i = 1$ and $\sum_{r_i=1} \omega_i \pi_i^{-1}(\hat{\mathbf{h}}_i - n^{-1}\sum_{j=1}^n \hat{\mathbf{h}}_j) = \mathbf{0}$ are computed, where $\hat{\mathbf{h}}_i = \mathbf{h}_i(\hat{\boldsymbol{\theta}}_0)$ and $\hat{\boldsymbol{\theta}}_0$ is the solution to $\sum_{r_i=1} \pi_i^{-1}\mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0}$. In the second step, we can get the resulting EL estimator $\hat{\boldsymbol{\theta}}_{h1}$ by solving

$$\sum_{r_i=1} \hat{\omega}_i \pi_i^{-1}\mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0}.$$

Such a two-step algorithm was discussed in Chaudhuri et al. (2008) when the control function h_i does not depend on θ. Using $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{x}_i; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is any $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}$, in the two-step optimization is asymptotically equivalent to the original solution.

4. ESTIMATION WITH UNKNOWN RESPONSE PROBABILITY

We now consider the case when the response probability is known up to some parameter and has the known form

$$Pr(r = 1|\mathbf{X}, Y) = \pi(\mathbf{X}; \boldsymbol{\phi}_0),$$

for some $\boldsymbol{\phi}_0$. Thus, we assume that the response mechanism is ignorable. We also assume that there exists $\hat{\phi}$ such that

$$(18) \qquad \hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 = \frac{1}{n}\sum_{i=1}^n \mathbf{b}(\mathbf{x}_i, r_i; \boldsymbol{\phi}_0) + \mathbf{o}_p(n^{-1/2}),$$

for some function $\mathbf{b}$ with $\mathrm{E}\{\mathbf{b}(\mathbf{X}_i, r_i; \boldsymbol{\phi}_0)\} = \mathbf{0}$ and $\mathrm{V}\{\mathbf{b}(\mathbf{X}_i, r_i; \boldsymbol{\phi}_0)\} = \mathbf{V}_b$, where $\mathbf{V}_b$ is positive definite.

If the true response probability $\pi_i = \pi_i(\boldsymbol{\phi}_0)$ is estimated by $\hat{\pi}_i = \pi_i(\hat{\boldsymbol{\phi}})$, then the proposed EL estimator can be described as maximizing (12) subject to
(19)

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \hat{\pi}_i^{-1}\left\{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\right\} = \mathbf{0},$$

and

$$(20) \qquad \sum_{r_i=1} \omega_i \hat{\pi}_i^{-1}\mathbf{U}(\boldsymbol{\theta}; \mathbf{x}_i, y_i) = \mathbf{0}.$$

The following theorem presents some asymptotic properties of the proposed EL estimator.

Theorem 4.1. *Let $\hat{\boldsymbol{\phi}}$ be a $\sqrt{n}$-consistent estimator of $\boldsymbol{\phi}_0$, satisfying (18). Let $\hat{\boldsymbol{\theta}}_{h2}$ be obtained by maximizing (12) subject to the constraints (19) and (20). Under the same regularity conditions as Theorem 3.1 and (18), we have*

(21)

$$\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0$$
$$= \boldsymbol{\tau}\frac{1}{n}\sum_{i=1}^n \left\{\frac{r_i}{\pi_i}\mathbf{U}_i(\boldsymbol{\theta}_0) - \mathbf{B}\left(\frac{r_i}{\pi_i} - 1\right)\tilde{\mathbf{h}}_i(\boldsymbol{\theta}_0) - \mathbf{C}\mathbf{b}_i(\boldsymbol{\phi}_0)\right\}$$
$$+ \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{B}$ is defined in (14), $\boldsymbol{\tau} = -\{\mathrm{E}(\partial\mathbf{U}/\partial\boldsymbol{\theta})\}$, $\mathbf{C} = \mathrm{E}\{\pi^{-1}(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})(\partial\pi/\partial\boldsymbol{\phi})^\top\}$, and $\mathbf{b}_i(\boldsymbol{\phi}_0) = \mathbf{b}(\mathbf{x}_i, r_i; \boldsymbol{\phi}_0)$ defined in (18). Hence, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h2}),$$

where $\mathbf{V}_{h2} = \boldsymbol{\tau}\boldsymbol{\Omega}_{h2}\boldsymbol{\tau}^\top$ and $\boldsymbol{\Omega}_{h2} = \mathrm{V}\{r\pi^{-1}(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}}) + \mathbf{B}\tilde{\mathbf{h}} - \mathbf{C}\mathbf{b}\}$.

A consistent variance estimator of $\mathbf{V}_{h2}$ can be constructed by

$$\hat{\mathbf{V}}_{h2} = \hat{\boldsymbol{\tau}}\hat{\boldsymbol{\Omega}}_{h2}\hat{\boldsymbol{\tau}}^\top,$$

where $\hat{\boldsymbol{\tau}} = -[n^{-1}\sum_{i=1}^n r_i\hat{\pi}_i^{-1}\{\partial\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h2})/\partial\boldsymbol{\theta}\}]^{-1}$ and $\hat{\boldsymbol{\Omega}}_{h2} = (n-1)^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^{\otimes 2}$, where $\boldsymbol{\eta}_i = r_i\hat{\pi}_i^{-1}\{\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h2}) - \hat{\mathbf{B}}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h2})\} + \hat{\mathbf{B}}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h2}) - \hat{\mathbf{C}}\mathbf{b}_i(\hat{\boldsymbol{\phi}})$,

$$\hat{\mathbf{C}} = n^{-1}\sum_{i=1}^n r_i\hat{\pi}_i^{-2}\left\{\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h2}) - \hat{\mathbf{B}}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h2})\right\}(\partial\hat{\pi}_i/\partial\boldsymbol{\phi})^\top,$$

$$\hat{\mathbf{B}} = \hat{\mathrm{E}}(\mathbf{U}\mathbf{h}^\top/\pi)\left\{\hat{\mathrm{E}}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi)\right\}^{-1},$$

where

$$\hat{\mathrm{E}}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi) = n^{-1}\sum_{i=1}^n r_i\hat{\pi}_i^{-2}\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h2})\tilde{\mathbf{h}}_i^\top(\hat{\boldsymbol{\theta}}_{h2}),$$

$$\hat{\mathrm{E}}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi) = n^{-1}\sum_{i=1}^n r_i\hat{\pi}_i^{-2}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h2})\tilde{\mathbf{h}}_i^\top(\hat{\boldsymbol{\theta}}_{h2}),$$

with $\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}_{h2}) = \mathbf{h}_i(\hat{\boldsymbol{\theta}}_{h2}) - \hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\mu}}_h = n^{-1}\sum_{i=1}^n \mathbf{h}_i(\hat{\boldsymbol{\theta}}_{h2})$. Comparing (21) with (14), we have an extra term, $-\mathbf{C}\mathbf{b}_i(\boldsymbol{\phi}_0)$, in the linearization. This is because we have additional randomness due to estimating parameter $\boldsymbol{\phi}_0$.

Remark 4.1. If we use $\mathbf{h} = a\mathbf{h}^* = a\mathrm{E}(\mathbf{U}|\mathbf{X})$ in the constraint (19) for some constant $a \neq 0$, we have $\mathbf{B} = \mathrm{E}(\mathbf{U}\mathbf{h}^\top/\pi)\mathrm{E}^{-1}(\mathbf{h}\mathbf{h}^\top/\pi) = a^{-1}\mathbf{I}$ and

$$\mathbf{C} = \mathrm{E}\left\{\pi^{-1}(\mathbf{U} - \mathbf{B}\mathbf{h})(\partial\pi/\partial\phi)^\top\right\}$$
$$= \mathrm{E}\left[\mathrm{E}\left\{\pi^{-1}(\mathbf{U} - \mathbf{h}^*)(\partial\pi/\partial\phi)^\top|\mathbf{X}\right\}\right] = \mathbf{0}.$$

Thus, the asymptotic variance is equal to

$$\mathbf{V}_{h2} = \boldsymbol{\tau}\left\{\mathrm{E}\left(\frac{\mathbf{U}\mathbf{U}^\top}{\pi}\right) - \mathrm{E}\left(\frac{1-\pi}{\pi}\mathbf{h}^*\mathbf{U}^\top\right)\right\}\boldsymbol{\tau}^\top,$$

which is equal to the lower bound in (17) when the propensity score is known. Under the optimal choice of $\mathbf{h}$, the lower bound for the asymptotic variance is achieved regardless of whether the propensity score is known or estimated.

