Semiparametric Efficient and Robust Estimation of an Unknown Symmetric Population Under Arbitrary Sample Selection Bias

Yanyuan MA, Mijeong KIM, and Marc G. GENTON

We propose semiparametric methods to estimate the center and shape of a symmetric population when a representative sample of the population is unavailable due to selection bias. We allow an arbitrary sample selection mechanism determined by the data collection procedure, and we do not impose any parametric form on the population distribution. Under this general framework, we construct a family of consistent estimators of the center that is robust to population model misspecification, and we identify the efficient member that reaches the minimum possible estimation variance. The asymptotic properties and finite sample performance of the estimation and inference procedures are illustrated through theoretical analysis and simulations. A data example is also provided to illustrate the usefulness of the methods in practice.

KEY WORDS: Efficiency; Nonrandom data; Robustness; Semiparametric model; Skewness; Symmetric distribution.

1. INTRODUCTION

Estimating the center of a population is probably one of the most elementary problems in statistics. When the population is symmetric, the center can be equivalently represented by the population mean or population median, and their estimators and corresponding statistical properties are well understood. However, these familiar methods are based on a key assumption, namely that we observe a random representative sample from the population. When the sampling procedure involves some selection mechanism, that is, when the sample obtained is no longer a representative sample of the original population, the problem of estimating the population center is no longer so simple.

More specifically, let X be a random variable that is symmetrically distributed in a population with center μ, which we want to estimate. Assume that a representative sample from this population is not obtained due to various reasons. Instead, only a biased sample from a specific data collection procedure is available. Let the observed biased sample be $X_1, \ldots, X_n$, where the X_i's are independent and identically distributed (iid). Then, we can write the probability density function (pdf) of one observation as

$$g(x;\mu,\boldsymbol{\beta},f) = c(\boldsymbol{\beta})f(x-\mu)w(x-\mu;\boldsymbol{\beta}) = \frac{f(x-\mu)w(x-\mu;\boldsymbol{\beta})}{\int f(t)w(t;\boldsymbol{\beta})dt}, \quad (1)$$

where we use w to capture the selection mechanism, and we use f to denote the original symmetric yet unspecified population pdf of X. Here $c(\boldsymbol{\beta}) = 1/\int f(t)w(t;\boldsymbol{\beta})dt$ is a normalizing constant. Note that in Equation (1), other than being even, the specific form of f is not known. Thus, f can be, for example, a Normal or Student's t pdf or any other symmetric pdf. The sampling bias is described by the multiplicative factor w, which essentially reweights the observation by taking into account the effect of the data collection procedure. The functional form of w is completely decided by the selection process and is not subject to any artifactual restrictions. We consider the situation where w is a function of the centered data $x-\mu$ instead of the uncentered data x, because otherwise, the biased sample from the selection procedure can be used directly as if no sampling bias existed in estimating μ. Considering that the selection mechanism may also contain some aspects that are not known in advance, we allow for an additional unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p-1}$ in the selection function w. Finally, to avoid imposing additional constrains on w, we incorporate a normalizing constant $c(\boldsymbol{\beta})$ in Equation (1). If desired, one can also view $c(\boldsymbol{\beta})w(x-\mu;\boldsymbol{\beta}) = w(x-\mu;\boldsymbol{\beta})/\int f(t)w(t;\boldsymbol{\beta})dt$ as a weight function.

Selection bias issues have been acknowledged and modeled extensively (see, e.g., Rao 1985, and more recently Arellano-Valle, Branco, and Genton 2006). The biased sample model in Equation (1) represents one of the most general situations for such models. When special restrictions are further imposed on either the population model f or the selection function w, it reduces to various special models in the literature. For example, when w satisfies an antisymmetric property, $w(x-\mu;\boldsymbol{\beta}) + w(\mu-x;\boldsymbol{\beta}) = 1$ for all $x \in \mathbb{R}$, Copas and Li (1997), Arnold and Beaver (2002), Azzalini and Capitanio (2003), Ma and Genton (2004), Wang, Boyer, and Genton (2004), Arellano-Valle and Genton (2007, 2008) and many others have described the types of selection mechanisms that lead to Equation (1). When f is further assumed to belong to the elliptical family, Equation (1) is reduced to the generalized skew-elliptical distributions (Genton and Loperfido 2005), which include the well-known skew-normal distribution introduced by Azzalini (1985; see the edited book by Genton 2004, and the review by Azzalini 2005, and references therein, for

Yanyuan Ma, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *ma@stat.tamu.edu*). Mijeong Kim, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *mjkim@stat.tamu.edu*). Marc G. Genton is Professor, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia (E-mail: *marc.genton@kaust.edu.sa*). This research was partially supported by NSF grants DMS-0906341, DMS-1007504, and DMS-1100492; NINDS grant R01-NS073671; and by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).

Journal of the American Statistical Association
September 2013, Vol. 108, No. 503, Theory and Methods
DOI: 10.1080/01621459.2013.816184

further details). Other special cases of Equation (1) that do not satisfy the antisymmetric property of w include extended skew-elliptical distributions (Arellano-Valle and Genton 2010a) and their specific members such as extended skew-t distributions (Arellano-Valle and Genton 2010b) and extended skew-normal distributions (Azzalini 1985). The link between extended skew-elliptical distributions and Heckman-type selection models has been recently described by Marchenko and Genton (2012).

A familiar example of samples subject to selection bias was given by Cameron and Trivedi (2010), where they considered a dataset of ambulatory expenditures from the 2001 Medical Expenditure Panel Survey. Because the patients' decision to use the ambulatory service is related to the potential medical cost, the data contained in the survey form a biased sample due to the hidden selection process. Cameron and Trivedi (2010) assumed a normal distribution of the ambulatory expenditures had there been no selection process, and they further modeled the selection process from a normal distribution as well. Their formulation corresponds to assuming f to be normal in Equation (1) and w to be a normal cumulative distribution function (cdf). By relaxing the normality assumption on both f and w, the ambulatory expenditure data can be more flexibly described by a less restrictive model (1). Intuitively, one can consider the potential ambulatory expense X, distributed as $f(X-\mu)$, and the alternative medical cost Y, distributed as $h(Y)$ with cdf H. In practice, a patient or his/her relative would decide to use the ambulatory service if the benefit associated with Y is smaller than the benefit associated with X, that is, $b_Y(Y) \le b_X(X-\mu)$, where b_X, b_Y denote the corresponding benefit functions associated with the two expenditures. This can be described by $w(X-\mu) = \mathrm{pr}\{Y \le a + b(X-\mu)\} = H\{a + b(X-\mu)\}$ if pure benefit is considered, where a, b capture the joint effect of typical deductible and copay associated with an insurance policy, or a more general form $w(X-\mu) = \mathrm{pr}\{b_Y(Y) \le b_X(X-\mu)\}$. Thus, the observed ambulatory expenditures included in the survey are not a representative random sample from f, but a biased one that has a weighted form given in Equation (1). Estimating and studying the corresponding inference on μ using the biased sample is the main purpose of this article.

Next, we demonstrate that the wrong parametric model for f can lead to serious bias on the estimate of the center μ. Consider a random sample $X_1, \ldots, X_n$ from a skew-normal distribution (Azzalini 1985), a particular version of model (1), that is, from a distribution with probability density

$$2\phi(x)\Phi(\beta x), \tag{2}$$

which has center $\mu = 0$. We can estimate μ with the maximum likelihood estimator (MLE) based on the skew-normal parametric model (2). Now suppose that the random sample is in fact from a distribution with probability density

$$2t_5(x)\Phi(\beta x), \tag{3}$$

which also has center $\mu = 0$, but with original symmetric population density $f = t_5$, a Student's t with 5 degrees of freedom. In this case, the MLE of μ based on the skew-normal parametric model (2) is biased as we illustrate next with a simulation experiment. We simulate a 1000 random samples from each of the two data-generating mechanisms (2) and (3) with $n = 400$ and $\beta = 3$. In both cases, we estimate μ based on the skew-normal MLE. Figure 1 presents the boxplots of the 1000 estimates of μ in each of the two cases. It is apparent that when f is misspecified to be $f_0 = \phi$ instead of the correct $f = t_5$, then the estimates of μ are biased. This motivates the development of semiparametric estimators of μ that do not rely on a specific shape of the symmetric density f.

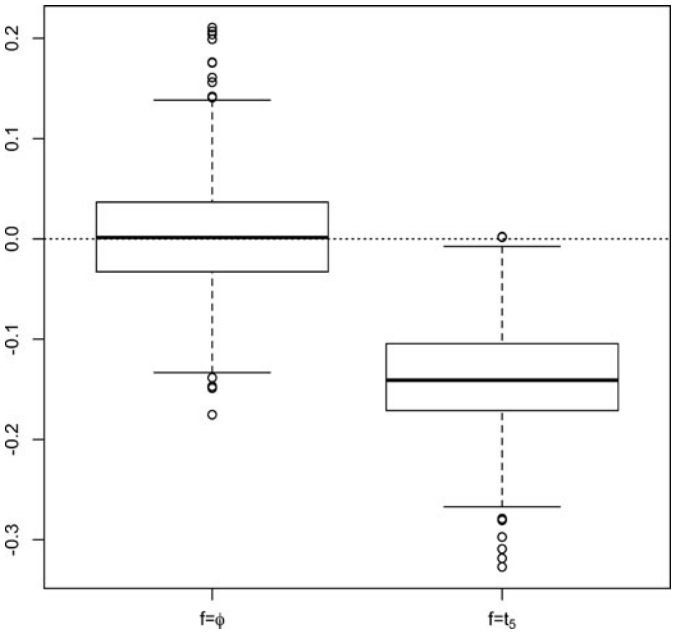

Figure 1. Boxplots of 1000 MLEs of the center μ based on a skew-normal parametric model. Left: correct model when $f = \phi$. Right: incorrect model when $f = t_5$. True value $\mu = 0$ is indicated by the horizontal dotted line.

We organize the rest of the article as follows. In Section 2, we construct a class of consistent estimators of μ that are general and robust to model misspecification on f using a semiparametric approach. We further consider the estimation efficiency issue and construct the semiparametric efficient member of this class by incorporating nonparametric estimation procedures of the shape of f in Section 3. The asymptotic properties of both the consistent and efficient estimators are derived in Section 4. We conduct numerical experiments via simulations and the ambulatory expenditure data analysis in Section 5. We finish the article with a discussion in Section 6. Technical details are collected in the Appendix.

2. CONSISTENT ESTIMATION UNDER MISSPECIFIED F

2.1 The Estimator Family

Model (1) contains several unknown quantities, including the parameter of our central interest μ, the additional parameters $\boldsymbol{\beta} \in \mathbb{R}^{p-1}$ related to the selection process, and the unspecified symmetric density function f. Writing $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ as the finite dimensional parameter, and treating the unknown symmetric density function f as an infinite dimensional nuisance parameter, we can consider Equation (1) as a semiparametric model. Here, although our essential interest is only in μ, we decide to include $\boldsymbol{\beta}$ as part of the parameter vector to estimate instead of treating it as part of the nuisance parameters. This is

because estimating $\boldsymbol{\beta}$ along with μ does not impose too much complexity, and we have the additional benefit of obtaining the estimator of $\boldsymbol{\beta}$ as a byproduct.

Although $X_1, \ldots, X_n$ do not form an iid sample from $f(X-\mu)$, once the selection mechanism is taken into account, they are iid observations with pdf (1). Thus, semiparametric methods described by Bickel et al. (1993) and Tsiatis (2006) become applicable. The central result of the semiparametric approach is to describe the consistent estimators via a nuisance tangent space orthogonal complement $\Lambda^{\perp}$, and to understand the asymptotic properties of the estimators through their matching members in $\Lambda^{\perp}$. For model (1), we explicitly derived in the Appendix that

$$\Lambda^{\perp} = \{\mathbf{v}(X-\mu) : \mathbf{v}(z)w(z;\boldsymbol{\beta}) + \mathbf{v}(-z)w(-z;\boldsymbol{\beta}) = \mathbf{0}\ a.s., \mathbf{v} \in \mathbb{R}^p\}.$$

In the Appendix and throughout the rest of the article, we use a subindex $_0$ to denote the true values of the parameters or the true functions, and write the projection of a function $\mathbf{h}$ onto a space A as $\Pi(\mathbf{h}|A)$ and let $\mathbf{c}^{\otimes 2} = \mathbf{c}\mathbf{c}^{\mathrm{T}}$ for any vector or matrix $\mathbf{c}$.

