EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR NONLINEAR REGRESSIONS AND MODELS UNDER SAMPLE SELECTION BIAS

A Dissertation

by

MI JEONG KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Statistics

EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR NONLINEAR REGRESSIONS AND MODELS UNDER SAMPLE SELECTION BIAS

A Dissertation

by

MI JEONG KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Yanyuan Ma
Committee Members,	Marc G. Genton
	Mohsen Pourahmadi
	Guido Kanschat
Head of Department,	Simon Sheather

August 2012

Major Subject: Statistics

ABSTRACT

Efficient Semiparametric Estimators for Nonlinear Regressions
and Models Under Sample Selection Bias. (August 2012)
Mi Jeong Kim, B.S., Ewha University;
M.S., Ewha University
Chair of Advisory Committee: Dr. Yanyuan Ma

We study the consistency, robustness and efficiency of parameter estimation in different but related models via semiparametric approach. First, we revisit the second-order least squares estimator proposed in Wang and Leblanc (2008) and show that the estimator reaches the semiparametric efficiency. We further extend the method to the heteroscedastic error models and propose a semiparametric efficient estimator in this more general setting. Second, we study a class of semiparametric skewed distributions arising when the sample selection process causes sampling bias for the observations. We begin by assuming the anti-symmetric property to the skewing function. Taking into account the symmetric nature of the population distribution, we propose consistent estimators for the center of the symmetric population. These estimators are robust to model mis-specification and reach the minimum possible estimation variance. Next, we extend the model to permit a more flexible skewing structure. Without assuming a particular form of the skewing function, we propose both consistent and efficient estimators for the center of the symmetric population using a semiparametric method. We also analyze the asymptotic properties and derive the corresponding inference procedures. Numerical results are provided to support the results and illustrate the finite sample performance of the proposed estimators.

To my parents

ACKNOWLEDGMENTS

First and foremost, I would like to thank God. Whenever I felt frustrated, I meditated on the words He said to me right before I started my Ph.D. study. It is Joshua 1:9, "Be strong and courageous. Do not be afraid; do not be discouraged, for the Lord your God will be with you wherever you go." Looking back on hard times, I feel my God was with me as He was with Moses.

I am forever grateful to my supervisor, Dr. Yanyuan Ma, for her guidance, encouragement and patience shown to me during the development of this dissertation and my years as a Ph.D. student. Her great insight into statistical methodology has led me to do statistical research. I was touched by her warm heart and great care for me. I also would like to thank Dr. Marc Genton for his guidance throughout my research. In particular, I improved my statistical knowledge and insight through the Skew-Tea Seminar he lead. Thanks to Dr. Mohsen Pourahmadi, who encouraged me and advised me to do presentations well. I will miss his humor. I also would like to thank Dr. Guido Kanschat in Math Department for serving on my committee.

I would like to thank the faculty of the Department of Statistics at Texas A&M University. Special thanks go to my colleagues, Tanya, Deep, Pryia, Xinxin and Mr. Jincheol Park for good times and valuable discussions during my years as a Ph.D student.

Finally, I would like to thank my parents, my sister and my brother for their constant love and support.

TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION	1
	A. Overview of Semiparametric Theory	2
	B. The Second-order Least Squares Estimator	3
	C. Sample Selection Bias	4
II	THE EFFICIENCY OF THE SECOND-ORDER NONLINEAR LEAST SQUARES ESTIMATOR AND ITS EXTENSION*	7
	A. Efficiency Results	9
	B. Extension	12
	C. Numerical Results	15
	D. Discussion	19
III	SEMIPARAMETRIC ESTIMATION OF THE CENTER OF AN UNKNOWN SYMMETRIC POPULATION UNDER SELECTION BIAS	22
	A. Estimation	24
	1. Semiparametric Derivation	24
	2. Robust Estimation Family	25
	3. Efficient Estimation	29
	4. Population Density Estimation	33
	B. Simulations	34
	C. Data Example	38
	D. Discussion	39
IV	EFFICIENT AND ROBUST ESTIMATION USING BIASED SAMPLES	45
	A. Consistent Estimation Under Misspecified f	47
	1. The Estimator Family	47
	2. Locally Efficient Estimators and Their Robustness	49
	B. Efficiency Considerations	50
	1. Improving the Estimation Efficiency	50
	2. Efficient Estimation of μ	52
	C. Asymptotic Properties	54