According to Remark 4.1, the choice of $\hat{\boldsymbol{\phi}}$ does not make any difference in the asymptotic variance of $\hat{\boldsymbol{\theta}}_{h2}$, as long as $\mathbf{h} \propto \mathrm{E}(\mathbf{U} \mid \mathbf{X})$ is used in (19). If $\mathbf{h} \propto \mathrm{E}(\mathbf{U} \mid \mathbf{X})$ does not hold, then the choice of $\hat{\boldsymbol{\phi}}$ makes a difference. While the MLE of $\boldsymbol{\phi}_0$ is a popular choice, it does not necessarily lead to the optimal estimator. To see this, let $\hat{\boldsymbol{\phi}}_q$ be a consistent estimator of $\boldsymbol{\phi}_0$ that can be obtained by solving the following equation:

$$(22) \qquad \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi_i(\boldsymbol{\phi})} - 1\right\}\mathbf{q}_i(\boldsymbol{\phi}) = \mathbf{0},$$

where $\mathbf{q}_i(\boldsymbol{\phi})$ is an arbitrary function to make the solution of (22) unique. Note that equation (22) can be called the calibration equation in the sense that the estimator for the mean of $\mathbf{q}_i$, using the propensity score, is equal to the sample mean of $\mathbf{q}_i$. The MLE of $\boldsymbol{\phi}_0$ also belongs to the class because it satisfies (22) with $\mathbf{q}_i = \pi_i\mathrm{logit}(\pi_i)$. Under some regularity conditions, we have $\hat{\boldsymbol{\phi}}_q \to^p \boldsymbol{\phi}_0$, regardless of the choice of $\mathbf{q}_i$. For example, instead of using the maximum likelihood method, we can use $\mathbf{q}_i(\boldsymbol{\phi}) = (1, \mathbf{x}_i^\top)^\top$ in (22) to estimate $\boldsymbol{\phi}_0$. Such estimation is particularly useful when individual values of $\mathbf{x}_i$ are not available outside the respondents, but the sample mean of $\mathbf{x}_i$ is available from an external source.

The following theorem discusses the optimal choice of $\mathbf{q}$ in the calibration equation (22).

Theorem 4.2. *Let $\hat{\boldsymbol{\phi}}_q$ be the estimator which solves (22) and satisfies $\hat{\boldsymbol{\phi}}_q \to^p \boldsymbol{\phi}_0$. Under the same regularity conditions as Theorem 4.1, we have*

$$(23) \qquad \hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0$$
$$= \boldsymbol{\tau}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\pi_i}\mathbf{U}_i(\boldsymbol{\theta}_0) - (\frac{r_i}{\pi_i} - 1)\left\{\mathbf{B}\tilde{\mathbf{h}}_i(\boldsymbol{\theta}_0)\right.\right.$$
$$\left.\left. + \mathbf{C}\mathbf{S}^{-1}\mathbf{q}(\boldsymbol{\phi}_0)\right\}\right] + \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{B}$ is defined in (14), $\boldsymbol{\tau} = -\{\mathrm{E}(\partial\mathbf{U}/\partial\boldsymbol{\theta})\}$, $\mathbf{S} = \mathrm{E}\{\pi^{-1}\mathbf{q}(\partial\pi/\partial\phi)^\top\}$, $\mathbf{C} = \mathrm{E}\{\pi^{-1}(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})(\partial\pi/\partial\phi)^\top\}$.

Hence, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h2}),$$

where $\mathbf{V}_{h2} = \boldsymbol{\tau}\boldsymbol{\Omega}_{h2}\boldsymbol{\tau}^\top$, and $\boldsymbol{\Omega}_{h2} = \mathrm{V}\{r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q})\}$. In addition, we have

$$\mathbf{V}_{h2} \geq \boldsymbol{\tau}\mathrm{V}\left\{r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathbf{h}^*\right\}\boldsymbol{\tau}^\top$$

with equality if $\boldsymbol{\alpha}^\top\mathbf{q} = \mathbf{h}^* - \mathbf{B}\tilde{\mathbf{h}}$ for some $\boldsymbol{\alpha}$.

In Theorem 4.2, the meaning of $\mathbf{h}^* - \mathbf{B}\tilde{\mathbf{h}}$ is the residual for the regression of $\mathbf{h}^* = \mathrm{E}(\mathbf{U}|\mathbf{X})$ on $\tilde{\mathbf{h}}$. If $\mathbf{h} \propto \mathbf{h}^*$, then the residual is equal to zero and the lower bound is achieved, as discussed in Remark 4.1. If $\mathbf{h} \propto \mathbf{h}^*$ does not hold, we cannot achieve the lower bound and the efficiency can be improved by how well $\mathbf{q}$ explains the conditional expectation $\mathrm{E}(\mathbf{U} \mid \mathbf{X})$. In the extreme case of $\mathbf{h} \equiv \mathbf{0}$, the choice of $\mathbf{q} \propto \mathbf{h}^*$ achieves the lower bound while the choice of $\mathbf{q}_i \propto \pi_i\mathbf{h}^*$, which corresponds to the maximum likelihood estimation of $\boldsymbol{\phi}_0$, does not achieve the lower bound and thus leads to less efficient estimation.

5. NONPARAMETRIC ESTIMATION OF THE RESPONSE MECHANISM

In this section, we consider nonparametric estimation of the response probability. For simplicity, we assume the response mechanism is ignorable, $\pi(\mathbf{x}) = \mathrm{Pr}(r = 1|\mathbf{x})$, and consider estimation of π nonparametrically. To this end, we consider kernel estimation of the response model as below:

$$(24) \qquad \hat{\pi}_H(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{X}_i)r_i}{\sum_{j=1}^{n} K_H(\mathbf{x} - \mathbf{X}_j)},$$

where $K_H(\mathbf{s})$ is the kernel function which satisfies certain regularity conditions and H is the bandwidth. In addition, we define $K_H(\mathbf{s}, \mathbf{t}) = K\{(\mathbf{s} - \mathbf{t})/H\}$. Let $f(\mathbf{x})$ be the probability density function of $\mathbf{X}$. In addition to regularity conditions (C1)–(C5) in Appendix A.1, we also assume the following regularity conditions:

(C6) $f(\mathbf{x})$ and $\pi(\mathbf{x})$ have bounded partial derivatives with respect to $\mathbf{x}$ up to an order q with $q \geq 2$, $2q > d_x$ almost surely, where d_x is the dimension of $\mathbf{x}$.

(C7) The kernel function $K(\mathbf{s})$ is a probability density function such that

  (a) It is bounded and has compact support.

  (b) $\int K(s_1, \ldots, s_{d_x})ds_1 \cdots ds_{d_x} = 1$,

  (c) $\int s_i^l K(s_1, \ldots, s_{d_x})ds_1 \cdots ds_{d_x} = 0$ for any $i = 1, \ldots, d_x$ and $1 \leq l < q$.

  (d) $\int s_i^q K(s_1, \ldots, s_{d_x})ds_1 \cdots ds_{d_x} \neq 0$.

(C8) $nH^{2d_x} \to \infty$, $\sqrt{n}H^q \to 0$, as $n \to \infty$.

(C9) $1 > \pi(\mathbf{x}) > d > 0$ almost surely.

Conditions (C6)–(C8) are common conditions used for nonparametric problems. In condition (C8), we used $\sqrt{n}H^q \to 0$ to control the bias due to kernel smoothing, and $nH^{2d_x} \to \infty$ is used to produce consistent estimation of the conditional distribution as well as control the convergence rate of response probability estimation. Condition (C9) is used to avoid extreme propensity scores. Wang and Chen (2009) used the kernel smoothing method to estimate the conditional distribution function. They claimed that even though the curse of dimensionality is an issue for estimating the conditional distribution function, it does not play a leading order effect on estimating the finite dimensional parameter θ as long as the bias of the kernel estimator is controlled. Similar results will hold in our EL method when estimating propensity scores nonparametrically.