Members of the space $\Lambda^{\perp}$ can be used to construct estimating equations, and the resulting estimator that solves the corresponding estimating equation has its influence function being the normalized version of this member. Here, it is of interest to consider the special situation when $w = 1$. This corresponds to the classical representative random sample case when there is no selection bias issue. In this case, $\Lambda^{\perp} = \{v(X-\mu) : v(z) + v(-z) = 0 \text{ a.s.}\}$. We can easily see that by choosing $v(X-\mu) = X-\mu$, we obtain the sample mean estimator as the center estimator, and by choosing $v(X-\mu) = \mathrm{sign}(X-\mu)$, we obtain the sample median estimator. Both estimators are consistent under the symmetry assumption, and the fact that the median is more robust to outliers than the mean is reflected in that $\mathrm{sign}(X-\mu)$ is a bounded function, while $X-\mu$ is not. Comparing the general case with an arbitrary w and the special case of $w = 1$, we can view the criterion in $\Lambda^{\perp}$ as a tilted version of the antisymmetric requirement of $v(z) + v(-z) = 0$.

2.2 Locally Efficient Estimators and Their Robustness

The form of $\Lambda^{\perp}$ allows a large selection of the function $\mathbf{v}$. For example, taking any p-component odd function of z and dividing it by $w(z;\boldsymbol{\beta})$ yields a valid $\mathbf{v}$. With this vast amount of choices, we further scale down the problem to investigate a class of estimators that has the potential of reaching asymptotic efficiency, yet is robust against possible model misspecification regarding f. Our approach is through deriving the efficient score, which is the orthogonal projection of the score function onto the space $\Lambda^{\perp}$. The score function, denoted $\mathbf{S}_{\boldsymbol{\theta}}$, is defined as $\partial \log g(x;\boldsymbol{\theta}, f)/\partial\boldsymbol{\theta}$, which has the explicit form

$$\mathbf{S}_{\boldsymbol{\theta}} = \begin{pmatrix} S_{\mu} \\ \mathbf{S}_{\boldsymbol{\beta}} \end{pmatrix} = \left\{ -\frac{f_0'(x-\mu)}{f_0(x-\mu)} - \frac{w'(x-\mu;\boldsymbol{\beta})}{w(x-\mu;\boldsymbol{\beta})}, \frac{\mathbf{w}_{\beta}(x-\mu;\boldsymbol{\beta})^{\mathrm{T}}}{w(x-\mu;\boldsymbol{\beta})} - \frac{\int f_0(t)\mathbf{w}_{\beta}(t;\boldsymbol{\beta})^{\mathrm{T}}dt}{\int f_0(t)w(t;\boldsymbol{\beta})dt} \right\}^{\mathrm{T}},$$

where we write $w'(\cdot;\boldsymbol{\beta}) = \partial w(x;\boldsymbol{\beta})/\partial x|_{x=\cdot}$ and $\mathbf{w}_{\boldsymbol{\beta}}(\cdot;\boldsymbol{\beta}) = \partial w(x;\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{x=\cdot}$. Further projecting $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$, we show in the Appendix that the efficient score $\mathbf{S}_{\mathrm{eff}} = \Pi(\mathbf{S}_{\boldsymbol{\theta}}|\Lambda^{\perp})$ is

$$\begin{aligned}\mathbf{S}_{\mathrm{eff}}(x;\boldsymbol{\theta}, f_0) &= \left\{ \frac{-2f_0'(x-\mu)w(-x+\mu;\boldsymbol{\beta})}{f_0(x-\mu)\{w(x-\mu;\boldsymbol{\beta}) + w(-x+\mu;\boldsymbol{\beta})\}} \right. \\ &\quad + \frac{w'(x-\mu;\boldsymbol{\beta}) + w'(-x+\mu;\boldsymbol{\beta})}{w(x-\mu;\boldsymbol{\beta}) + w(-x+\mu;\boldsymbol{\beta})} - \frac{w'(x-\mu;\boldsymbol{\beta})}{w(x-\mu;\boldsymbol{\beta})} \\ &\quad \left. - \frac{\mathbf{w}_{\beta}(x-\mu;\boldsymbol{\beta}) + \mathbf{w}_{\beta}(-x+\mu;\boldsymbol{\beta})}{w(x-\mu;\boldsymbol{\beta}) + w(-x+\mu;\boldsymbol{\beta})} + \frac{\mathbf{w}_{\beta}(x-\mu;\boldsymbol{\beta})}{w(x-\mu;\boldsymbol{\beta})} \right\}.\end{aligned}$$

Because our goal is to search for locally efficient estimators that are robust to model misspecification, we borrow the form of the efficient score and propose the following estimation procedure. We first postulate a density model for X that is symmetric around a center μ. We write this model $f^*(X-\mu)$. Of course f^* may not reflect the true distribution of X, hence we do not need to have $f^*(t) = f_0(t)$. We then estimate $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ through solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{S}_{\mathrm{eff}}(X_i;\boldsymbol{\theta}, f^*) = \mathbf{0}. \tag{4}$$

Obviously, if we postulate a correct model, that is, if $f^*(t) = f_0(t)$, then the above estimating equation yields the efficient estimator, hence we achieve the optimal efficiency. This is why the estimator is named "locally efficient." On the other hand, if the postulated model is incorrect, that is, if $f^*(t) \neq f_0(t)$, we find that the last $p-1$ components of the difference $\mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}, f^*) - \mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}, f_0)$ is zero, and the first component satisfies

$$\begin{aligned}&E\left\{\mathbf{e}_1^{\mathrm{T}}\mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}, f^*) - \mathbf{e}_1^{\mathrm{T}}\mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}, f_0)\right\} \\ &\quad = c(\boldsymbol{\beta}) \int \frac{2\{f_0'(t)f^*(t) - f^{*'}(t)f_0(t)\}w(t;\boldsymbol{\beta})w(-t;\boldsymbol{\beta})}{f^*(t)\{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})\}}dt,\end{aligned}$$

where $\mathbf{e}_1$ is a length p vector with 1 in the first component and zero everywhere else. Because f_0, f^* are even functions, f_0', $f^{*'}$ are odd functions. Hence, the above integrand is an odd function, and the expectation is therefore zero. Thus, we have found that $E\{\mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}, f^*)\} = \mathbf{0}$ regardless of the choice of f^*. In other words, the estimator obtained from Equation (4) has an additional robustness property, in that even if the model for f is misspecified, the resulting estimator is still consistent.

3. EFFICIENCY CONSIDERATIONS

3.1 Improving the Estimation Efficiency

Different choices in postulating the model f^* provide many different consistent estimators for $\boldsymbol{\theta}$. In practice, a natural question to ask is which f^* is the best choice? From the estimation variability point of view, postulating $f^* = f_0$ is certainly the optimal choice because then we can obtain the efficient estimator. However, it requires extremely good luck to happen to have $f^* = f_0$. Thus, one might need to compromise between optimality and feasibility, and look to improve the estimation efficiency in a class of possible models of f^*. One convenient way is to index the class by a parameter $\boldsymbol{\gamma}$, which can be a vector, and postulate $f^*(x-\mu;\boldsymbol{\gamma})$ as a model family instead of one fixed model. For example, one may postulate a normal model with mean μ, while leaving the variance undecided. In this case,

γ is the variance. Or, one may postulate a Student's t distribution family with mean μ, while leaving both the variance and degrees of freedom unspecified. In this case, $\boldsymbol{\gamma}$ contains both the variance and the degrees of freedom.

Of course, the unspecified parameter $\boldsymbol{\gamma}$ also needs to be estimated. To this end, we can calculate the score with respect to $\boldsymbol{\gamma}$ to obtain the nuisance score vector

$$\mathbf{S}_{\boldsymbol{\gamma}}(x;\boldsymbol{\theta},\boldsymbol{\gamma},f^*) = \frac{\partial \log g(x;\boldsymbol{\theta},\boldsymbol{\gamma},f^*)}{\partial \boldsymbol{\gamma}} = \frac{\partial f^*(x-\mu;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}}{f^*(x-\mu;\boldsymbol{\gamma})} - \frac{\int \partial f^*(t;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\, w(t;\boldsymbol{\beta})dt}{\int f^*(t;\boldsymbol{\gamma})w(t;\boldsymbol{\beta})dt}.$$

We can then augment Equation (4) with $\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\gamma}}(X_i;\boldsymbol{\theta},\boldsymbol{\gamma},f^*) = \mathbf{0}$ to form the extended estimating equation to solve for $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\theta}}$ jointly. The final estimator based on the partially postulated model $f^*(x-\mu;\boldsymbol{\gamma})$ certainly retains the robustness property, in that even if the postulated family does not contain the true pdf $f_0(x-\mu)$ as its member, the consistency is still retained. The comparative benefit with respect to a fully postulated model $f^*(x-\mu)$ is that we only need the family to contain $f_0(x-\mu)$ to achieve the optimal efficiency.

An additional remark we would like to make regarding the postulated family of models is about the estimation of $\boldsymbol{\gamma}$. Specifically, the uncertainty of the postulated model represented by the additional parameter $\boldsymbol{\gamma}$ and the subsequent estimation of $\boldsymbol{\gamma}$ do not incur a price to pay regarding estimating μ or $\boldsymbol{\theta}$. In other words, if we had used a completely determined model $f^*(x-\mu;\boldsymbol{\gamma}_0)$ and proceeded to obtain the estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}_0}$, versus if we had used a partially specified model $f^*(x-\mu;\boldsymbol{\gamma})$ and proceeded to estimate $\boldsymbol{\gamma}$ to obtain $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\theta}}_{\widehat{\boldsymbol{\gamma}}}$, the estimation variabilities of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}_0}$ and $\widehat{\boldsymbol{\theta}}_{\widehat{\boldsymbol{\gamma}}}$ are the same asymptotically. This property will be studied more carefully in Section 4.

3.2 Efficient Estimation of μ

When we postulate a family $f^*(x-\mu;\boldsymbol{\gamma})$ instead of one single function $f^*(x-\mu)$, we have a better chance of capturing the true $f_0(x-\mu)$ hence a better chance of achieving efficiency. Likewise, when we increase the flexibility of the family of $f^*(x-\mu;\boldsymbol{\gamma})$, our chance of achieving the efficiency further increases. Thus, naturally, if we can find a most flexible family so that it has the best chance of including $f_0(x-\mu)$, then the chance of achieving optimal efficiency will also be maximized. This most flexible way of postulating a family turns out to be the nonparametric modeling. Using a properly constructed nonparametric estimator of $f_0(x-\mu)$, we can indeed reach the optimal efficiency. Specifically, we recommend to estimate the function $f(t)$ through a refined kernel density estimator that takes advantage of the symmetry of $f(t)$. The explicit form of the refined kernel estimator we propose is

$$\widetilde{f}(t;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \frac{K_h(X_i-\mu-t)+K_h(X_i-\mu+t)}{w(t;\boldsymbol{\beta})+w(-t;\boldsymbol{\beta})}, \quad (5)$$

where $K_h(t) = K(t/h)/h$, K is a kernel function, and h is a bandwidth. The form of the estimator in Equation (5) guarantees that $\widetilde{f}(t;\boldsymbol{\theta})$ is indeed symmetric. However, $\widetilde{f}(t;\boldsymbol{\theta})$ does not necessarily integrate to 1, hence it may not be a valid pdf estimator. Nevertheless, a closer look at the efficient score reveals that f_0 (or replaced with f^*) and its derivative appear, respectively, on the numerator and denominator in $\mathbf{S}_{\text{eff}}$ simultaneously; hence, the normalizing constant in front of $\widetilde{f}(t;\boldsymbol{\theta})$ does not have any impact on the final estimator for $\boldsymbol{\theta}$. On the other hand, we would like to point out that although $f_0(t)$ does not rely on $\boldsymbol{\theta}$, our refined nonparametric kernel estimator does involve $\boldsymbol{\theta}$. This implies that a profile type of estimator is needed in our final construction. Specifically, our algorithm for the efficient estimator is the following:

Step 1. Choose a symmetric density function f^*. Obtain $\widetilde{\boldsymbol{\theta}}$ through solving Equation (4).
Step 2. Obtain $\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})$ from Equation (5).
Step 3. Obtain $\widehat{\boldsymbol{\theta}}$ through solving Equation (4) with f^* replaced by $\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})$ obtained in Step 2.

We point out that in the above Step 3, $\widetilde{\boldsymbol{\theta}}$ is known and it is $\widetilde{\boldsymbol{\theta}}$ that appears inside the $\widetilde{f}$ function, not $\boldsymbol{\theta}$. Hence, in terms of solving Equation (4) in Step 3, it is completely equivalent to the estimating equation solving procedure in Step 1. Thus, the above three-steps procedure is much simpler than the conventional profile procedure. Of course, if we wish, we can choose to iterate Steps 2 and 3 using the most recently obtained $\boldsymbol{\theta}$ estimate to replace $\widetilde{\boldsymbol{\theta}}$. Such an iterative procedure falls into the conventional profile category. Although with or without iteration the first-order asymptotic properties of $\widehat{\boldsymbol{\theta}}$ are identical, their finite sample performance is often slightly different. As is often observed in semiparametric problems, the estimation and inference of $\boldsymbol{\theta}$ is very insensitive to the bandwidth h. A large range of h can be applied including the classical nonparametric optimal bandwidth. Thus, in practice, one can often use a default bandwidth h calculated under the normal density or perform an initial cross-validation to obtain h.