CHAPTER Page

D. Numerical Performance 58
1. Simulations 58
2. Ambulatory Expenditures Data 61
E. Discussion 62

V CONCLUSION 70

REFERENCES 72

APPENDIX A 77

APPENDIX B 81

APPENDIX C 100

VITA 112

LIST OF TABLES

TABLE		Page
I	Simulation results for exponential and growth mean model with homoscedastic error. The average and sample variance (VAR) of 1000 OLS, SLS and SE estimators, as well as the average of the 1000 estimated variances (VAR1) for the SE estimator are presented. Results are based on a sample size of 200.	17
II	Simulation results for exponential and growth mean model with homoscedastic error. The average and sample variance (VAR) of 1000 OLS, SLS and SE estimators, as well as the average of the 1000 estimated variances (VAR1) for the SE estimator are presented. Results are based on a sample size of 500.	18
III	Simulation results for exponential and growth mean model with heteroscedastic error. The average and sample variance (VAR) of 1000 WLS, SE1 and SE2 estimators are presented. SE1 is the SE estimator with the true moment models, and SE2 the wrong moment models. Median of the 1000 estimated variances (VAR1) for SE1 and SE2 are calculated. Results are based on a sample size of 400.	19
IV	Simulation results for exponential and growth mean model with heteroscedastic error. The average and sample variance (VAR) of 1000 WLS, SE1 and SE2 estimators are presented. SE1 is the SE estimator with the true moment models, and SE2 the wrong moment models. Median of the 1000 estimated variances (VAR1) for SE1 and SE2 are calculated. Results are based on a sample size of 1000.	20
V	Median of 1000 estimates of μ and β and their sample standard deviation (sd) and average of the estimated standard deviation ($\widehat{\text{sd}}$) for simulations 1-4. True values are $\mu_0 = 4$, $\beta_0 = 2$. Results based on sample size $n = 500$.	36

TABLE Page

VI Five estimates of μ and β and their estimated variance and standard deviation for the ambulatory expenditures data. 39

VII Results of the four simulation studies. Mean, sample standard deviation (sd), average of the estimated standard deviation ($\widehat{\text{sd}}$), and the 95% coverage probabilities of μ and β are reported. Results are obtained with sample size $n = 500$ and 1000 simulations. 68

VIII Five estimates of μ and β and their estimated standard deviation for the ambulatory expenditures data, under two weighting function models. The first weighting function is $w(x, \beta) = e^{-e^{-\beta x}}$, the second weighting function is $w(x, \beta) = \frac{e^{-\beta x}}{(1+e^{-\beta x})^2}$. 69

LIST OF FIGURES

FIGURE Page

1 Pointwise quantile curves from simulation 1. 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 in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively. 40

2 Pointwise quantile curves from simulation 2. 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 in simulation 2. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively. 41

3 Pointwise quantile curves from simulation 3. 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 in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively. 42

4 Pointwise quantile curves from simulation 4. 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 in simulation 4. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively. 43

5 The estimated densities of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ for the ambulatory expenditures data. The estimated sample density curve is overlayed on the histogram of the observations. 44

FIGURE Page

6 Pointwise quantile curves from simulation 1. 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 in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively. 63

7 Pointwise quantile curves from simulation 2. 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 in simulation 2. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively. 64

8 Pointwise quantile curves from simulation 3. 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 in simulation 3. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively. 65

9 Pointwise quantile curves from simulation 4. 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 in simulation 4. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively. 66

10 The estimated densities of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ for the ambulatory expenditures data (right), under the first (upper) and second (lower) weighting functions. The estimated sample density curve is overlayed on the histogram of the observations. 67