Under those regularity conditions, the proposed empirical likelihood method can be constructed similarly by maximizing (12) subject to

(25)

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \hat{\pi}_{i,H}^{-1} \left\{ \mathbf{h}_i(\boldsymbol{\theta}) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}) \right\} = 0,$$
$$\sum_{r_i=1} \omega_i \hat{\pi}_{i,H}^{-1} \mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0},$$

where $\hat{\pi}_{i,H} = \hat{\pi}_H(\mathbf{x}_i)$.

The following theorem presents some asymptotic properties of the proposed EL estimator of $\boldsymbol{\theta}_0$ using nonparametric response probability (24).

Theorem 5.1. *Let $\hat{\boldsymbol{\theta}}_{h3}$ be the empirical likelihood estimator that is obtained by maximizing (12) subject to (25). Under the regularity conditions (C1)–(C9), we have*

(26)

$$\hat{\boldsymbol{\theta}}_{h3} - \boldsymbol{\theta}_0$$
$$= -\boldsymbol{\tau} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{r_i}{\pi_i} \mathbf{U}_i(\boldsymbol{\theta}_0) - \frac{1}{n} \sum_{i=1}^n \left(\frac{r_i}{\pi_i} - 1 \right) \mathbf{h}_i^*(\boldsymbol{\theta}_0) \right\}$$
$$+ \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{h}^*(\boldsymbol{\theta}_0) = \mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\}$ and $\boldsymbol{\tau} = -\{\mathrm{E}(\partial \mathbf{U}/\partial \boldsymbol{\theta})\}^{-1}$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Furthermore, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h3} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h3}),$$

where $\mathbf{V}_{h3} = \boldsymbol{\tau}\boldsymbol{\Omega}_{h3}\boldsymbol{\tau}^\top$ and $\boldsymbol{\Omega}_{h3} = \mathrm{V}\{r\pi^{-1}(\mathbf{U} - \mathbf{h}^*) + \mathbf{h}^*\}$.

By Theorem 5.1, the asymptotic variance of $\hat{\boldsymbol{\theta}}_{h3}$ using nonparametric $\hat{\pi}_{i,H}$ is equal to the semiparametric lower bound in (17). Note that the linearization in (26) does not depend on the choice of h in (25). This means that the same result (26) can be achieved for different choices of $\mathbf{h}$. This

is because, according to (40) in the proof of Theorem 5.1 in Appendix A.6, we have

$$\frac{1}{n} \sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} \tilde{\mathbf{h}}_i - \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{h}}_i = \mathbf{o}_p(n^{-1/2}),$$

where $\tilde{\mathbf{h}} = \mathbf{h} - \boldsymbol{\mu}_h$ and $\mathbf{h}$ is any arbitrary function of $\mathbf{x}$ with some moment conditions described in Appendix A.1. In addition, the second constraint in (25) can be written as $\sum_{r_i=1} \omega_i \hat{\pi}_{i,H}^{-1}(\tilde{\mathbf{h}}_i - n^{-1} \sum_{i=1}^n \tilde{\mathbf{h}}_i) = \mathbf{0}$, which implies that we can safely remove it. As was also pointed out by Hirano et al. (2003), the inverse probability weighted estimator is globally semiparametrically efficient and the calibration step does not improve the efficiency. Thus, using the nonparametric estimator (24) of the response probability, the EL solution can be written as maximizing (12) subject to

(27)
$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \hat{\pi}_{i,H}^{-1} \mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0},$$

which is equivalent to obtain $\hat{\boldsymbol{\theta}}_{h3}$ by solving $n^{-1} \sum_{i=1}^n r_i \hat{\pi}_{i,H}^{-1} \mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{0}$.

According to Theorem 5.1, a consistent variance estimator for $\mathbf{V}_{h3}$ is $\hat{\mathbf{V}}_{h3} = \hat{\boldsymbol{\tau}}\hat{\boldsymbol{\Omega}}_{h3}\hat{\boldsymbol{\tau}}^\top$, where $\hat{\boldsymbol{\tau}} = -\{n^{-1} \sum_{i=1}^n r_i \hat{\pi}_{i,H}^{-1} \partial \mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h3})/\partial \boldsymbol{\theta}\}^{-1}$ and $\hat{\boldsymbol{\Omega}}_{h3} = (n-1)^{-1} \sum_{i=1}^n (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^{\otimes 2}$, where $\boldsymbol{\eta}_i = r_i \hat{\pi}_{i,H}^{-1}(\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h3}) - \hat{\mathbf{h}}_i^*) + \hat{\mathbf{h}}_i^*$, where $\hat{\mathbf{h}}_i^*$ is a consistent estimator of $\mathrm{E}(\mathbf{U}|\mathbf{X})$ under the working model. In the special case when $\boldsymbol{\theta}_0$ satisfies $\mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$, then a version of Wilk's Theorem can be established as below.

Theorem 5.2. *Assume that the regularity conditions in Theorem 5.1 hold and $\boldsymbol{\theta}_0$ satisfies $\mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$. Let $R_n(\boldsymbol{\theta}_0) = 2\{l(\hat{\boldsymbol{\theta}}_{h3}) - l(\boldsymbol{\theta}_0)\}$, where $l(\boldsymbol{\theta}) = \sum_{r_i=1} \log\{\omega_i(\boldsymbol{\theta})\}$ with $\omega_i(\boldsymbol{\theta})$ obtained by maximizing (12) subject to (27). Then, as $n \to \infty$,*

$$R_n(\boldsymbol{\theta}_0) \to^d \chi_p^2,$$

where p is the dimension of $\boldsymbol{\theta}$.

According to Theorem 5.2, we can construct a Wilk-type confidence region for $\boldsymbol{\theta}_0$ without calculating the variance estimator, if $\mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$ holds. For example, if $\mathrm{E}(Y|\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\theta}$, then $\mathbf{U} = (Y - \mathbf{X}^\top \boldsymbol{\theta})\mathbf{X}$ satisfies $\mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\} = 0$. If $\mathrm{E}(\mathbf{U}|\mathbf{X}) \neq \mathbf{0}$, but we know $\mathrm{E}(\mathbf{U}|\mathbf{X})$, the result for Theorem 5.2 holds by replacing $\mathbf{U}$ with $\mathbf{U}^* = \mathbf{U} - \mathrm{E}(\mathbf{U}|\mathbf{X})$. Alternatively, a resampling method, such as bootstrap or jackknife, can be used to construct a confidence region for $\boldsymbol{\theta}_0$.

6. SIMULATION STUDY

In the simulation study, the following two models were considered to generate the samples:

[A] $x_i \overset{iid}{\sim} N(1,1), \quad z_i \overset{iid}{\sim} N(0,1), \quad e_i \sim^{iid} exp(1) - 1$, and $y_i = 0.5 + 0.5x_i + 0.5z_i + e_i$.

[B] (x_i, z_i, e_i) are the same as in [A] and $y_i = 0.8(x_i - 0.5)^2 + 0.8e_i$.

For each model, $B = 2,000$ Monte Carlo samples of size $n = 200$ were independently generated. In addition, two different response mechanisms were used to generate r_i, the response indicator function for y_i. The response mechanisms are

[M1] $P(r = 1|x, y) = \exp(-0.5 + x)/[1 + \exp(-0.5 + x)]$.
[M2] $P(r = 1|x, y) = (0.3 + 0.175|x|)I(|x| < 1.5) + I(|x| \geq 1.5)$.

Thus, the two models are ignorable and the response rate is about 0.6 in both response mechanisms. We are interested in estimating $\theta_0 = E(Y)$, which is the population mean of Y. Thus, we use $U(\theta) = Y - \theta$. We assume that the working model for $E(y \mid x, z)$ is linear in x and z. That is, $E(y \mid x, z) = \beta_0 + \beta_1 x + \beta_2 z$. Also, the working model for $\pi(x) = E(r \mid x)$ is the logistic regression model with logit$\{\pi(x)\} = \phi_0 + \phi_1 x$. That is, even when the true response mechanism is [M2], we use the logistic regression model to obtain $\hat{\pi}_i = \exp(\hat{\phi}_0 + \hat{\phi}_1 x_i)/\{1 + \exp(\hat{\phi}_0 + \hat{\phi}_1 x_i)\}$. Thus, we have the following four possible scenarios:

1: Both working models are correct. That is, the samples are generated by [A] and [M1].
2: Only the outcome regression model, the model for $E(y \mid x)$, is correct. That is, the samples are generated by [A] and [M2].
3: Only the response probability model, the model for $E(r \mid x)$, is correct. That is, the samples are generated by [B] and [M1].
4: Both models are incorrect. That is, the samples are generated by [B] and [M2].