As far as our original goal of estimating the population center μ is concerned, we have obtained the most efficient estimator. Our final remark is about the nonparametric estimation of f_0. Obviously, once we have the efficient estimator $\widehat{\boldsymbol{\theta}}$, plugging it into Equation (5) with a cross-validation selected bandwidth h will in turn provide a valid nonparametric estimation of f_0, up to a normalizing constant. In fact, for the purpose of the nonparametric estimation of f_0, merely using a consistent estimator $\widetilde{\boldsymbol{\theta}}$ in Equation (5) works equally well. This is because as a nonparametric estimator, $\widetilde{f}$ has slower rate than root-n; hence, as long as root-n consistency is retained, the variance involved in estimating $\boldsymbol{\theta}$ has no first-order effect. In other words, plugging $\widetilde{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\theta}}$, or even $\boldsymbol{\theta}_0$ all yield the same nonparametric estimator $\widetilde{f}$ to its first asymptotic order. Finally, to correct for the normalizing constant, we can simply perform a numerical integration procedure to obtain $\widehat{c}^{-1} = \int \widetilde{f}(t;\widehat{\boldsymbol{\theta}})dt$, and form $\widehat{f}(t) = \widehat{c}\widetilde{f}(t;\widehat{\boldsymbol{\theta}})$.

4. ASYMPTOTIC PROPERTIES

We have proposed a class of estimators that are consistent under misspecification of f. To improve the estimation efficiency, we have allowed for an additional parameter $\boldsymbol{\gamma}$ in the specified model, as well as nonparametric estimation. We also provided a refined nonparametric kernel estimator of f. We now summarize the asymptotic properties of these various estimators in several theorems. The proofs are relegated to the Appendix.

Theorem 1. Assume $f^*(t)$ is a symmetric density function and $E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}, f^*)\} = \mathbf{0}$ has a unique root. Let

$$\mathbf{A} = E\left\{\frac{\partial \mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}_0, f^*)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}, \quad \mathbf{B} = E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}_0, f^*)^{\otimes 2}\}$$

be bounded nonsingular matrices. Then the estimator $\widetilde{\boldsymbol{\theta}}$, obtained by solving Equation (4) satisfies

$$n^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

in distribution when $n \to \infty$.

Theorem 1 is readily seen via a simple Taylor expansion, hence we omit its proof. A more interesting result concerns the additional parameter $\boldsymbol{\gamma}$ in f^* and its effect on $\boldsymbol{\theta}$, stated in Theorem 2.

Theorem 2. Assume $f^*(t;\boldsymbol{\gamma})$ is a family of symmetric density functions and

$$E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}, f^*(X-\mu;\boldsymbol{\gamma})\} = \mathbf{0}, \quad E\{\mathbf{S}_{\boldsymbol{\gamma}}(X;\boldsymbol{\theta},\boldsymbol{\gamma}, f^*)\} = \mathbf{0}$$

has a unique root. Denote by $\boldsymbol{\gamma}^*$ the $\boldsymbol{\gamma}$ component of the unique root. Let

$$\mathbf{A} = E\left[\frac{\partial \mathbf{S}_{\text{eff}}\{X;\boldsymbol{\theta}_0, f^*(X-\mu_0;\boldsymbol{\gamma}^*)\}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right],$$
$$\mathbf{B} = E[\mathbf{S}_{\text{eff}}\{X;\boldsymbol{\theta}_0, f^*(X-\mu_0;\boldsymbol{\gamma}^*)\}^{\otimes 2}]$$

be bounded nonsingular matrices. Then the estimator $\widetilde{\boldsymbol{\theta}}$, obtained through solving

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(X_i;\boldsymbol{\theta}, f^*(X-\mu;\boldsymbol{\gamma}) = \mathbf{0}, \quad \sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\gamma}}(X_i;\boldsymbol{\theta},\boldsymbol{\gamma}, f^*) = \mathbf{0}$$

satisfies

$$n^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

in distribution when $n \to \infty$.

Comparing the results in Theorem 1 and in Theorem 2, we can see that the two estimators have essentially identical properties. More specifically, postulating a family of models $f^*(t;\boldsymbol{\gamma})$ with an unknown parameter $\boldsymbol{\gamma}$ yields an estimator that is asymptotically equal to the estimator if we had postulated the fixed model $f^*(t;\boldsymbol{\gamma}^*)$. In other words, the variability associated with the estimation of $\boldsymbol{\gamma}$ does not have any impact on the variability in estimating the parameter of interest $\boldsymbol{\theta}$.

Instead of postulating a parametric model family for f and estimating $\boldsymbol{\gamma}$, the nonparametric alternative aims to estimate f in a model-free fashion. This is the philosophy behind the refined nonparametric estimator proposed in Section 3.2. We summarize the asymptotic properties of the estimator in Theorem 3 and provide the necessary conditions and proof in the Appendix. For notational brevity, we write $w_1(t;\boldsymbol{\beta}) = w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})$, $w_1'(t;\boldsymbol{\beta}) = \partial w_1(t;\boldsymbol{\beta})/\partial t$, $w_1''(t;\boldsymbol{\beta}) = \partial^2 w_1(t;\boldsymbol{\beta})/\partial t^2$.

Theorem 3. Let $c_2 = \int_{-1}^{1} s^2 K(s)ds$, $v_2 = \int_{-1}^{1} K^2(s)ds$, and $\widetilde{\boldsymbol{\theta}}$ be obtained from solving Equation (4). Under the regularity conditions C1–C4 given in the Appendix, the nonparametric estimator $\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})$ given in Equation (5) satisfies

$$\begin{aligned}
&\text{bias}\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\} \\
&\quad \equiv E\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\} - c(\boldsymbol{\beta}_0)f_0(t) = \frac{h^2 c(\boldsymbol{\beta}_0)c_2}{2} \\
&\qquad \times \left\{f_0''(t) + \frac{2f_0'(t)w_1'(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)} + \frac{f_0(t)w_1''(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)}\right\} + o(h^2) \\
&\text{var}\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\} \\
&\quad = \frac{c(\boldsymbol{\beta}_0)}{nhw_1(t;\boldsymbol{\beta}_0)}\left\{v_2 f_0(t) + \frac{2I(|t|<h)}{w_1(t;\boldsymbol{\beta}_0)}\right. \\
&\qquad \left. \times \int_0^{1-\frac{|t|}{h}} K(s-t/h)K(s+t/h)f_0(hs)w_1(hs;\boldsymbol{\beta}_0)ds\right\} \\
&\qquad + o\{(nh)^{-1}\} \\
&\quad \le \frac{2c(\boldsymbol{\beta}_0)v_2 f_0(t)}{nhw_1(t;\boldsymbol{\beta}_0)} + o\{(nh)^{-1}\}.
\end{aligned}$$

The estimator $\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})$ is intended to be an estimator for $f_0(t)$ without adjusting the normalizing constant, hence our quantification of bias takes this into account. The integration in the variance expression in Theorem 3 is a bounded quantity under the regularity conditions, hence the nonparametric estimator $\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})$ has the classical bias and variance properties. Because the only a priori information we have about f is its symmetry, this does not come as a surprise. The bias and variance properties subsequently guarantee that the mean squared error (MSE) and mean integrated squared error (MISE) also have the classical nonparametric rates. Similarly, one can easily take derivative of the estimator $\widetilde{f}$ to obtain a nonparametric estimator $\widetilde{f}'$. It is easy to see that the derivative estimator will also have the classical bias and variance rates. Theorem 3 prepares the results in Theorem 4.

Theorem 4. Let $X_1, \ldots, X_n$ be iid with density (1) and let $\widetilde{\boldsymbol{\theta}}$ be an initial estimator obtained from solving Equation (4). Let $\widetilde{f}(t;\boldsymbol{\theta})$ be given by Equation (5) for any t and any $\boldsymbol{\theta}$. Assume $E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}, f_0)\} = \mathbf{0}$ has a unique root and $\widehat{\boldsymbol{\theta}}$ satisfies

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}\{X_i;\widehat{\boldsymbol{\theta}}, \widetilde{f}(X_i - \widehat{\mu};\widetilde{\boldsymbol{\beta}})\} = \mathbf{0}.$$

It then follows that when $n \to \infty$, under the regularity conditions C1–C4 listed in the Appendix, $\widehat{\boldsymbol{\theta}}$ is the semiparametric efficient estimator and it satisfies

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(\mathbf{0}, [E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}_0, f_0)^{\otimes 2}\}]^{-1})$$

in distribution when $n \to \infty$.

In terms of estimating $\boldsymbol{\theta}$, Theorem 4 contains the strongest result regarding estimation efficiency. It clearly states that as long as we incorporate a suitable nonparametric estimation of f, even if this nonparametric estimation is conducted using an initial root-n consistent estimator of $\boldsymbol{\theta}$, the efficient estimator will still be achieved in model (1).

Table 1. Results of simulation studies for five semiparametric estimators of μ when the selection function w is known

	Antisymmetric w				General w			
	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg
$f_0 = \phi$								
est1	4.0061	0.1541	0.1549	95.4	3.9889	0.1784	0.1809	96.1
est2	4.0024	0.1539	0.1587	95.3	3.9878	0.1792	0.1872	96.4
est3	4.0058	0.1539	0.1550	95.5	3.9890	0.1786	0.1809	96.1
est4	4.0015	0.1537	0.1597	95.6	3.9758	0.1849	0.1826	95.5
est5	4.0050	0.1557	0.1547	94.7	3.9873	0.1787	0.1816	95.4
$f_0 = t_5$								
est1	3.9965	0.1437	0.1409	95.7	3.9893	0.1803	0.1776	94.5
est2	4.0014	0.1306	0.1313	95.5	3.9904	0.1559	0.1605	94.6
est3	3.9981	0.1429	0.1404	95.9	3.9893	0.1809	0.1776	94.4
est4	4.0016	0.1300	0.1316	95.5	3.9966	0.1733	0.1810	94.6
est5	3.9980	0.1300	0.1319	95.0	3.9952	0.1644	0.1614	94.1

NOTE: Mean, sample standard deviation (sd), average of the estimated standard deviation ($\widehat{\text{sd}}$), and the 95% coverage probabilities of $\mu = 4$ are reported, when the selection function is antisymmetric (left) and is general (right), and the symmetric density f_0 is normal (ϕ) and is Student's t with 5 degrees of freedom (t_5). Results are obtained with sample size $n = 500$ and 1000 simulations.

5. NUMERICAL EXAMPLES

5.1 Simulations With Known Selection Function

We illustrate the finite sample performance of the estimators proposed in Sections 2 and 3 through a series of simulation studies. In each of them, we generated 1000 datasets, each with sample size $n = 500$, from model (1) with different choices of the symmetric pdf and the selection function.

The first set of simulations contain two separate studies. In the first study, the true $f_0(x)$ pdf is normal with mean $\mu = 4$ and standard deviation $\gamma = 3$. The selection function is of logit-type and is either antisymmetric around 1/2, where $w(x;\beta) = (0.15 + 0.85e^{\beta x})/(1 + e^{\beta x})$, or is general, where $w(x;\beta) = (0.15 + 0.85e^{-1+\beta x})/(1 + e^{-1+\beta x})$. In both cases, the true β value is 2. We implemented five different semiparametric estimators of μ on each of the simulated datasets.

In the first estimator, we proposed the true normal density f_0 as the posited model for f to form the corresponding estimating equation. This means in the estimating Equation (4), we adopt $f^* = f_0$, and solve it to obtain $\widehat{\boldsymbol{\theta}}$. Note that this is the oracle estimator. Because proposing a true model is not likely achievable in practice, we further implemented a second estimator, where we plug in a wrong form for f. Specifically, we adopted a Student's t with 5 degrees of freedom density function with standard deviation 3 as f^* and plugged it into the estimating Equation (4) to obtain the second estimator. To further increase the flexibility of these two estimators, we also implemented the third and fourth estimators. In these two estimators, the function f^* contains an unknown scale parameter and hence is not fully specified. Specifically, in the third estimator, we used a normal model for f^*, and in the fourth estimator, we used a Student's t with 5 degrees of freedom model for f^*. In both cases, the standard deviation of the model is left unspecified, and is treated as a nuisance parameter estimated using the methods described in Section 3.1. Finally, we also implemented the fully nonparametric estimator described in Section 3.2 as our fifth estimator.

The second study in this set of simulations contains exactly the same five estimators. The difference from the first study is that now the data are generated from a Student's t distribution with 5 degrees of freedom. Thus, the second and fourth estimators now contain the true f function or true f model, while the first and third estimators contain a misspecified f function or model.