Under this setup, we considered eight estimators of θ_0.

1. QZ: The EL estimator of Qin and Zhang (2007) using $\mathbf{h}_i = (x_i, z_i)^\top$ in (8).
2. CLQ: The EL estimator of Chen et al. (2008) using $\mathbf{h}_i = (x_i, z_i)^\top$ in (9).
3. QZL: The EL estimator of Qin et al. (2009), which is obtained by maximizing $\sum_{i=1}^n \log(\omega_i)$, subject to $\sum_{i=1}^n \omega_i = 1$, $\sum_{i=1}^n \omega_i r_i \hat{\pi}_i^{-1}(y_i - \theta) = 0$, and $\sum_{i=1}^n \omega_i(r_i \hat{\pi}_i^{-1} - 1)\mathbf{h}_i = \mathbf{0}$, with $\mathbf{h}_i = (x_i, z_i)^\top$ and $\hat{\pi}_i$ computed by the MLE of ϕ.
4. NEW (MLE): The proposed EL estimator using $\mathbf{h}_i = (x_i, z_i)^\top$ in (19), where $\hat{\pi}_i$ is computed by the MLE of ϕ.
5. NEW (CAL): The proposed EL estimator using $\mathbf{h}_i = (x_i, z_i)^\top$ in (19), where $\hat{\pi}_i$ is computed by the calibration method on $(1, x)$. That is, $\hat{\phi}$ is computed by solving (22) with $\mathbf{q}_i = (1, x_i)^\top$.
6. NEW (NP1): The proposed EL estimator using nonparametric estimator (24) of $\pi(x) = P(r = 1 \mid x)$ and $\mathbf{h}_i = (x_i, z_i)^\top$ in (19). In addition, we used the Gaussian kernel and the reference bandwidth $H = 1.06\hat{\sigma}_x n^{-1/5}$, where $\hat{\sigma}_x$ is the estimated standard deviation of x_i in the sample.
7. NEW (NP2): The proposed EL estimator using nonparametric estimator (24) of $\pi(x) = P(r = 1 \mid x)$ without using $\mathbf{h}_i = (x_i, z_i)^\top$ in (19). We used the same kernel density and bandwidth as NEW (NP1).

Table 1 presents the Monte Carlo biases, variances, and mean square errors of the eight estimators under the four difference scenarios. Under Scenario 1, when both the outcome regression model and the response probability model are correct, the simulation results in Table 1 show that all the estimators are comparable since they all achieve the semiparametric lower bound except for CLQ, as discussed in Remark 3.1 and Remark 4.1. The NP2 method also shows some efficiency loss because the nonparametric propensity estimator does not make use of z_i information. Under Scenario 2, when only the outcome regression model is correct, the CLQ estimator shows a large bias, suggesting that the CLQ estimator is not robust against the failure of the response model. In terms of efficiency, the QZ method, QZL method, and the proposed EL estimators show the smallest variances. Under Scenario 3, when only the response probability model is correct, the biases are all negligible. The QZL estimator is more efficient than the proposed EL estimator using maximum-likelihood estimation, which is discussed in Remark 3.1. The nonparametric estimators, NEW (NP1) and NEW (NP2), show good efficiency because they automatically achieve the lower bound in (17) without correctly specifying the outcome regression model, which is consistent with the theory in Theorem 5.1. The NEW (NP2) is slightly more efficient than NEW (NP1) because it does not use calibration on the wrong outcome regression model. When both models are incorrect, as in Scenario 4, the nonparametric estimators still show negligible bias because they estimate the response probability consistently. In terms of efficiency, the nonparametric estimators are quite comparable because they achieve the semi-parametric lower bound.

7. CONCLUDING REMARKS

We have considered a respondent's weighting approach of empirical likelihood in the sense that the empirical likelihood weights are created only for the respondents, rather than being created for the entire original sample. Unlike the full-sample-weighting approach considered in Qin et al. (2009), the proposed method does not enjoy the nice property of the limiting chi-squared distribution of the likelihood ratio statistics. However, because the empirical likelihood weights are created only for the respondents, the proposed method is more useful in the situations where the individual values of the auxiliary variable are not available outside the respondents. For example, in voluntary surveys, the demographic information of the population can be easily incorporated into the proposed empirical likelihood method. Real data application using the proposed empirical likelihood approach will be presented elsewhere.

Table 1. Biases, variances and mean squared errors (MSE) of the estimators under four different scenarios in simulation study

Scenario	Method	Bias	Var	MSE
1	QZ	0.00	0.0129	0.0130
	CLQ	−0.00	0.0147	0.0147
	QZL	0.00	0.0130	0.0131
	NEW(MLE)	0.00	0.0130	0.0130
	NEW(CAL)	0.00	0.0130	0.0130
	NEW(NP1)	0.00	0.0130	0.0131
	NEW(NP2)	0.04	0.0129	0.0147
2	QZ	0.00	0.0119	0.0119
	CLQ	−0.16	0.0189	0.0467
	QZL	0.01	0.0116	0.0118
	NEW(MLE)	0.00	0.0119	0.0119
	NEW(CAL)	0.00	0.0119	0.0119
	NEW(NP1)	0.00	0.0120	0.0120
	NEW(NP2)	0.01	0.0128	0.0130
3	QZ	−0.03	0.0231	0.0247
	CLQ	−0.01	0.0210	0.0213
	QZL	−0.02	0.0197	0.0202
	NEW(MLE)	−0.02	0.0220	0.0227
	NEW(CAL)	−0.02	0.0216	0.0221
	NEW(NP1)	−0.05	0.0169	0.0196
	NEW(NP2)	−0.00	0.0156	0.0157
4	QZ	0.26	0.0307	0.0988
	CLQ	0.39	0.0602	0.2140
	QZL	0.22	0.0240	0.0762
	NEW(MLE)	0.29	0.0285	0.1150
	NEW(CAL)	0.30	0.0319	0.1255
	NEW(NP1)	0.03	0.0169	0.0180
	NEW(NP2)	0.04	0.0168	0.0189

APPENDIX A. APPENDIX SECTION

In the following appendix section, we provide the regularity conditions as well as the proofs of Theorems 3.1, 4.1, 4.2, 5.1, 5.2, and Corollary 3.1.

A.1 Regularity conditions for Theorem 3.1

(C1) $\boldsymbol{\theta}_0 \in \Theta$ is the unique solution to $\mathrm{E}\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\theta})\} = \mathbf{0}$, and Θ is compact; $\partial \mathbf{U}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is continuous at each $\boldsymbol{\theta} \in \Theta$ and $\mathrm{E}\{\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbf{g}(\mathbf{X}, Y; \boldsymbol{\theta})\|^\alpha\}$ is finite for some $\alpha > 2$, where $\mathbf{g}(\mathbf{X}, Y; \boldsymbol{\theta}) = (\mathbf{U}^\top(\mathbf{X}, Y; \boldsymbol{\theta}), \mathbf{h}^\top(\mathbf{X}; \boldsymbol{\theta}))^\top$.

(C2) The partial derivative $\partial \mathbf{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a continuous functions of $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_0$ almost everywhere.

(C3) $\|\mathbf{g}(\mathbf{X}, Y; \boldsymbol{\theta})\|^3$, $\|\partial \mathbf{g}(\mathbf{X}, Y; \boldsymbol{\theta})/\partial \boldsymbol{\theta}\|$, $\|\partial^2 \mathbf{g}(\mathbf{X}, Y; \boldsymbol{\theta})/ (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top)\|$, are bounded by some integrable function $\mathbf{G}(\mathbf{X}, Y)$.

(C4) The $p \times p$ matrix $\mathrm{E}\{\partial \mathbf{U}(\mathbf{X}, Y; \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\}$ has full column rank p. Also, $\mathrm{V}\{\mathbf{U}(\mathbf{X}, Y; \boldsymbol{\theta})\}$ and $\mathrm{E}(\mathbf{h}\mathbf{h}^\top/\pi)$ are positive definite in the neighborhood of $\boldsymbol{\theta}_0$.