The results of this set of simulations are given in Table 1. It is evident that regardless of whether a true or false f function is adopted, regardless of whether f is fully specified or partially specified or even completely unspecified, all five estimators yield estimators with very small biases. It is also clear that when additional nuisance parameters are included in the third and fourth estimators, the resulting estimation variability almost remains unchanged in comparison with their simpler versions, that is, the first and second estimators. Although the estimators engaging a false f model (second and fourth estimators in the first study and first and third estimators in the second study) can lead to efficiency loss, this loss is only noticeable in the second study. The asymptotic optimality of the fifth estimator is also reflected in these studies in that its sample variance is comparable with that of the oracle estimator. Finally, the inference results are reasonably precise, reflected in the closeness between the sample standard deviation and their estimated version, and the closeness of the 95% coverage to the nominal value.

5.2 Simulations With Selection Function Containing an Unknown Parameter

To further study the properties of our proposed estimators, we conducted a second set of simulation studies where the selection functions are the same as in Section 5.1 but the parameter β is now treated as unknown and therefore needs to be estimated along with other parameters. All other designs of the simulations are kept unchanged from the first set in Section 5.1.

The results parallel to that of the first set are presented in the upper half of Table 2, in an identical layout. Similar claims can

Table 2. Results of simulation studies for five semiparametric estimators of μ when the selection function w contains an unknown parameter β

	Antisymmetric w				General w			
	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg
$f_0 = \phi$								
est1	4.0167	0.2057	0.2033	96.0	4.0111	0.1945	0.1900	95.9
est2	4.0221	0.2139	0.2091	95.9	4.0130	0.1992	0.1956	95.7
est3	4.0166	0.2065	0.2033	96.1	4.0111	0.1935	0.1900	95.9
est4	4.0241	0.2127	0.2116	95.5	4.0064	0.1888	0.1929	95.9
est5	4.0139	0.2100	0.2083	95.7	4.0107	0.1995	0.1906	96.0
$f_0 = t_5$								
est1	3.9966	0.2413	0.2237	95.7	3.9993	0.1976	0.2009	96.5
est2	4.0111	0.2095	0.1970	94.7	4.0058	0.1690	0.1812	95.9
est3	3.9968	0.2412	0.2249	95.6	3.9993	0.1975	0.2005	96.4
est4	4.0112	0.2096	0.1989	94.1	4.0000	0.1653	0.1873	96.6
est5	4.0059	0.2135	0.2068	94.4	4.0056	0.1851	0.1842	96.5
	$\widehat{\beta}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\beta}$	sd	$\widehat{\text{sd}}$	95% cvg
$f_0 = \phi$								
est1	1.9793	0.6289	0.5862	93.9	1.9479	0.5313	0.5440	92.6
est2	1.9759	0.6257	0.5954	93.0	1.9607	0.5307	0.5472	92.5
est3	1.9775	0.6289	0.5882	93.8	1.9479	0.5310	0.5440	92.6
est4	1.9654	0.6277	0.6053	92.4	1.9599	0.5262	0.5500	93.2
est5	1.9734	0.6328	0.5949	93.4	1.9553	0.5242	0.5450	92.7
$f_0 = t_5$								
est1	1.9886	0.6999	0.7222	96.4	1.9582	0.5433	0.5655	93.8
est2	1.9594	0.6256	0.6498	94.3	1.9688	0.5194	0.5441	92.9
est3	1.9875	0.7006	0.7256	96.4	1.9576	0.5443	0.5655	93.7
est4	1.9578	0.6222	0.6496	93.6	1.9815	0.5237	0.5537	94.8
est5	1.9564	0.6426	0.6628	94.4	1.9737	0.5246	0.5511	93.0

NOTE: Mean, sample standard deviation (sd), average of the estimated standard deviation ($\widehat{\text{sd}}$), and the 95% coverage probabilities of $\mu = 4$ and of $\beta = 2$ are reported, when the selection function is antisymmetric (left) and is general (right), and the symmetric density f_0 is normal (ϕ) and is Student's t with 5 degrees of freedom (t_5). Results are obtained with sample size $n = 500$ and 1000 simulations.

be made regarding the finite sample bias. However, in this set of simulations, the efficiency loss caused by engaging a false f model is better manifested, where the second and fourth estimators in the first study, and the first and third estimators in the second study resulted in the largest sample standard deviation among all five estimators. The optimality of the nonparametric fifth estimator is still quite clear in the situation when the selection function is antisymmetric around 1/2, while it is less obvious for the general selection function, indicating that when additional parameters in the selection function are involved, the first-order asymptotic results regarding estimation efficiency may require a sample size larger than $n = 500$. In the lower half of Table 2, we provide the estimation and inference results regarding the parameter β in the selection function. Generally speaking, the estimates show very small finite sample bias, while the inference results are somewhat less precise in comparison to the μ estimation. We have investigated this issue further, and found that this is a common observation in the parametric model, where both the f model and the w model are parametrically specified. In this parametric setting, the difficulty of the estimation and inference regarding β is caused by a flat likelihood function as a function of β (see Branco, Genton, and Liseo 2013, and references therein). Thus, when the model is further weakened to semiparametric in our setting, it is not surprising that the inference of β is even more difficult.

To provide a visual inspection of the various simulation settings and the results of the function estimation, we provided the plots of both $\widehat{f}$ (left panels) and $\widehat{g}$ (right panels) in Figure 2 for $f_0 = \phi$ and in Figure 3 for $f_0 = t_5$. Although, for the purpose of estimating the center μ, our theoretical results have shown that the bandwidth selection is a secondary issue and can be handled crudely, it is a rather important aspect for estimating the ultimate f and g function themselves. To this end, we used an indirect cross-validation procedure (Savchuk, Hart, and Sheather 2010) to select the bandwidth. In summary, the indirect cross-validation procedure follows the general idea of the classical cross-validation, except that it uses a special kernel function during the bandwidth selection procedure. This special kernel is a linear combination of two normal densities and can sometimes yield a negative value. Although this type of kernel function is almost never used in nonparametric estimation, it turns out to have superior features for bandwidth selection purpose (see Savchuk, Hart, and Sheather 2010, for further investigation of indirect cross-validation). The selected bandwidth is then transformed back to the proper scale according to the asymptotically optimal bandwidth formula based on the kernel function used in the density estimation procedure. Regarding our simulation results, we would like to point out that in all these cases the f and g curves are visually rather different; hence, this provides intuitive evidence that the selection bias should not be

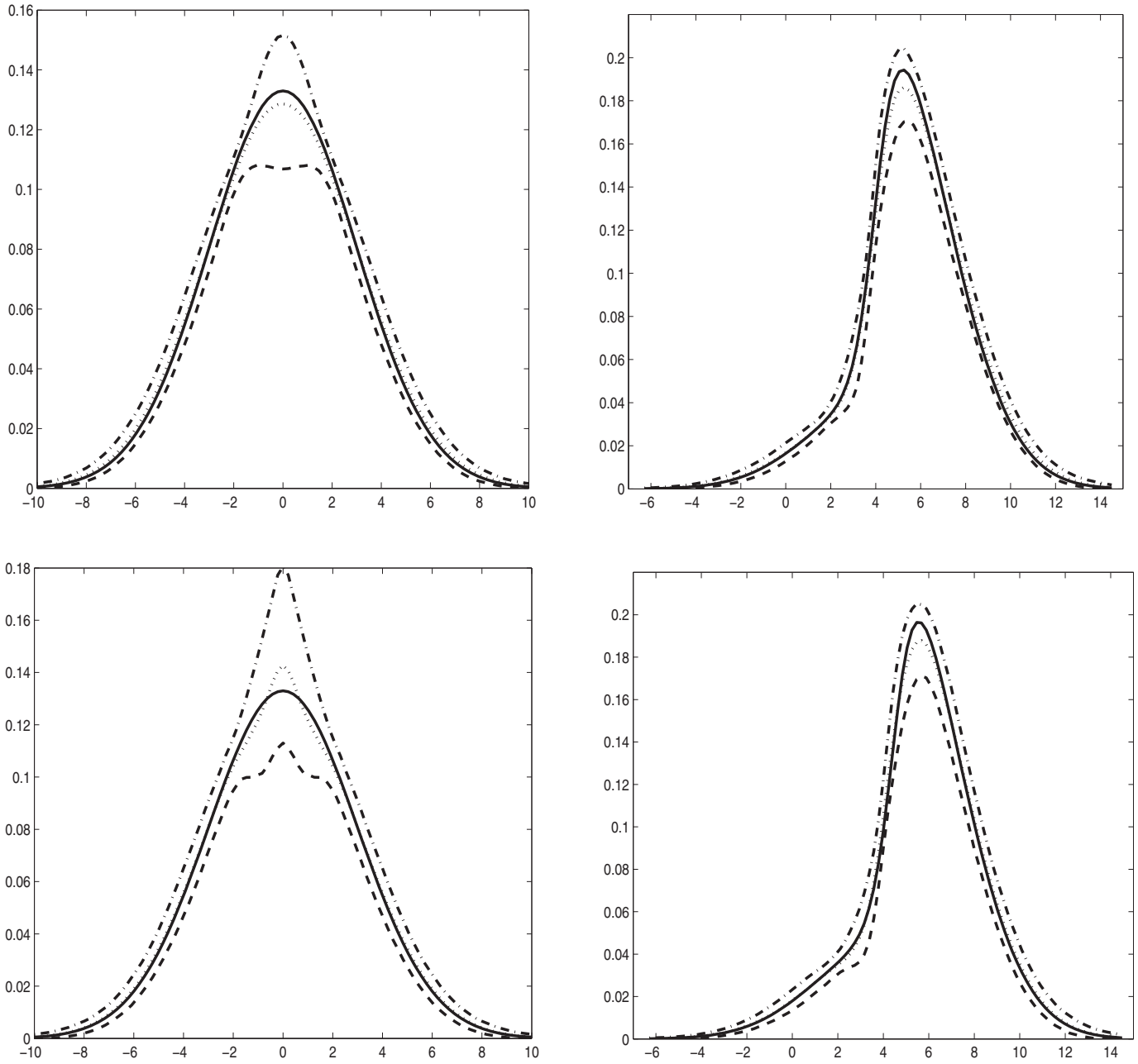

Figure 2. Pointwise quantile curves from Simulation 2 when $f_0 = \phi$. In each plot, the solid line is the true density and the other three curves are the median (dotted), 5% (dashed), and 95% (dot-dashed) quantile curves of all 1000 density estimates derived from Table 2. The left and right panels correspond to the underlying population density, $f(x)$, and the observed selected sample density, $g(x)$, respectively. The top panels are for an antisymmetric w and the bottom panels are for a general w.

ignored. Obviously, when the true f_0 function is normal, the nonparametric estimation performs much better than when f_0 is Student's t with 5 degrees of freedom, where the tails are much heavier.

5.3 Ambulatory Expenditures Data

The ambulatory expenditures data mentioned in the Introduction consists of $n = 2802$ observations. To take into account the possible selection bias, we fit model (1) with the selection function being a general probit model $w(x; \beta_1, \beta_2) = \Phi(\beta_1 + \beta_2 x)$, and its corresponding antisymmetric version by setting $\beta_1 = 0$. The selection functions here are chosen with the intention of capturing the possible behavior patterns when the decision of using the ambulatory service is made. In particular, the antisymmetric probit model is motivated by the assumption of normality used by Cameron and Trivedi (2010), and the general probit model is its natural generalization.

We performed the analysis on the logarithm of the data and implemented five semiparametric estimators for the center μ, respectively, with a posited normal model for f with a fixed standard deviation 1.4107 (this is the sample standard deviation), a posited Student's t with 5 degrees of freedom model for f with standard deviation 1.4107, a posited normal model for f with an unspecified standard deviation, a posited Student's t with 5 degrees of freedom model for f with an unspecified standard deviation, and a nonparametrically estimated f.