(C5) $\pi(\mathbf{x}, y) > d > 0$, $p(\mathbf{x}) = \mathrm{E}\{\pi(\mathbf{x}, y)|\mathbf{x}\} \neq 1$ almost surely.

A.2 Proof of Theorem 3.1

To discuss the asymptotic properties of the EL estimator, we write

(28)
$$\hat{\mathbf{Q}}_1(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{r_i=1} \frac{\pi_i^{-1} \mathbf{U}_i(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top \pi_i^{-1} \{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}},$$

and
(29)
$$\hat{\mathbf{Q}}_2(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{r_i=1} \frac{\pi_i^{-1} \{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}}{1 + \boldsymbol{\lambda}^\top \pi_i^{-1} \{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}}.$$

By using a similar argument of Lemma 1 of Qin and Lawless (1994), it can be shown that $(\hat{\boldsymbol{\theta}}_{h1}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top \to^p (\boldsymbol{\theta}_0^\top, \mathbf{0}^\top)^\top$.

To prove the asymptotic normality of $\hat{\boldsymbol{\theta}}_{h1}$, by (28) and (29), we have

$$\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) = \frac{1}{n} \sum_{r_i=1} \pi_i^{-1} \mathbf{U}_i(\boldsymbol{\theta}_0),$$

$$\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) = \frac{1}{n} \sum_{r_i=1} \pi_i^{-1} \left\{\mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0)\right\},$$

$$\frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{r_i=1} \pi_i^{-1} \frac{\partial \mathbf{U}_i}{\partial \boldsymbol{\theta}},$$

$$\frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}} = -\frac{1}{n} \sum_{r_i=1} \frac{1}{\pi_i^2} \mathbf{U}_i \left\{\mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0)\right\}^\top,$$

$$\frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

$$\frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}} = -\frac{1}{n} \sum_{r_i=1} \frac{1}{\pi_i^2} \left\{\mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0)\right\}^{\otimes 2}.$$

According to conditions (C1)–(C5) and $(\hat{\boldsymbol{\theta}}_{h1}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top \to^p (\boldsymbol{\theta}_0^\top, \mathbf{0}^\top)^\top$, we can apply the standard arguments using Taylor expansion to get

$$\mathbf{0} = \hat{\mathbf{Q}}_1(\hat{\boldsymbol{\theta}}_{h1}, \hat{\boldsymbol{\lambda}}) = \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) + \frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0)$$
$$+ \frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}^\top}(\hat{\boldsymbol{\lambda}} - \mathbf{0}) + \mathbf{o}_p(\delta_n),$$

and

$$\mathbf{0} = \hat{\mathbf{Q}}_2(\hat{\boldsymbol{\theta}}_{h1}, \hat{\boldsymbol{\lambda}}) = \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) + \frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0)$$
$$+ \frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}^\top}(\hat{\boldsymbol{\lambda}} - \mathbf{0}) + \mathbf{o}_p(\delta_n),$$

where $\delta_n = \|\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0\| + \|\hat{\boldsymbol{\lambda}}\|$. Thus, we have

$$\begin{pmatrix} \hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} - \mathbf{0} \end{pmatrix} = -\mathbf{S}_n^{-1} \begin{pmatrix} \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{o}_p(\delta_n) \\ \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{o}_p(\delta_n) \end{pmatrix},$$

where

$$\mathbf{S}_n = \begin{pmatrix} \partial\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})/\partial\boldsymbol{\theta} & \partial\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0})/\partial\boldsymbol{\lambda} \\ \partial\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})/\partial\boldsymbol{\theta} & \partial\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0})/\partial\boldsymbol{\lambda} \end{pmatrix}.$$

Because of the existence of moments, we have

$$\mathbf{S_n} \to^p \begin{pmatrix} \mathrm{E}(\partial\mathbf{U}/\partial\boldsymbol{\theta}) & -\mathrm{E}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi) \\ \mathbf{0} & -\mathrm{E}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi) \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

Since $\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) = n^{-1} \sum_{r_i=1} \pi_i^{-1} \mathbf{U}_i = \mathbf{O}_p(n^{-\frac{1}{2}})$ and $\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) = n^{-1} \sum_{r_i=1} \pi_i^{-1} (\mathbf{h}_i - n^{-1} \sum_{i=1}^n \mathbf{h}_i) = \mathbf{O}_p(n^{-\frac{1}{2}})$, we have $\delta_n = O_p(n^{-\frac{1}{2}})$ and

$$\begin{pmatrix} \hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} - \mathbf{0} \end{pmatrix} = -\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) \\ \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) \end{pmatrix} + \mathbf{o}_p(\delta_n).$$

So, after some algebra, we have

$$\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0 = \mathbf{S}_{11}^{-1} \left\{ -\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}) \right\} + \mathbf{o}_p(n^{-1/2}).$$

Hence, by the existence of second moments, we get

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h1} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h1}),$$

where $\mathbf{V}_{h1} = \boldsymbol{\tau}\boldsymbol{\Omega}_{h1}\boldsymbol{\tau}^\top$ with $\boldsymbol{\tau} = \{\mathrm{E}(\partial\mathbf{U}/\partial\boldsymbol{\theta})\}^{-1}$, $\boldsymbol{\Omega}_{h1} = \mathrm{V}\{r(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})/\pi + \mathbf{B}\tilde{\mathbf{h}}\}$ and $\mathbf{B} = \mathrm{E}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi)\{\mathbf{E}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi)\}^{-1}$.

A.3 Proof of Corollary 3.1

Note that $\boldsymbol{\Omega}_{h1}$ in (16) satisfies

$$\begin{aligned} \boldsymbol{\Omega}_{h1} &= \mathrm{E}\left\{r\mathbf{U}/\pi - \mathbf{B}(r/\pi - 1)\tilde{\mathbf{h}}\right\}^{\otimes 2} \\ &= \mathrm{E}\Big\{r\mathbf{U}/\pi - (r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X}) + (r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X}) \\ &\quad - \mathbf{B}(r/\pi - 1)\tilde{\mathbf{h}}\Big\}^{\otimes 2} \\ &= \mathrm{E}\left\{r\mathbf{U}/\pi - (r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X})\right\}^{\otimes 2} \\ &\quad + \mathrm{E}\Big[\left\{r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathrm{E}(\mathbf{U}|\mathbf{X})\right\} \\ &\quad \times (r\pi^{-1} - 1)\left\{\mathrm{E}(\mathbf{U}^\top|\mathbf{X}) - \tilde{\mathbf{h}}^\top\mathbf{B}^\top\right\}\Big] \\ &\quad + \mathrm{E}\Big[(r\pi^{-1} - 1)\left\{\mathrm{E}(\mathbf{U}|\mathbf{X}) - \mathbf{B}\tilde{\mathbf{h}}\right\} \\ &\quad \times \left\{r\pi^{-1}\mathbf{U}^\top - (r\pi^{-1} - 1)\mathrm{E}(\mathbf{U}^\top|\mathbf{X})\right\}\Big] \\ &\quad + \mathrm{E}\left\{(r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X}) - \mathbf{B}(r/\pi - 1)\tilde{\mathbf{h}}\right\}^{\otimes 2} \end{aligned}$$

$$\begin{aligned} &= \mathrm{E}\left\{r\mathbf{U}/\pi - (r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X})\right\}^{\otimes 2} \\ &\quad + \mathrm{E}\left\{(r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X}) - \mathbf{B}(r/\pi - 1)\tilde{\mathbf{h}}\right\}^{\otimes 2} \\ &\geq \mathrm{E}\left\{r\mathbf{U}/\pi - (r/\pi - 1)\mathrm{E}(\mathbf{U}|\mathbf{X})\right\}^{\otimes 2}, \end{aligned}$$

where the equality is achieved when $\mathbf{h} \propto \mathbf{h}^* = \mathrm{E}(\mathbf{U}|\mathbf{X})$. The asymptotic variance of $\hat{\boldsymbol{\theta}}_{h1}$ achieved at $\mathbf{h} \propto \mathbf{h}^* = \mathrm{E}(\mathbf{U}|\mathbf{X})$ is equal to

$$\mathbf{V}_{h^*} = \boldsymbol{\tau}\left\{\mathrm{E}\left(\frac{\mathbf{U}\mathbf{U}^\top}{\pi}\right) - \mathrm{E}\left(\frac{1 - \pi}{\pi}\mathbf{h}^*\mathbf{U}^\top\right)\right\}\boldsymbol{\tau}^\top.$$

A.4 Proof of Theorem 4.1

To discuss the asymptotic properties of the EL estimator, we write

(30)

$$\begin{aligned} &\hat{\mathbf{Q}}_1(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\phi}) \\ &= \frac{1}{n}\sum_{r_i=1} \frac{\pi_i(\boldsymbol{\phi})^{-1}\mathbf{U}_i(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top\pi_i(\boldsymbol{\phi})^{-1}\{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}}, \end{aligned}$$

and

(31)

$$\begin{aligned} &\hat{\mathbf{Q}}_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\phi}) \\ &= \frac{1}{n}\sum_{r_i=1} \frac{\pi_i(\boldsymbol{\phi})^{-1}\{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}}{1 + \boldsymbol{\lambda}^\top\pi_i(\boldsymbol{\phi})^{-1}\{\mathbf{h}_i(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})\}}. \end{aligned}$$

Hence $(\hat{\boldsymbol{\theta}}_{h2}, \hat{\boldsymbol{\lambda}})$ is the solution defined by equation $\hat{\mathbf{Q}}_1(\hat{\boldsymbol{\theta}}_{h2}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\phi}}) = \mathbf{0}$ and $\hat{\mathbf{Q}}_2(\hat{\boldsymbol{\theta}}_{h2}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\phi}}) = \mathbf{0}$, where $\hat{\boldsymbol{\phi}}$ is defined in (18). By using a similar argument as that for Theorem 3.1, we can prove $(\hat{\boldsymbol{\theta}}_{h2}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top \to^p (\boldsymbol{\theta}_0^\top, \mathbf{0}^\top)^\top$. Next, we want to prove the asymptotic normality of $\hat{\boldsymbol{\theta}}_{h2}$. According to (30) and (31), we have

$$\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0) = \frac{1}{n}\sum_{r_i=1} \pi_i^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0),$$

$$\hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0) = \frac{1}{n}\sum_{r_i=1} \pi_i^{-1}\left\{\mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0)\right\},$$

$$\frac{\partial\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial\boldsymbol{\theta}} = \frac{1}{n}\sum_{r_i=1} \pi_i^{-1}\frac{\partial\mathbf{U}_i}{\partial\boldsymbol{\theta}},$$

$$\frac{\partial\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial\boldsymbol{\lambda}} = -\frac{1}{n}\sum_{r_i=1} \frac{1}{\pi_i^2}\mathbf{U}_i\left\{\mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0)\right\}^\top,$$

$$\frac{\partial\hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial\boldsymbol{\phi}} = -\frac{1}{n}\sum_{r_i=1} \frac{1}{\pi_i^2}\frac{\partial\pi_i}{\partial\boldsymbol{\phi}}\mathbf{U}_i\mathbf{U}_i^\top,$$

$$\frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

$$\frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\lambda}}$$
$$= -\frac{1}{n} \sum_{r_i=1} \frac{1}{\pi_i^2} \left\{ \mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0) \right\}^{\otimes 2},$$
$$\frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}$$
$$= -\frac{1}{n} \sum_{r_i=1} \frac{1}{\pi_i^2} \frac{\partial \pi_i}{\partial \boldsymbol{\phi}} \left\{ \mathbf{h}_i(\boldsymbol{\theta}_0) - n^{-1} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}_0) \right\}^{\otimes 2}.$$

By using Taylor expansion around $(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)$, we have

$$\mathbf{0} = \hat{\mathbf{Q}}_1(\hat{\boldsymbol{\theta}}_{h2}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\phi}})$$
$$= \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0) + \frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0)$$
$$+ \frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\lambda}^\top}(\hat{\boldsymbol{\lambda}} - \mathbf{0}) + \frac{\partial \hat{\mathbf{Q}}_1(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^\top}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$$
$$+ \mathbf{o}_p(\delta_n),$$

and

$$\mathbf{0} = \hat{\mathbf{Q}}_2(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\phi}})$$
$$= \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0) + \frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0)$$
$$+ \frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\lambda}^\top}(\hat{\boldsymbol{\lambda}} - \mathbf{0}) + \frac{\partial \hat{\mathbf{Q}}_2(\boldsymbol{\theta}_0, \mathbf{0}, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^\top}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$$
$$+ \mathbf{o}_p(\delta_n),$$

where $\delta_n = \|\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0\| + \|\hat{\boldsymbol{\lambda}}\| + \|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\|$. By using a similar argument as the proof of Theorem 3.1 and by (18), after some algebra, we have

$$\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0 = -\mathbf{S}_{11}^{-1} n^{-1} \sum_{i=1}^n \left\{ \frac{r_i}{\pi_i} \mathbf{U}_i(\boldsymbol{\theta}_0) - \mathbf{B}(\frac{r_i}{\pi_i} - 1)\tilde{\mathbf{h}}_i(\boldsymbol{\theta}_0) \right.$$
$$\left. - \mathbf{C}\mathbf{b}_i(\boldsymbol{\phi}_0) \right\} + \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{B} = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}$, $\mathbf{C} = \mathrm{E}\{\pi^{-1}(\mathbf{U} - \mathbf{B}\tilde{\mathbf{h}})(\partial \pi / \partial \boldsymbol{\phi})'\}$ and $\mathbf{S}_{11} = \mathrm{E}(\partial \mathbf{U}/\partial \boldsymbol{\theta})$, $\mathbf{S}_{12} = -\mathrm{E}(\mathbf{U}\tilde{\mathbf{h}}^\top/\pi)$ and $\mathbf{S}_{22} = -\mathrm{E}(\tilde{\mathbf{h}}\tilde{\mathbf{h}}^\top/\pi)$. Hence, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h2}),$$

where $\mathbf{V}_{h2} = \mathbf{S}_{11}^{-1} \mathrm{V}\{r\pi^{-1}\mathbf{U} - \mathbf{B}(r\pi^{-1} - 1)\tilde{\mathbf{h}} - \mathbf{C}\mathbf{b}\}\mathbf{S}_{11}^{-1}$.

A.5 Proof of Theorem 4.2

Because $\hat{\boldsymbol{\phi}}_q \to^p \boldsymbol{\phi}_0$, under some moment conditions, by using Taylor expansion, we have

$$\hat{\boldsymbol{\phi}}_q - \boldsymbol{\phi}_0 = \frac{1}{n} \sum_{i=1}^n \mathbf{b}_i(\boldsymbol{\phi}_0) + \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{b}_i(\boldsymbol{\phi}_0) = \mathbf{S}^{-1}(r_i \pi_i^{-1} - 1)\mathbf{q}_i$, and $\mathbf{S} = \mathrm{E}\{\pi^{-1}\mathbf{q}(\partial \pi / \partial \boldsymbol{\phi})^\top\}$. Using the result of Theorem 4.1, we can get (23). So, under the existence of moments, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{h2} - \boldsymbol{\theta}_0) \to^d N(0, \mathbf{V}_{h2}),$$

where $\mathbf{V}_{h2} = \boldsymbol{\tau} \boldsymbol{\Omega}_{h2} \boldsymbol{\tau}^\top$ and $\boldsymbol{\Omega}_{h2} = \mathrm{V}\{r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q})\}$. Using a similar argument as the proof of Corollary 3.1, we have

$$\boldsymbol{\Omega}_{h2} = \mathrm{V}\left\{ r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q}) \right\}$$
$$= \mathrm{E}\left\{ r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q}) \right\}^{\otimes 2}$$
$$= \mathrm{E}\left\{ r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathbf{h}^* + (r\pi^{-1} - 1)\mathbf{h}^* \right.$$
$$\left. - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q}) \right\}^{\otimes 2}$$
$$= \mathrm{E}\left\{ r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathbf{h}^* \right\}^{\otimes 2}$$
$$+ \mathrm{E}\left\{ (r\pi^{-1} - 1)\mathbf{h}^* - (r\pi^{-1} - 1)(\mathbf{B}\tilde{\mathbf{h}} + \mathbf{C}\mathbf{S}^{-1}\mathbf{q}) \right\}^{\otimes 2}$$
$$\geq \mathrm{E}\left\{ r\pi^{-1}\mathbf{U} - (r\pi^{-1} - 1)\mathbf{h}^* \right\}^{\otimes 2}.$$

The equality holds when $\mathbf{h}^* - \mathbf{B}\tilde{\mathbf{h}} - \mathbf{C}\mathbf{S}^{-1}\mathbf{q} = \mathbf{0}$, which implies $\mathbf{q} = \mathbf{S}(\mathbf{C}^\top \mathbf{C})^{-1}\mathbf{C}^\top(\mathbf{h}^* - \mathbf{B}\tilde{\mathbf{h}})$. Hence, the optimal choice of $\mathbf{q}$ is $\boldsymbol{\alpha}^\top \mathbf{q} = \mathbf{h}^* - \mathbf{B}\tilde{\mathbf{h}}$ for some $\boldsymbol{\alpha}$.