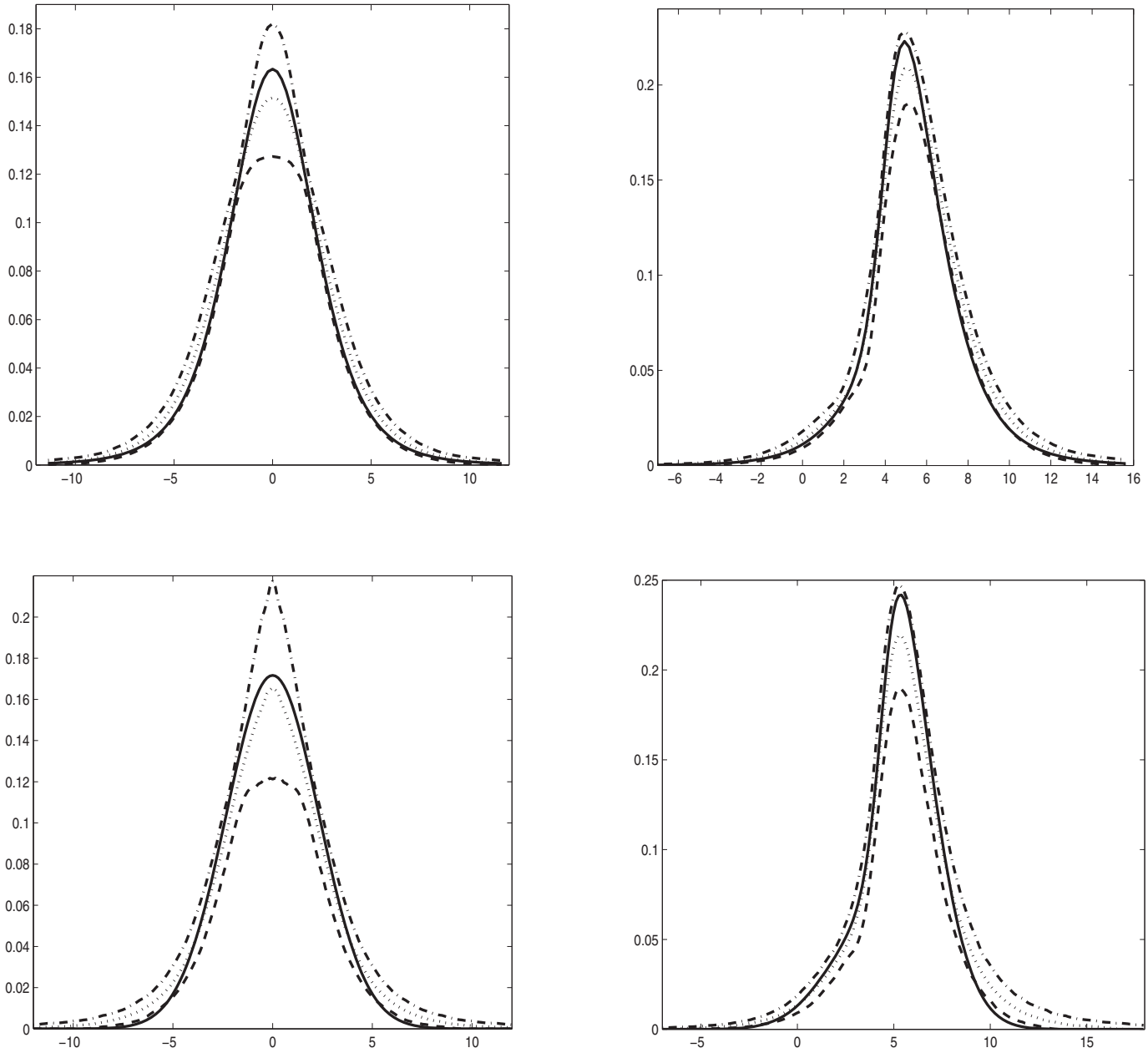

Figure 3. Pointwise quantile curves from Simulation 2 when $f_0 = t_5$. In each plot, the solid line is the true density and the other three curves are the median (dotted), 5% (dashed), and 95% (dot-dashed) quantile curves of all 1000 density estimates derived from Table 2. The left and right panels correspond to the underlying population density, $f(x)$, and the observed selected sample density, $g(x)$, respectively. The top panels are for an antisymmetric w and the bottom panels are for a general w.

The results for the five estimators as well as the estimated standard deviations, in conjunction with the two different selection functions, are listed in Table 3. All estimates resulted in the population center to be significantly larger than the sample average of 6.56, an estimate that does not correct for sample selection bias. All estimates of β_1 and β_2 suggested skewness to the left. This yields a monotonically decreasing weighting function as the expenditure increases, indicating the increasing unwillingness of using the ambulatory service with the increase of the associated cost. This is an indication that patients or their family are less likely to opt for the ambulatory service when the situation tends to incur very large costs. The closeness of the center estimation under the two selection functions implies that an antisymmetric probit model is probably adequate to capture the selection pattern in this example. In addition, when a more general probit model is assumed, the variability of the center estimation increases due to the increased model complexity, especially for the four parametric-model-based estimators.

The estimated densities of the population distribution $\widehat{f}$ and the selected sample distribution $\widehat{g}$ are plotted in Figure 4 when the selection process is modeled through the general probit model. The corresponding plots under an antisymmetric probit model yield very similar curves hence are omitted. These nonparametric estimates used a bandwidth selected through indirect cross-validation procedure (Savchuck, Hart, and Sheather 2010), which is more reliable than the classical cross-validation

Table 3. Five semiparametric estimates of μ and β and their estimated standard deviation for the ambulatory expenditures data, under two selection function models w

	Antisymmetric w		General w	
	$\widehat{\mu}$	$\widehat{\mathrm{sd}}(\widehat{\mu})$	$\widehat{\mu}$	$\widehat{\mathrm{sd}}(\widehat{\mu})$
est1	8.0806	0.1862	7.7101	0.8818
est2	8.0805	0.1862	7.7101	0.8904
est3	8.0806	0.1861	7.7103	0.8264
est4	8.0804	0.1862	7.7101	0.9508
est5	7.9467	0.1456	7.9649	0.2098
	–	–	$\widehat{\beta}_1$	$\widehat{\mathrm{sd}}(\widehat{\beta}_1)$
est1	–	–	0.5913	1.4726
est2	–	–	0.5913	1.4827
est3	–	–	0.5910	1.3832
est4	–	–	0.5913	1.5796
est5	–	–	0.1821	0.4119
	$\widehat{\beta}_2$	$\widehat{\mathrm{sd}}(\widehat{\beta}_2)$	$\widehat{\beta}_2$	$\widehat{\mathrm{sd}}(\widehat{\beta}_2)$
est1	−1.0190	0.1523	−1.0542	0.1369
est2	−1.0190	0.1523	−1.0542	0.1361
est3	−1.0190	0.1523	−1.0542	0.1348
est4	−1.0189	0.1524	−1.0542	0.1383
est5	−0.9150	0.1133	−1.0312	0.1483

procedure. The estimated sample density curve is overlayed on the histogram of the observations and shows a good fit. The estimated density $\widehat{f}$ has a nonnormal and non-Student's t shape, hence confirming that it is wise to leave f completely unspecified.

6. DISCUSSION

We have proposed methods of estimation for the center of a symmetric population when a representative sample of the population is unavailable due to selection bias. Unlike previous studies, we have allowed an arbitrary sample selection mechanism determined by the data collection procedure, and we have not imposed any parametric form on the population distribution. Under this general framework, we have constructed a family of consistent estimators that is robust to population model misspecification, and identified the efficient member that reaches the minimum possible estimation variance. The asymptotic properties and finite sample performance of the estimation and inference procedures were illustrated through theoretical analysis and simulations. A data example about ambulatory expenditures was also provided to illustrate the usefulness of the methods in practice.

We have treated the case of model (1) where the pdf f is completely unspecified and the selection function w is assumed to have a known parametric form. An alternative setting is when f has a known parametric form, whereas the selection mechanism is somewhat hidden; hence, the selection function w is unknown. Such models, with the additional antisymmetric assumption on w, have been investigated by Ma, Genton, and Tsiatis (2005), Ma and Hart (2007), and Azzalini, Genton, and Scarpa (2010).

One may also consider some further generalization of our current model. For example, one possibility is to relax the symmetry assumption on f, and hence leave f completely arbitrary. In this context, the notion of "center" is of course not well defined. Depending on the circumstance, one might be interested in estimating the mean, or the median, or the mode. Our preliminary investigation indicates that interestingly, these different "center-like" statistics require different estimation procedures, and these procedures are not a generalization of the methodology developed here. Research along this line can be quite interesting and promising.

It is worth pointing out that because a parametric form is not assumed on f, the structure of the data relies largely on the selection model w. In fact, the selection procedure is entirely

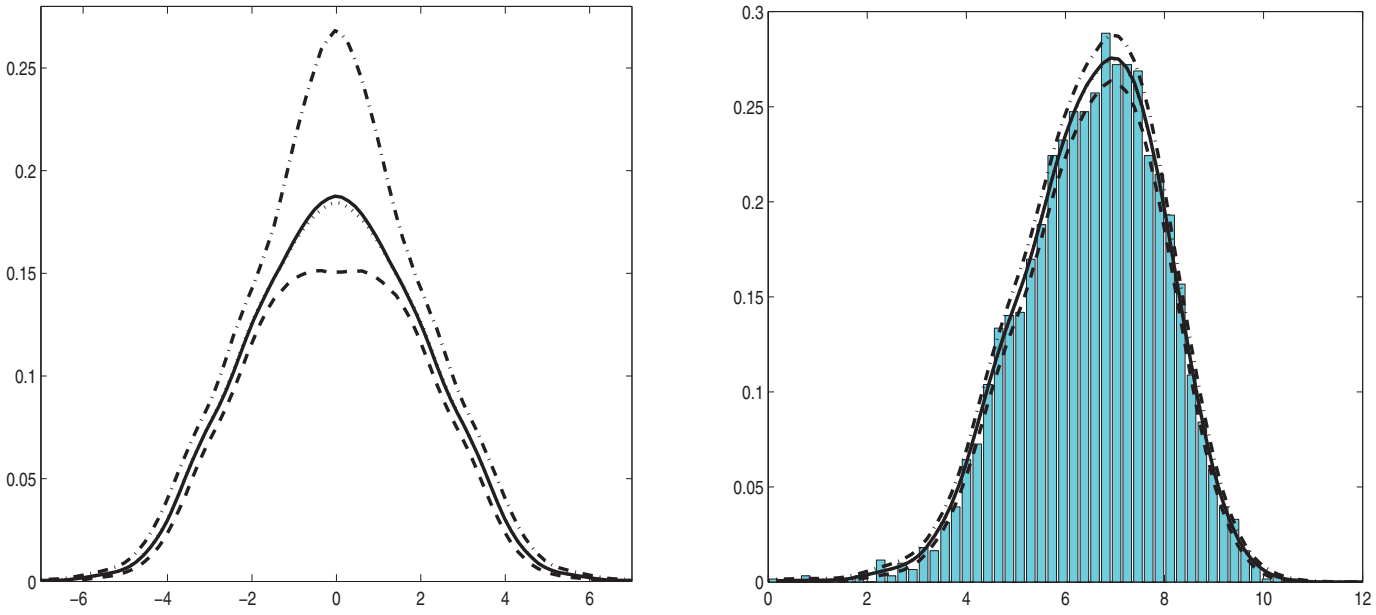

Figure 4. The estimated densities of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ for the ambulatory expenditures data (right), under a general selection function. The estimated sample density curve is overlayed on the histogram of the observations, along with the median (dotted), 5% (dashed), and 95% (dot-dashed) quantile curves of the pointwise confidence bands. The online version of this figure is in color.

Table 4. Results of simulation studies for five semiparametric estimators of μ when the selection function w is misspecified (true selection logit, misspeficied to probit)

	Antisymmetric w				General w			
	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg
$f_0 = \phi$								
est1	4.0090	0.1990	0.1992	95.4	4.0057	0.1958	0.1896	96.3
est2	4.0029	0.2097	0.2051	95.6	4.0051	0.1959	0.1946	95.4
est3	4.0086	0.1988	0.1996	95.5	4.0060	0.1964	0.1896	96.2
est4	4.0051	0.2096	0.2075	95.5	4.0042	0.1901	0.1943	95.6
est5	4.0068	0.2047	0.2011	95.1	4.0088	0.1996	0.1898	95.8
$f_0 = \phi$								
est1	3.9815	0.2198	0.2305	95.6	3.9474	0.1654	0.1687	92.3
est2	3.9844	0.2353	0.2324	94.0	3.9268	0.1782	0.1776	92.5
est3	3.9812	0.2195	0.2305	95.5	3.9475	0.1654	0.1702	92.3
est4	3.9860	0.2368	0.2327	94.3	3.9248	0.1751	0.1775	92.1
est5	3.9832	0.2320	0.2298	93.8	3.9520	0.1763	0.1668	90.5

NOTE: Mean, sample standard deviation (sd), average of the estimated standard deviation ($\widehat{\text{sd}}$), and the 95% coverage probabilities of $\mu = 4$ are reported, when the selection function is antisymmetric (left) and is general (right). The symmetric density f_0 is normal. Results are obtained with sample size $n = 500$ and 1000 simulations.

captured in this function, hence it is intuitive to expect that a misspecified selection model would lead to biased estimation. This is indeed what we observed in our numerical experiment. We performed a simple example, where even though the true selection process follows a logistic pattern, we assumed a probit model. Depending on the true selection parameter value β, we see different levels of estimation bias. We listed two sets of such results in Table 4 as an illustration. Although under the first set of selection parameters (upper half of Table 4), the estimators exhibit some robustness to the misspecification of w; this property quickly vanishes when we change the selection parameter values (lower half of Table 4).