A.6 Proof of Theorem 5.1

For simplicity, we assume $r = 1$, $q = 2$ and $d_x = 1$ in the following proof. A similar proof can be obtained for other cases. In order to prove Theorem 5.1, we first prove the following Lemma:

Lemma A.1. *Let $\hat{\pi}_{i,H}$ be the kernel estimator of $\pi(x_i)$ which is defined in (24). For the choice of $h^* = E(U|X)$, where U is the estimating function defined in (1), we have*
(32)
$$\frac{1}{n} \sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i = \frac{1}{n} \sum_{i=1}^n \frac{r_i}{\pi_i}(U_i - h_i^*) + \frac{1}{n} \sum_{i=1}^n h_i^* + o_p(n^{-1/2}).$$

Proof. By using the standard arguments in the kernel smoothing method, we have

$$(33) \quad E\left\{ \frac{1}{n} \sum_{j=1}^n K_H(X_i, X_j) \right\} = f(X_i) + O(H^2)$$

and

$$(34) \quad E\left\{ \frac{1}{n} \sum_{j=1}^n r_j K_H(X_i, X_j) \right\} = \pi(X_i)f(X_i) + O(H^2).$$

According to (33), (34) and by using Taylor expansion, we have

(35)
$$\frac{\sum_{j=1}^n K_H(X_i, X_j)}{\sum_{j=1}^n r_j K_H(X_i, X_j)}$$

$$= \frac{1}{\pi(X_i)} + \frac{1}{\pi(X_i)f(X_i)} \left\{ \frac{1}{n}\sum_{j=1}^n K_H(X_i, X_j) - f(X_i) \right\}$$

$$- \frac{1}{\pi^2(X_i)f(X_i)} \left\{ \frac{1}{n}\sum_{j=1}^n r_j K_H(X_i, X_j) - \pi(X_i)f(X_i) \right\}$$

$$+ O(H^2)$$

$$= \frac{1}{\pi(X_i)} + \frac{1}{n}\sum_{j=1}^n \frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)} \left\{ 1 - \frac{r_j}{\pi(X_i)} \right\} + O(H^2).$$

By (35) and because of $nH^4 \to 0, nH^2 \to \infty$, we have

$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i$$

$$= \frac{1}{n}\sum_{i=1}^n r_i \frac{\sum_{j=1}^n K_H(X_i, X_j)}{\sum_{j=1}^n r_j K_H(X_i, X_j)} U_i$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n r_i U_i \frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}$$

$$\times \left\{ 1 - \frac{r_j}{\pi(X_i)} \right\} + O(H^2)$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n^2 H}\sum_{i=1}^n U_i \frac{K(0)}{\pi(X_i)f(X_i)}$$

$$\times \left\{ r_i - \frac{r_i}{\pi(X_i)} \right\}$$

$$+ \frac{1}{n(n-1)}\sum_{i \neq j} r_i U_i \frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)} \left\{ 1 - \frac{r_j}{\pi(X_i)} \right\}$$

$$+ O(H^2)$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n(n-1)}\sum_{i \neq j} r_i U_i \frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}$$

$$\times \left\{ 1 - \frac{r_j}{\pi(X_i)} \right\} + o_p(n^{-1/2})$$

so,

(36)
$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n(n-1)}\sum_{i \neq j} h(Z_i, Z_j) + o_p(n^{-1/2}),$$

where $Z_i = (X_i, Y_i, r_i)$ and

(37)
$$h(Z_i, Z_j) = \frac{1}{2}\left[r_i U_i \frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)} \left\{ 1 - \frac{r_j}{\pi(X_i)} \right\} \right.$$

$$\left. + r_j U_j \frac{K_H(X_j, X_i)}{\pi(X_j)f(X_j)} \left\{ 1 - \frac{r_i}{\pi(X_j)} \right\} \right]$$

$$\triangleq \frac{1}{2}(\zeta_{ij} + \zeta_{ji}).$$

According to (36) and (37), we know that $\sum_{i \neq j} h(Z_i, Z_j)/\{n(n-1)\}$ is the U-statistics. Let $s = (X_j - X_i)/H$, by $nH^2 \to \infty, nH^4 \to 0$ and according to Taylor expansion, we have

(38) $E(\zeta_{ij}|Z_i)$
$$= \frac{r_i U_i}{\pi(X_i)f(X_i)} \frac{1}{H} \int K\left(\frac{X_j - X_i}{H}\right) \left\{ 1 - \frac{\pi(X_j)}{\pi(X_i)} \right\}$$

$$\times f(X_j)dX_j$$

$$= \frac{r_i U_i}{\pi(X_i)f(X_i)} \int K(s) \left\{ 1 - \frac{\pi(X_i + Hs)}{\pi(X_i)} \right\}$$

$$\times f(X_i + Hs)ds$$

$$= O(H^2),$$

and

(39) $E(\zeta_{ji}|Z_i)$
$$= \frac{1}{H} \int \frac{U_j}{f(X_j)} K\left(\frac{X_j - X_i}{H}\right) \left\{ 1 - \frac{r_i}{\pi(X_j)} \right\}$$

$$\times f(X_j, Y_j)dX_j dY_j$$

$$= \int \frac{U_j}{f(X_i + Hs)} K(s) \left\{ 1 - \frac{r_i}{\pi(X_i + Hs)} \right\}$$

$$\times f(X_i + Hs, Y_j)ds dY_j$$

$$= \left\{ 1 - \frac{r_i}{\pi(X_i)} \right\} h_i^* + O(H^2).$$

According to (36), (37), (38), (39) and by the theory of U-statistics, see Serfling (1980), Chapter 5, we have

$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i = \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n}\sum_{i=1}^n \left\{ 1 - \frac{r_i}{\pi(X_i)} \right\} h_i^*$$

$$+ o_p(n^{-1/2}). \qquad \square$$

Similarly to Lemma A.1, it can be shown that

(40)
$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} \tilde{h}_i = \frac{1}{n}\sum_{i=1}^n \tilde{h}_i + o_p(n^{-1/2}).$$

By the same argument for proof of Theorem 4.1, under certain conditions, it can be shown that $\hat{\theta}_{h3} \to^p \theta_0$, and the asymptotic normality property:

$$\hat{\theta}_{h3} - \theta_0$$

$$= S_{11}^{-1} \left\{ -\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i(\theta_0) + B\frac{1}{n}\sum_{i=1}^n \left(\frac{r_i}{\hat{\pi}_{i,H}} - 1\right)\tilde{h}_i(\theta_0) \right\}$$

$$+ o_p(n^{-1/2}),$$

where B and S_{11} are defined in the proof of Theorem 4.1. By using (32) and (40), we have

$$\hat{\theta}_{h3} - \theta_0 = S_{11}^{-1}\left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}U_i(\theta_0) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{\pi_i}-1\right)h_i^*(\theta_0)\right\}$$
$$+ o_p(n^{-1/2}).$$

A.7 Proof of Theorem 5.2

Let

$$\hat{\mathbf{Q}}_3(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{r_i=1}\frac{\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta})}.$$

We can write

$$R_n(\boldsymbol{\theta}_0) = 2\Bigg[\sum_{r_i=1}\log\left\{1 + \boldsymbol{\lambda}_0^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0)\right\}$$
$$- \sum_{r_i=1}\log\left\{1 + \hat{\boldsymbol{\lambda}}^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\hat{\boldsymbol{\theta}}_{h3})\right\}\Bigg],$$

where $\boldsymbol{\lambda}_0$ is the solution of $\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) = \mathbf{0}$, by using a similar argument from Theorem 3.1, we have $\boldsymbol{\lambda}_0 \to^p \mathbf{0}$. Thus, by using Taylor expansion around $\mathbf{0}$, we have

$$\mathbf{0} = \hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0) = \hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) + \frac{\partial\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0})}{\partial\boldsymbol{\lambda}_0}\boldsymbol{\lambda}_0 + \mathbf{o}_p(\|\boldsymbol{\lambda}_0\|).$$

According to Lemma A.1 in the proof of Theorem 5.1, we have $\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) = n^{-1}\sum_{r_i=1}\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0) = \mathbf{O}_p(n^{-1/2})$, so $\|\boldsymbol{\lambda}_0\| = \mathbf{O}_p(n^{-1/2})$. Hence, we have

$$(41) \qquad \boldsymbol{\lambda}_0 = -\mathbf{S}_{11}^{*-1}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{o}_p(n^{-1/2}),$$

where $\mathbf{S}_{11}^* = -\mathrm{E}(\mathbf{U}\mathbf{U}^\top/\pi)$. Also, by using Taylor expansion around $\boldsymbol{\lambda}_0 = \mathbf{0}$, we have

$$(42) \qquad 2\sum_{r_i=1}\log\left\{1 + \boldsymbol{\lambda}_0^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0)\right\}$$
$$= 2\sum_{r_i=1}\boldsymbol{\lambda}_0^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0)$$
$$- \sum_{r_i=1}\boldsymbol{\lambda}_0^\top\hat{\pi}_{i,H}^{-2}\mathbf{U}_i(\boldsymbol{\theta}_0)\mathbf{U}_i^\top(\boldsymbol{\theta}_0)\boldsymbol{\lambda}_0 + \mathbf{o}_p(1).$$

By the existence of moments, we have
$$(43)$$
$$\partial\hat{\mathbf{Q}}(\boldsymbol{\theta}_0, \mathbf{0})/\partial\boldsymbol{\lambda} = -n^{-1}\sum_{r_i=1}\hat{\pi}_{i,H}^{-2}\mathbf{U}_i(\boldsymbol{\theta}_0)\mathbf{U}_i^\top(\boldsymbol{\theta}_0) \to^p \mathbf{S}_{11}^*.$$

By plugging (41) into (42) and according to (43), we have

$$(44) \qquad 2\sum_{r_i=1}\log\left\{1 + \boldsymbol{\lambda}_0^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\boldsymbol{\theta}_0)\right\}$$
$$= -n\hat{\mathbf{Q}}_3^\top(\boldsymbol{\theta}_0, \mathbf{0})\mathbf{S}_{11}^{*-1}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{o}_p(1).$$

Similarly, by the same argument of Qin and Lawless (1994) and by using Taylor expansion around $\hat{\boldsymbol{\lambda}} = \mathbf{0}$, we have

$$(45) \qquad 2\sum_{r_i=1}\log\left\{1 + \hat{\boldsymbol{\lambda}}^\top\hat{\pi}_{i,H}^{-1}\mathbf{U}_i(\hat{\boldsymbol{\theta}})\right\}$$
$$= -n\hat{\mathbf{Q}}_3^\top(\boldsymbol{\theta}_0, \mathbf{0})\mathbf{S}_{11}^{*-1}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0})$$
$$+ n\hat{\mathbf{Q}}_3^\top(\boldsymbol{\theta}_0, \mathbf{0})\mathbf{S}_{11}^{*-1}\mathbf{S}_{12}^*\mathbf{S}_{22.1}^{*-1}\mathbf{S}_{21}^*\mathbf{S}_{11}^{*-1}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0})$$
$$+ \mathbf{o}_p(1),$$

where $\mathbf{S}_{12}^* = \mathrm{E}\{\partial\mathbf{U}(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\}, \mathbf{S}_{21}^* = \mathbf{S}_{12}^{*\top}$, and $\mathbf{S}_{22.1}^* = \mathbf{S}_{21}^*\mathbf{S}_{11}^{*-1}\mathbf{S}_{12}^*$. So, by (44) and (45), we have

$$R_n(\boldsymbol{\theta}_0) = -n\hat{\mathbf{Q}}_3^\top(\boldsymbol{\theta}_0, \mathbf{0})\mathbf{S}_{11}^{*-1}\mathbf{S}_{12}^*\mathbf{S}_{22.1}^{*-1}\mathbf{S}_{21}^*\mathbf{S}_{11}^{*-1}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0})$$
$$+ \mathbf{o}_p(1).$$

In addition, by the existence of moments and $\mathbf{h}^* = \mathrm{E}\{\mathbf{U}(\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$, we have

$$\sqrt{n}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) \to^d N(0, \mathbf{V}_{Q_3}),$$

where $\mathbf{V}_{Q_3} = \mathrm{E}(\mathbf{U}\mathbf{U}^\top/\pi) = -\mathbf{S}_{11}^*$. Hence, we have

$$R_n(\boldsymbol{\theta}_0) = -\sqrt{n}\hat{\mathbf{Q}}_3^\top(\boldsymbol{\theta}_0, \mathbf{0})\mathbf{S}_{11}^{*-1/2}\mathbf{S}_{11}^{*-1/2}\mathbf{S}_{12}^*\mathbf{S}_{22.1}^{*-1}\mathbf{S}_{21}^*\mathbf{S}_{11}^{*-1/2}$$
$$\times \sqrt{n}\mathbf{S}_{11}^{*-1/2}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) + \mathbf{o}_p(1).$$

Because $-\sqrt{n}\mathbf{S}_{11}^{*-1/2}\hat{\mathbf{Q}}_3(\boldsymbol{\theta}_0, \mathbf{0}) \to^d N(0, \mathbf{I})$, and $-\mathbf{S}_{11}^{*-1/2}\mathbf{S}_{12}^*\mathbf{S}_{22.1}^{*-1}\mathbf{S}_{21}^*\mathbf{S}_{11}^{*-1/2}$ is an idempotent matrix with trace p, we have $R_n(\boldsymbol{\theta}_0) \to^d \chi_p^2$.

REFERENCES

CATTANEO, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J. Econometrics* **155** 138–154. MR2607191

CHANG, T. and KOTT, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95** 555–571. MR2443175

CHAUDHURI S., HANDCOCK, M. S., and RENDALL, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B Stat. Methodol* **70**, 311–328. MR2424755

CHEN, S. X., LEUNG, D. H. Y., and QIN, J. (2008). Improving semi-parametric estimation by using surrogate data. *J. R. Stat. Soc. Ser. B Stat. Methodol* **70**, 803–823. MR2523905

CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89**, 81–87.

DA SILVA, D. N. and OPSOMER, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Surv. Methodol.* **35**, 165–176.

DURRANT, G. B. and SKINNER, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Surv. Methodol.* **32**, 25–36.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189. MR1995826

KIM, J. K. and KIM, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.* **35**, 501–514. MR2381396

KIM, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statist. Sinica* **19**, 145–157. MR2487882

KIM, J. K. and YU, C. L. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data. *J. Amer. Statist. Assoc.* **106**, 157–165. MR2816710

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249. MR0946049

OWEN, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120. MR1041387

OWEN, A. B. (2001). *Empirical Likelihood*. New York: Chapman and Hall/CRC.

QIN, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.* **21**, 1182–1196. MR1241264

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–325. MR1272085

QIN, J., LEUNG, D., and SHAO, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.* **97**, 193–200. MR1947279

QIN, J. and ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Stat. Soc. Ser. B Stat. Methodol* **69**, 101–122. MR2301502

QIN, J., ZHANG, B., and LEUNG D. (2009). Empirical likelihood in missing data problems. *J. Amer. Statist. Assoc.* **104**, 1492–1503. MR2750574

ROBINS, J. M., ROTNITZKY, A., and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866. MR1294730

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592. MR0455196

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley. MR0595165

TAN, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika* **98**, 663–684. MR2836413

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203. MR0773161

WANG, Q. and RAO, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics* **30**, 896–924. MR1922545

WANG, D. and CHEN, S. X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.* **37**, 490–517. MR2488360

XUE, L. (2009). Empirical likelihood for linear models with missing responses. *J. Multivariate Anal.* **100**, 1353–1366. MR2514134

Sixia Chen

Westat

1600 Research blvd

Rockville, MD, 20850

USA

E-mail address: SixiaChen@westat.com