Finally, in practice, further data complications can occur, in that besides selection bias, the observations can be further censored or missing. With the help of the methods developed here, extensions to handle such additional data features become feasible. For example, one can incorporate an inverse probability weighting or augmented inverse probability weighting technique on the estimating Equation (4) to handle censored observations or more generally missing at random issues.

APPENDIX

A.1 Derivation of $\Lambda^{\perp}$

To prepare for the derivation of $\Lambda^{\perp}$, we first show that the nuisance tangent space of Equation (1) is

$$\Lambda = \left\{ \mathbf{u}(X-\mu) : \mathbf{u}(z) = \mathbf{u}(-z), \int_{-\infty}^{\infty} \mathbf{u}(t) f_0(t) w(t;\boldsymbol{\beta}) dt = \mathbf{0} \; a.s., \; \mathbf{u} \in \mathbb{R}^p \right\}.$$

To show the above result, we first write the right-hand side of the above expression as A, and then show $A \subset \Lambda$ and $\Lambda \subset A$.

To show $A \subset \Lambda$, assume that we have an arbitrary $\mathbf{u}(X-\mu) \in A$. Therefore, $\int_{-\infty}^{\infty} \mathbf{u}(t) f_0(t) w(t;\boldsymbol{\beta}) dt = \mathbf{0}$ and $\mathbf{u}$ is an even function. This obviously yields $E\{\mathbf{u}(X-\mu)\} = \mathbf{0}$. Consider a parametric submodel

$$g(X;\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{f(X-\mu;\boldsymbol{\gamma}) w(X-\mu;\boldsymbol{\beta})}{\int f(t;\boldsymbol{\gamma}) w(t;\boldsymbol{\beta}) dt},$$

where $f(z;\boldsymbol{\gamma}) = f_0(z)\{1 + e^{-2\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{u}(z)}\}^{-1} / \int f_0(t)\{1 + e^{-2\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{u}(t)}\}^{-1} dt$, $\boldsymbol{\gamma}$ is a finite dimensional nuisance parameter, and $\boldsymbol{\gamma} = \mathbf{0}$ yields the true model. Some algebra yields that

$$\left.\frac{\partial \log g(x;\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\mathbf{0}} = \left.\frac{\partial \log f(x-\mu_0;\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\mathbf{0}} - \left.\frac{\partial \log \int f(t;\boldsymbol{\gamma}) w(t;\boldsymbol{\beta}) dt}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\mathbf{0}} = \mathbf{u}(x-\mu).$$

Hence, $\mathbf{u}(X-\mu)$ is a nuisance score vector of a particular submodel, that is, $\mathbf{u}(X-\mu) \in \Lambda$.

We now show $\Lambda \subset A$. Consider an arbitrary element of Λ, which is the nuisance score of a corresponding parametric submodel

$$g(X;\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{f(X-\mu;\boldsymbol{\gamma}) w(X-\mu;\boldsymbol{\beta})}{\int f(t;\boldsymbol{\gamma}) w(t;\boldsymbol{\beta}) dt}.$$

Then, we can write it as

$$\mathbf{u}(x-\mu) = \frac{\partial f(x-\mu;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}}{f_0(x-\mu)} - \frac{\int \partial f(t;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} w(t;\boldsymbol{\beta}) dt}{\int f_0(t) w(t;\boldsymbol{\beta}) dt}.$$

Since $f(z;\boldsymbol{\gamma}) = f(-z;\boldsymbol{\gamma})$, we have $\partial f(z;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} = \partial f(-z;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$ for any $\boldsymbol{\gamma}$. This implies

$$\frac{\partial f(z;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}}{f(z;\boldsymbol{\gamma})} = \frac{\partial f(-z;\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}}{f(-z;\boldsymbol{\gamma})}$$

for any $\boldsymbol{\gamma}$. The second term in the expression of $\mathbf{u}(x-\mu)$ is a constant. Thus, we obtain $\mathbf{u}(z) = \mathbf{u}(-z)$. Simple algebra can verify that

$$\int_{-\infty}^{\infty} \mathbf{u}(t) f_0(t) w(t;\boldsymbol{\beta}) dt = \mathbf{0}.$$

Thus, we have shown $\Lambda \subset A$.

We are now ready to demonstrate the form of $\Lambda^{\perp}$. Again, we prove the form of $\Lambda^{\perp}$ by defining a space $L = \{\mathbf{v}(X-\mu) : \mathbf{v}(z)w(z;\boldsymbol{\beta}) + \mathbf{v}(-z)w(-z;\boldsymbol{\beta}) = \mathbf{0} \;\; \text{a.s.}, \mathbf{v} \in \mathbb{R}^p\}$, and showing that $L \subset \Lambda^{\perp}$ and

$\Lambda^\perp \subset L$. We point out that for any function $\mathbf{u} \in \Lambda$, we have the relation

$$E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} = \int_0^\infty \mathbf{u}(z)\{\mathbf{v}^{\mathrm{T}}(z)w(z;\boldsymbol{\beta}) + \mathbf{v}^{\mathrm{T}}(-z)w(-z;\boldsymbol{\beta})\}c(\boldsymbol{\beta})f_0(z)dz.$$

In addition, the normalizing constant can be expressed as

$$c(\boldsymbol{\beta}) = \left[\int_0^\infty f_0(t)\{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})\}\,dt\right]^{-1}.$$

We first show that $L \subset \Lambda^\perp$. For any function $\mathbf{v}(X-\mu) \in L$ and any function $\mathbf{u} \in \Lambda$, we have

$$E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} = \int_0^\infty \mathbf{u}(t)\{\mathbf{v}^{\mathrm{T}}(t)w(t;\boldsymbol{\beta}) + \mathbf{v}^{\mathrm{T}}(-t)w(-t;\boldsymbol{\beta})\}c(\boldsymbol{\beta})f_0(t)dt = \mathbf{0}$$

by the definition of L. Hence, $\mathbf{v}(X-\mu) \perp \Lambda$. In addition,

$$\begin{aligned} E\{\mathbf{v}(X-\mu)\} &= \int_{-\infty}^\infty \mathbf{v}(t)c(\boldsymbol{\beta})f_0(t)w(t;\boldsymbol{\beta})dt \\ &= \int_0^\infty c(\boldsymbol{\beta})f_0(t)\{\mathbf{v}(t)w(t;\boldsymbol{\beta}) + \mathbf{v}(-t)w(-t;\boldsymbol{\beta})\}dt = \mathbf{0} \end{aligned}$$

due to the definition of L as well. The above two equalities ensure that $\mathbf{v}(X-\mu) \in \Lambda^\perp$, hence $L \subset \Lambda^\perp$.

We now show that $\Lambda^\perp \subset L$. Suppose $\mathbf{v}(X-\mu) \in \Lambda^\perp$, then $E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} = \mathbf{0}$ for any $\mathbf{u}(X-\mu) \in \Lambda$. Let

$$\begin{aligned} \mathbf{u}_1(z) &= \frac{\mathbf{v}(z)w(z;\boldsymbol{\beta}) + \mathbf{v}(-z)w(-z;\boldsymbol{\beta})}{w(z;\boldsymbol{\beta}) + w(-z;\boldsymbol{\beta})}, \\ \mathbf{u}(z) &= \mathbf{u}_1(z) - E\{\mathbf{u}_1(X-\mu)\}, \end{aligned}$$

where

$$\begin{aligned} E\{\mathbf{u}_1(X-\mu)\} &= c(\boldsymbol{\beta})\int_0^\infty \mathbf{u}_1(z)f_0(z)\{w(z;\boldsymbol{\beta}) + w(-z;\boldsymbol{\beta})\}dz \\ &= c(\boldsymbol{\beta})\int_0^\infty \{\mathbf{v}(z)w(z;\boldsymbol{\beta}) + \mathbf{v}(-z)w(-z;\boldsymbol{\beta})\}f_0(z)dz \\ &= E\{\mathbf{v}(X-\mu)\}. \end{aligned}$$

We have $\mathbf{u}(z) = \mathbf{u}(-z)$ and $\int_0^\infty \mathbf{u}(z)f_0(z)\{w(z;\boldsymbol{\beta}) + w(-z;\boldsymbol{\beta})\}dz = \mathbf{0}$, so $\mathbf{u}(z) \in \Lambda$. Some algebra yields

$$\begin{aligned} &E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} \\ &\quad = c(\boldsymbol{\beta})\int_0^\infty \frac{\{\mathbf{v}(t)w(t;\boldsymbol{\beta}) + \mathbf{v}(-t)w(-t;\boldsymbol{\beta})\}^{\otimes 2}f_0(t)}{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})}dt \\ &\qquad - [E\{\mathbf{v}(X-\mu)\}]^{\otimes 2}, \end{aligned}$$

where for any vector $\mathbf{c}$, $\mathbf{c}^{\otimes 2} = \mathbf{c}\mathbf{c}^{\mathrm{T}}$. Since $\mathbf{v}(z) \in \Lambda^\perp$, we have $E\{\mathbf{v}(X-\mu)\} = \mathbf{0}$. Hence, the relation $E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} = \mathbf{0}$ yields

$$\int_0^\infty \frac{\{\mathbf{v}(t)w(t;\boldsymbol{\beta}) + \mathbf{v}(-t)w(-t;\boldsymbol{\beta})\}^{\otimes 2}f_0(t)}{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})}dt = \mathbf{0}.$$

Hence, we have $\mathbf{v}(t)w(t;\boldsymbol{\beta}) + \mathbf{v}(-t)w(-t;\boldsymbol{\beta}) = \mathbf{0}$ *a.s.* This indicates that $\mathbf{v}(X-\mu) \in L$, hence $\Lambda^\perp \subset L$.

A.2 Derivation of the Efficient Score $\mathbf{S}_{\mathrm{eff}}$

Define

$$\begin{aligned} u_1(t) &= -\frac{f_0'(t)\{w(t;\beta) - w(-t;\beta)\}}{f_0(t)\{w(t;\beta) + w(-t;\beta)\}} - \frac{w'(t;\beta) + w'(-t;\beta)}{w(t;\beta) + w(-t;\beta)}, \\ v_1(t) &= \frac{-2f_0'(t)w(-t;\beta)}{f_0(t)\{w(t;\beta) + w(-t;\beta)\}} + \frac{w'(t;\beta) + w'(-t;\beta)}{w(t;\beta) + w(-t;\beta)} - \frac{w'(t;\beta)}{w(t;\beta)}. \end{aligned}$$

Then we have $S_\mu = u_1(x-\mu) + v_1(x-\mu)$. In the following, we show that $u_1(x-\mu) \in \Lambda$ and $v_1(x-\mu) \in \Lambda^\perp$. To show $u_1(x-\mu) \in \Lambda$, we can easily verify that $u_1(t) = u_1(-t)$ and

$$\begin{aligned} &\int_{-\infty}^\infty u_1(t)f_0(t)w(t;\beta)dt \\ &\quad = \int_0^\infty u_1(t)f_0(t)\{w(t;\beta) + w(-t;\beta)\}dt \\ &\quad = -\int_0^\infty f_0'(t)\{w(t;\beta) - w(-t;\beta)\}dt \\ &\qquad - \int_0^\infty \{w'(t;\beta) + w'(-t;\beta)\}f_0(t)dt \\ &\quad = -\int_0^\infty \left[\frac{\partial f_0(t)\{w(t;\beta) - w(-t;\beta)\}}{\partial t}\right]dt = 0. \end{aligned}$$

Hence, $u_1(x-\mu) \in \Lambda$. To show $v_1(x-\mu) \in \Lambda^\perp$, we can easily verify that $v_1(t)w(t;\beta) + v_1(-t)w(-t;\beta) = 0$. Combining the above results, we obtain that $\Pi(S_\mu|\Lambda^\perp) = v_1(x-\mu)$.

Now we decompose $\mathbf{S}_{\boldsymbol{\beta}}$. Define

$$\begin{aligned} \mathbf{u}_2(t) &= \frac{\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta}) + \mathbf{w}_{\boldsymbol{\beta}}(-t;\boldsymbol{\beta})}{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})} - \frac{\int f_0(t)\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta})dt}{\int f_0(t)w(t;\boldsymbol{\beta})dt}, \\ \mathbf{v}_2(t) &= -\frac{\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta}) + \mathbf{w}_{\boldsymbol{\beta}}(-t;\boldsymbol{\beta})}{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})} + \frac{\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta})}{w(t;\boldsymbol{\beta})}. \end{aligned}$$

Then we have $\mathbf{S}_{\boldsymbol{\beta}} = \mathbf{u}_2(x-\mu) + \mathbf{v}_2(x-\mu)$. In the following, we show that $\mathbf{u}_2(x-\mu) \in \Lambda$ and $\mathbf{v}_2(x-\mu) \in \Lambda^\perp$. Obviously, $\mathbf{u}_2(t) = \mathbf{u}_2(-t)$ and

$$\begin{aligned} \int_{-\infty}^\infty \mathbf{u}_2(t)f_0(t)w(t;\boldsymbol{\beta})dt &= \int_0^\infty \mathbf{u}_2(t)f_0(t)\{w(t;\boldsymbol{\beta}) + w(-t;\boldsymbol{\beta})\}dt \\ &= \int_0^\infty f_0(t)\{\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta}) + \mathbf{w}_{\boldsymbol{\beta}}(-t;\boldsymbol{\beta})\}dt \\ &\quad - \int_{-\infty}^\infty f_0(t)\mathbf{w}_{\boldsymbol{\beta}}(t;\boldsymbol{\beta})dt = \mathbf{0}. \end{aligned}$$

Thus, $\mathbf{u}_2(x-\mu) \in \Lambda$. To show $\mathbf{v}_2(x-\mu) \in \Lambda^\perp$, we can easily verify that

$$\mathbf{v}_2(t)w(t;\boldsymbol{\beta}) + \mathbf{v}_2(-t)w(-t;\boldsymbol{\beta}) = \mathbf{0}.$$

Hence, $\mathbf{v}_2(t) \in \Lambda^\perp$. Combining the above results, we obtain that $\Pi(\mathbf{S}_{\boldsymbol{\beta}}|\Lambda^\perp) = \mathbf{v}_2(x-\mu)$.

Combining $\Pi(S_\mu|\Lambda^\perp)$ and $\Pi(\mathbf{S}_{\boldsymbol{\beta}}|\Lambda^\perp)$, we obtain the desired form of the efficient score.

A.3 Proof of Theorem 2

Obviously at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we have $E\{\mathbf{S}_{\mathrm{eff}}(X;\boldsymbol{\theta}_0, f^*(X-\mu_0;\boldsymbol{\gamma})\} = \mathbf{0}$ for any $\boldsymbol{\gamma}$. Hence, the unique solution is $(\boldsymbol{\theta}_0^{\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$. For simplicity, we denote $\boldsymbol{\alpha} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, the roots of the estimating equation as $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, the unique root $\boldsymbol{\alpha}_0 = (\boldsymbol{\theta}_0^{\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{S}(X;\boldsymbol{\alpha}, f^*) = \left[\mathbf{S}_{\mathrm{eff}}\{X;\boldsymbol{\theta}, f^*(X-\mu;\boldsymbol{\gamma})\}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}(X;\boldsymbol{\theta}, \boldsymbol{\gamma}, f^*)^{\mathrm{T}}\right]^{\mathrm{T}}$. Then a standard Taylor expansion yields

$$\begin{aligned} \mathbf{0} &= n^{-1/2}\sum_{i=1}^n \mathbf{S}(X_i;\widetilde{\boldsymbol{\alpha}}, f^*) = n^{-1/2}\sum_{i=1}^n \mathbf{S}(X_i;\boldsymbol{\alpha}_0, f^*) \\ &\quad + n^{-1}\sum_{i=1}^n \frac{\partial \mathbf{S}(X_i;\boldsymbol{\alpha}^*, f^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} n^{1/2}(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0), \end{aligned}$$

where $\boldsymbol{\alpha}^*$ is on the interval connecting $\boldsymbol{\alpha}_0$ and $\widetilde{\boldsymbol{\alpha}}$. This yields

$$\begin{aligned} n^{1/2}(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) &= -n^{-1/2}\left[E\left\{\frac{\partial \mathbf{S}(X_i;\boldsymbol{\alpha}_0, f^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right\}\right]^{-1} \\ &\quad \times \sum_{i=1}^n \mathbf{S}(X_i;\boldsymbol{\alpha}_0, f^*) + o_p(1). \end{aligned} \tag{A.1}$$

Note that the upper-left $p \times p$ block of $E\left\{\partial \mathbf{S}(X_i;\boldsymbol{\alpha}_0, f^*)/\partial \boldsymbol{\alpha}^{\mathrm{T}}\right\}$ is the $\mathbf{A}$ matrix defined in Theorem 2. The remaining upper-right block

satisfies

$$E\left\{\frac{\partial \mathbf{S}_{\text{eff}}(X;\boldsymbol{\alpha}_0,\, f^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\right\} = -E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\alpha}_0,\, f^*)\mathbf{S}_{\boldsymbol{\gamma}}(X;\boldsymbol{\alpha}_0,\, f^*)^{\mathrm{T}}\} = \mathbf{0},$$

where the last equality is because $\mathbf{S}_{\boldsymbol{\gamma}}$ is an element of the nuisance tangent space while $\mathbf{S}_{\text{eff}}$ is orthogonal to this space. Thus, extracting the first p components from Equation (A.1), we have

$$n^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -n^{-1/2}\mathbf{A}^{-1}\sum_{i=1}^{n}\mathbf{S}_{\text{eff}}\{X_i;\boldsymbol{\theta}_0,\, f^*(X_i-\mu_0;\boldsymbol{\gamma}^*)\} + o_p(1),$$

which subsequently proves Theorem 2. □

A.4 List of Regularity Conditions

C1: The symmetric function f_0 is twice differentiable with a compact support, f_0 and f_0' are bounded away from zero and ∞, and $\int f_0^2(t)dt$, $\int (f_0')^2(t)dt$, $\int (f_0'')^2(t)dt$ are bounded.

C2: The selection function w satisfies $0 < w(t;\boldsymbol{\beta}_0) \le 1$ and is twice differentiable with respect to t on the support of f_0 and its first and second derivatives $w'(t;\boldsymbol{\beta}_0)$, $w''(t;\boldsymbol{\beta}_0)$ are bounded. Note that as long as w is bounded, we can always rescale it to achieve $w(t;\boldsymbol{\beta}_0) \le 1$.

C3: The kernel function K integrates to 1, is symmetric about 0, has support $(-1, 1)$, and is twice differentiable on $[-1, 1]$.

C4: The bandwidth satisfies $h = O(n^{-1/5})$. In fact, a bandwidth h satisfying $nh^2 \to \infty$, $h \to 0$ when $n \to \infty$ is already sufficient. This is a very large range and it certainly includes the optimal bandwidth of order $n^{-1/5}$.

Although these conditions are often satisfied in applications, they can be made weaker at the cost of further technicalities. For instance, C1 can be replaced by a weaker condition that requires the tails of the distribution of X_i's to be sufficiently thin at the cost of a much more tedious proof (see Ma and Hart 2007).

A.5 Proof of Theorem 3

To simplify the proof, we split the n observations into two groups, with sample sizes $n_1 = n - n^{1-\epsilon}$, $n_2 = n^{1-\epsilon}$, respectively, where ϵ is a sufficiently small positive number. Suppose that $\widetilde{\boldsymbol{\theta}}$ is obtained using the observations $X_{n_1+1}, \ldots, X_n$, and $\widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})$ is obtained using the observations $X_1, \ldots, X_{n_1}$ and $\widetilde{\boldsymbol{\theta}}$. From Theorem 1, $\widetilde{\boldsymbol{\theta}}$ satisfies $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n_2^{-1/2})$. To calculate the bias, we have

$$\begin{aligned}
&E\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\}\\
&= E\{\widetilde{f}(t;\boldsymbol{\theta}_0)\} + O\left(n_2^{-1/2}\right)\\
&= \frac{1}{hw_1(t;\boldsymbol{\beta}_0)}E\left\{K\left(\frac{X-\mu_0-t}{h}\right) + K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&\quad + O\left(n_2^{-1/2}\right) = \frac{c(\boldsymbol{\beta}_0)}{hw_1(t;\boldsymbol{\beta}_0)}\int_{t+\mu_0-h}^{t+\mu_0+h} K\left(\frac{x-\mu_0-t}{h}\right)\\
&\quad \times f_0(x-\mu_0)w(x-\mu_0;\boldsymbol{\beta}_0)dx + \frac{c(\boldsymbol{\beta}_0)}{hw_1(t;\boldsymbol{\beta}_0)}\\
&\quad \times \int_{\mu_0-t-h}^{\mu_0-t+h} K\left(\frac{x-\mu_0+t}{h}\right) f_0(x-\mu_0)w(x-\mu_0;\boldsymbol{\beta}_0)dx\\
&\quad + O\left(n_2^{-1/2}\right) = \frac{c(\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)}\int_{-1}^{1} K(s)\{f_0(t+hs)w(t+hs;\boldsymbol{\beta}_0)\\
&\quad + f_0(hs-t)w(hs-t;\boldsymbol{\beta}_0)\}ds + O\left(n_2^{-1/2}\right) = c(\boldsymbol{\beta}_0)f_0(t)\\
&\quad + \frac{h^2c(\boldsymbol{\beta}_0)c_2}{2}\left\{f_0''(t) + \frac{2f_0'(t)w_1'(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)} + \frac{f_0(t)w_1''(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)}\right\}\\
&\quad + o(h^2).
\end{aligned}$$

Thus, the bias is

$$\begin{aligned}
&\text{bias}\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\}\\
&= E\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\} - c(\boldsymbol{\beta}_0)f_0(t)\\
&= \frac{h^2c(\boldsymbol{\beta}_0)c_2}{2}\left\{f_0''(t) + \frac{2f_0'(t)w_1'(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)} + \frac{f_0(t)w_1''(t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)}\right\} + o(h^2).
\end{aligned}$$

To analyze the variance, we have

$$\begin{aligned}
&\text{var}\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\}\\
&= \text{var}\{\widetilde{f}(t;\boldsymbol{\theta}_0)\} + O\left(n_2^{-1}\right)\\
&= \text{var}\left[\frac{1}{w_1(t;\boldsymbol{\beta}_0)}\sum_{i=1}^{n_1}\frac{1}{n_1h}\left\{K\left(\frac{X_i-\mu_0-t}{h}\right)\right.\right.\\
&\quad \left.\left. + K\left(\frac{X_i-\mu_0+t}{h}\right)\right\}\right] + O\left(n_2^{-1}\right)\\
&= \frac{1}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right) + K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&\quad + O\left(n_2^{-1}\right)\\
&= \frac{1}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)\right\}\\
&\quad + \frac{1}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{var}\left\{K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&\quad + \frac{2}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right), K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&\quad + O\left(n_2^{-1}\right).
\end{aligned}$$

We can easily obtain

$$\begin{aligned}
&\frac{1}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)\right\}\\
&= \frac{c(\boldsymbol{\beta}_0)}{n_1hw_1^2(t;\boldsymbol{\beta}_0)}\int K^2(s)f_0(t+hs)w(t+hs;\boldsymbol{\beta}_0)ds + O\left(n_1^{-1}\right)\\
&= \frac{c(\boldsymbol{\beta}_0)v_2}{n_1hw_1^2(t;\boldsymbol{\beta}_0)}f_0(t)w(t;\boldsymbol{\beta}_0) + O\left(n_1^{-1}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{1}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{var}\left\{K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&= \frac{c(\boldsymbol{\beta}_0)v_2}{n_1hw_1^2(t;\boldsymbol{\beta}_0)}f_0(t)w(-t;\boldsymbol{\beta}_0) + O\left(n_1^{-1}\right).
\end{aligned}$$

The covariance term vanishes unless t satisfies $-h + |X-\mu_0| < t < h - |X-\mu_0|$. Thus, for $|t| \ge h$, the covariance term is zero. Otherwise, we have

$$\begin{aligned}
&\frac{2}{n_1h^2w_1^2(t;\boldsymbol{\beta}_0)}\text{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right), K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&= \frac{2c(\boldsymbol{\beta}_0)}{n_1hw_1^2(t;\boldsymbol{\beta}_0)}\int_{\frac{|t|}{h}-1}^{1-\frac{|t|}{h}} K(s-t/h)K(s+t/h)\\
&\quad \times f_0(hs)w(hs;\boldsymbol{\beta}_0)ds + O\left(n_1^{-1}\right)\\
&= \frac{2c(\boldsymbol{\beta}_0)}{n_1hw_1^2(t;\boldsymbol{\beta}_0)}\int_{0}^{1-\frac{|t|}{h}} K(s-t/h)K(s+t/h)\\
&\quad \times f_0(hs)w_1(hs;\boldsymbol{\beta}_0)ds + O\left(n_1^{-1}\right).
\end{aligned}$$

Obviously,

$$\begin{aligned}
&2\text{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right), K\left(\frac{X-\mu_0+t}{h}\right)\right\}\\
&\le \text{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)\right\} + \text{var}\left\{K\left(\frac{X-\mu_0+t}{h}\right)\right\},
\end{aligned}$$

hence the above integral is a bounded quantity. Combining the above results, we have

$$\begin{aligned}&\operatorname{var}\{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})\}\\&= \frac{c(\boldsymbol{\beta}_0)}{n_1 h w_1(t;\boldsymbol{\beta}_0)}\left\{v_2 f_0(t) + \frac{2I(|t|<h)}{w_1(t;\boldsymbol{\beta}_0)}\right.\\&\quad\left.\times\int_0^{1-\frac{|t|}{h}} K(s-t/h)K(s+t/h)f_0(hs)w_1(hs;\boldsymbol{\beta}_0)ds\right\}\\&\quad + o\{(n_1h)^{-1}\} \le \frac{2c(\boldsymbol{\beta}_0)v_2 f_0(t)}{n_1 h w_1(t;\boldsymbol{\beta}_0)} + o\{(n_1h)^{-1}\}.\end{aligned}$$

Note that $n_1/n \to 1$, hence we have proved the results. □

A.6 Proof of Theorem 4

To simplify the proof, we split the n observations into three groups, with sample sizes $n_1 = n - 2n^{1-\epsilon}$, $n_2 = n_3 = n^{1-\epsilon}$, respectively, where ϵ is a sufficiently small positive number. The data splitting technique helps to circumvent the complexity of correlations among different components in the estimation procedure. It is not necessary in practice. Let $\widetilde{\boldsymbol{\theta}}$ be an estimator obtained using the observations $X_{n_1+n_2+1}, \ldots, X_n$; let $\widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})$ be obtained using observations $X_{n_1+1}, \ldots, X_{n_1+n_2}$ and $\widetilde{\boldsymbol{\theta}}$; and let the final estimating equation be based on the observations $X_1, \ldots, X_{n_1}$. From Theorem 1, we have $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n_3^{-1/2})$.

We write the estimating equation as

$$\begin{aligned}0 &= n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{S}_{\text{eff}}\{X_i;\widehat{\boldsymbol{\theta}},\, \widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})\}\\&= n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{S}_{\text{eff}}(X_i;\boldsymbol{\theta}_0,\, f_0) + n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial\mathbf{S}_{\text{eff}}\{X_i;\boldsymbol{\theta}^*,\, \widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})\}}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\\&\quad\times n_1^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + n_1^{-1/2}\sum_{i=1}^{n_1}[\mathbf{S}_{\text{eff}}\{X_i;\boldsymbol{\theta}_0,\, \widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})\} - \mathbf{S}_{\text{eff}}(X_i;\boldsymbol{\theta}_0,\, f_0)],\end{aligned}$$

where $\boldsymbol{\theta}^* = \lambda\widehat{\boldsymbol{\theta}} + (1-\lambda)\boldsymbol{\theta}_0$ for $0 \le \lambda \le 1$. It is easy to see that

$$\begin{aligned}n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial\mathbf{S}_{\text{eff}}\{X_i;\boldsymbol{\theta}^*,\, \widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})\}}{\partial\boldsymbol{\theta}^T} &= E\left\{\frac{\partial\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}_0,\, f_0)}{\partial\boldsymbol{\theta}^T}\right\} + o_p(1)\\&= -E\{\mathbf{S}_{\text{eff}}(X;\boldsymbol{\theta}_0,\, f_0)^{\otimes 2}\} + o_p(1),\end{aligned}$$

where we used the results from Theorems 1 and 3 in the first equality and the last equality is because $\mathbf{S}_{\text{eff}}$ is the orthogonal projection of the score function to $\Lambda^{\perp}$. It remains to demonstrate that

$$n_1^{-1/2}\sum_{i=1}^{n_1}\left[\mathbf{S}_{\text{eff}}\{X_i;\boldsymbol{\theta}_0,\, \widetilde{f}(\cdot;\widetilde{\boldsymbol{\theta}})\} - \mathbf{S}_{\text{eff}}(X_i;\boldsymbol{\theta}_0,\, f_0)\right] = o_p(1),$$

or equivalently, using the explicit form of $\mathbf{S}_{\text{eff}}$, we need to show

$$n_1^{-1/2}\sum_{i=1}^{n_1}\left\{\frac{\widetilde{f}'(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}\times\frac{w(-X_i+\mu_0;\boldsymbol{\beta}_0)}{w_1(X_i-\mu_0;\boldsymbol{\beta}_0)} = o_p(1). \tag{A.2}$$

Consider the first moment of the left side of Equation (A.2). We have

$$\begin{aligned}&n_1^{-1/2}\sum_{i=1}^{n_1}E\left[\left\{\frac{\widetilde{f}'(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}\right.\\&\quad\left.\times\frac{w(-X_i+\mu_0;\boldsymbol{\beta}_0)}{w_1(X_i-\mu_0;\boldsymbol{\beta}_0)}\right] = n_1^{1/2}E\int\left\{\frac{\widetilde{f}'(t;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(t)}{f_0(t)}\right\}\\&\quad\times\frac{w(-t;\boldsymbol{\beta}_0)}{w_1(t;\boldsymbol{\beta}_0)}c(\boldsymbol{\beta}_0)f_0(t)w(t;\boldsymbol{\beta}_0)dt = 0,\end{aligned}$$

because the integrand is an odd function. Consider the second moment of the left side of Equation (A.2). We have

$$\begin{aligned}&E\left(\left[n_1^{-1/2}\sum_{i=1}^{n_1}\left\{\frac{\widetilde{f}'(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}\right.\right.\\&\quad\left.\left.\times\frac{2w(-X_i+\mu_0;\boldsymbol{\beta}_0)}{w_1(X_i-\mu_0;\boldsymbol{\beta}_0)}\right]^2\right)\\&= E\left[\left\{\frac{\widetilde{f}'(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(X_i-\mu_0;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}^2\frac{4w^2(-X_i+\mu_0;\boldsymbol{\beta}_0)}{w_1^2(X_i-\mu_0;\boldsymbol{\beta}_0)}\right]\\&= E\int\left\{\frac{\widetilde{f}'(t;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})} - \frac{f_0'(t)}{f_0(t)}\right\}^2\frac{4w^2(-t;\boldsymbol{\beta}_0)}{w_1^2(t;\boldsymbol{\beta}_0)}c(\boldsymbol{\beta}_0)f_0(t)w(t;\boldsymbol{\beta}_0)dt\\&\le 4c(\boldsymbol{\beta}_0)E\int\left[\left\{\frac{\widetilde{f}'(t;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})} - \frac{\widetilde{f}'(t;\boldsymbol{\theta}_0)}{\widetilde{f}(t;\boldsymbol{\theta}_0)}\right\}^2\right.\\&\quad\left.+\left\{\frac{\widetilde{f}'(t;\boldsymbol{\theta}_0)}{\widetilde{f}(t;\boldsymbol{\theta}_0)} - \frac{f_0'(t)}{f_0(t)}\right\}^2\right]f_0(t)dt,\end{aligned}$$

where we used

$$\frac{w(t;\beta_0)w(-t;\beta_0)}{w_1(t;\beta_0)} \le \frac{1}{2} \quad\text{and}\quad \frac{w(-t;\beta_0)}{w_1(t;\beta_0)} \le 1.$$

Using the delta method, we have

$$\begin{aligned}&E\int\left\{\frac{\widetilde{f}'(t;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})} - \frac{\widetilde{f}'(t;\boldsymbol{\theta}_0)}{\widetilde{f}(t;\boldsymbol{\theta}_0)}\right\}^2 f_0(t)dt\\&= E\int E\left[\left\{\frac{\widetilde{f}'(t;\widetilde{\boldsymbol{\theta}})}{\widetilde{f}(t;\widetilde{\boldsymbol{\theta}})} - \frac{\widetilde{f}'(t;\boldsymbol{\theta}_0)}{\widetilde{f}(t;\boldsymbol{\theta}_0)}\right\}^2\Big|X_{n_1+1},\ldots,X_{n_1+n_2}\right]f_0(t)dt\\&= E\{O_p(n_3^{-1})\} = o(1).\end{aligned}$$

On the other hand,

$$E\int\left\{\frac{\widetilde{f}'(t;\boldsymbol{\theta}_0)}{\widetilde{f}(t;\boldsymbol{\theta}_0)} - \frac{f_0'(t)}{f_0(t)}\right\}^2 f_0(t)dt$$

is the MISE of the nonparametric estimations and has order $O\{h^4 + (n_2h^3)^{-1}\} = o(1)$ for $h = O(n^{-1/5})$ following the results in Theorem 3. Thus, the second moment of the left side of Equation (A.2) converges to zero as $n \to \infty$. From the book by Serfling (2002, sec. 1.2.3), Equation (A.2) is indeed true.

Summarizing the above results, taking into account that $n_1 = n - 2n^{1-\epsilon}$ implies

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - n_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = o_p(1),$$

we have

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N_p(\mathbf{0}, [E\{\mathbf{S}_{\text{eff}}(X, \boldsymbol{\theta}_0, f_0)^{\otimes 2}\}]^{-1}).$$

□

[Received August 2011. Revised May 2013.]

REFERENCES

Arellano-Valle, R. B., Branco, M. D., and Genton, M. G. (2006), "A Unified View on Skewed Distributions Arising From Selections," *Canadian Journal of Statistics*, 34, 581–601. [1090]

Arellano-Valle, R. B., and Genton, M. G. (2007), "On the Exact Distribution of Linear Combinations of Order Statistics From Dependent Random Variables," *Journal of Multivariate Analysis*, 98, 1876–1894. [1090]

——— (2008), "On the Exact Distribution of the Maximum of Absolutely Continuous Dependent Random Variables," *Statistics and Probability Letters*, 78, 27–35. [1090]

——— (2010a), "Multivariate Unified Skew-Elliptical Distributions," *Chilean Journal of Statistics*, 1, 17–33. [1091]

——— (2010b), "Multivariate Extended Skew-*t* Distributions and Related Families," *Metron*, 68, 201–234. [1091]

Arnold, B. C., and Beaver, R. J. (2002), "Skewed Multivariate Models Related to Hidden Truncation and/or Selective Reporting," *Test*, 11, 7–54. [1090]

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178. [1090,1091]

——— (2005), "The Skew-Normal Distribution and Related Multivariate Families" (with discussion), *Scandinavian Journal of Statistics*, 32, 159–200. [1090]

Azzalini, A., and Capitanio, A. (2003), "Distributions Generated by Perturbation of Symmetry With Emphasis on a Multivariate Skew *t* Distribution," *Journal of the Royal Statistical Society*, Series B, 65, 367–389. [1090]

Azzalini, A., Genton, M. G., and Scarpa, B. (2010), "Invariance-Based Estimating Equations for Skew-Symmetric Distributions," *Metron*, 68, 275–298. [1099]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: The Johns Hopkins University Press. [1092]

Branco, M. D., Genton, M. G., and Liseo, B. (2013), "Objective Bayesian Analysis of Skew-*t* Distributions," *Scandinavian Journal of Statistics*, 40, 63–85. [1096]

Cameron, A. C., and Trivedi, P. K. (2010), *Microeconometrics Using Stata* (Rev. ed.), College Station, TX: Stata Press. [1091,1097]

Copas, J. B., and Li, H. G. (1997), "Inference From Non-Random Samples" (with discussion), *Journal of the Royal Statistical Society,* Series B, 59, 55–95. [1090]

Genton, M. G. (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality* (Ed. Vol.), Boca Raton, FL: Chapman & Hall/CRC. [1090]

Genton, M. G., and Loperfido, N. (2005), "Generalized Skew-Elliptical Distributions and Their Quadratic Forms," *Annals of the Institute of Statistical Mathematics*, 57, 389–401. [1090]

Ma, Y., and Genton, M. G. (2004), "A Flexible Class of Skew-Symmetric Distributions," *Scandinavian Journal of Statistics*, 31, 459–468. [1090]

Ma, Y., Genton, M. G., and Tsiatis, A. A. (2005), "Locally Efficient Semiparametric Estimators for Generalized Skew-Elliptical Distributions," *Journal of the American Statistical Association*, 100, 980–989. [1099]

Ma, Y., and Hart, J. (2007), "Constrained Local Likelihood Estimators for Semiparametric Skew-Normal Distributions," *Biometrika*, 94, 119–134. [1099,1102]

Marchenko, Y. V., and Genton, M. G. (2012), "A Heckman Selection-*t* Model," *Journal of the American Statistical Association*, 107, 304–317. [1091]

Rao, C. R. (1985), "Weighted Distributions Arising Out of Methods of Ascertainment: What Populations Does a Sample Represent?," in *A Celebration of Statistics: The ISI Centenary Volume*, eds. A. C. Atkinson and S.E. Fienberg, New York: Springer-Verlag, pp. 543–569. [1090]

Savchuk, O. Y., Hart, J. D., and Sheather, S. J. (2010), "Indirect Cross-Validation for Density Estimation," *Journal of the American Statistical Association*, 105, 415–423. [1096,1098]

Serfling, R. J. (2002), *Approximation Theorems of Mathematical Statistics*, New York: Wiley. [1103]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [1092]

Wang, J., Boyer, J., and Genton, M. G. (2004), "A Skew-Symmetric Representation of Multivariate Distributions," *Statistica Sinica*, 14, 1259–1270. [1090]