CHAPTER I

INTRODUCTION

We study the consistency, robustness and efficiency of parameter estimation in non-linear regressions and models under sample selection bias using the semiparametric method. In Chapter 2, we consider the model proposed in Wang and Leblanc (2008). Wang and Leblanc introduced a second-order least squares estimator which minimizes the distances of the response variable and the squared response variable to its first and second conditional moments simultaneously. We are interested that their estimator indeed reaches the efficiency bound in the sense of semiparametrics. We also propose the semiparametric efficient estimator and show its asymptotic properties under the same assumptions Wang and Leblanc made. Furthermore, we extend the model to the heteroscedastic error model and demonstrate our estimator is consistent, robust and efficient. In Chapter 3 and 4, we propose methods of estimation for the center of a symmetric population when a representative sample of the population is unavailable due to selection mechanism. We do not impose any parametric form on the population distribution. In Chapter 3, we assume the anti-symmetric property to the selection function, i.e., $\pi(t, \beta) + \pi(-t, \beta) = 1$ for all $t \in \mathbb{R}$. In Chapter 4, we consider a more general form of the selection function. Based on semiparametric theory and taking into account the symmetric nature of the population distribution, we propose consistent, robust, and efficient estimators. In order to improve the efficiency, we adopt a

---

The style of this dissertation follows *Annals of the Institute of Statistical Mathematics*.

modified nonparametric kernel density estimation. We demonstrate the theoretical properties of our estimators through asymptotic analysis. Numerical experiments are provided to advocate our theory and a real data analysis is also given to illustrate the applicability of the methods in practice.

In this chapter, we first present semiparametric theory in Section A, and introduce the second-order least squares estimator proposed in Wang and Leblanc (2008) in Section B. In Section C, we describe the selection bias issues in sampling.

A. Overview of Semiparametric Theory

A large class of estimators in semiparametric models is defined by regular asymptotically linear (RAL) estimators (Newey, 1990). An RAL estimator $\widehat{\theta}$ is uniquely linked to an influence function through

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i; \theta_0) + o_p(1),$$

where θ_0 denotes the true value of the finite-dimensional parameter of interest θ, and $\psi(X_i; \theta_0)$ is the influence function associated with the ith observation, X_i, $i = 1, \ldots, n$. For RAL estimators, the corresponding influence function satisfies $E(\psi) = 0$ and $E(\psi\psi^T)$ is finite and nonsingular. Here and later, when not explicitly pointed out, all the expectations are calculated under the density that defines the true data generation process. The relation between RAL estimators and influence functions allows one to construct RAL estimators through finding influence functions.

A geometric approach taken in Bickel et al. (1993) views all the mean zero, finite variance functions as a Hilbert space, where inner product is defined to be the covariance. The Hilbert space is then decomposed into a nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$. It further establishes that every non-zero function in

$\Lambda^{\perp}$ can be normalized to a valid influence function, and the orthogonal projection of the score function to $\Lambda^{\perp}$ corresponds to the efficient influence function: the influence function with the minimum variance.

To find the nuisance tangent space Λ, we first construct parametric submodels which are models included in the semiparametric model and contain the true density. The nuisance tangent space Λ is the closed subspace generated by the linear span of the nuisance score functions of all such parametric submodels. Then $\Lambda^{\perp}$ is the orthogonal complement of Λ in the Hilbert space.

B. The Second-order Least Squares Estimator

Wang and Leblanc considered the general regression model

$$Y = m(X;\beta) + \epsilon,$$

where $Y \in \mathbb{R}$ is the response variable, $X \in \mathbb{R}^k$ is the predictor variable, $\beta \in \mathbb{R}^p$ is the unknown regression parameter and ϵ is the random error satisfying $E(\epsilon|X) = 0$ and $E(\epsilon^2|X) = \sigma^2$. Y and ϵ are assumed to have finite fourth moments. The parameter vector and the parameter space are denoted as $\theta = (\beta^T, \sigma^2)^T$ and $\Theta \subset \mathbb{R}^{p+1}$, respectively. The true parameter value is denoted by $\theta_0 = (\beta_0^T, \sigma_0^2)^T \in \Theta$.

Assume (Y_i, X_i^T), $i = 1, 2, \cdots, n$ is an *i.i.d.* random sample. Then the second-order least squares estimator (SLSE) $\widehat{\theta}$ is defined as the measurable function that minimizes

$$Q_n(\theta) = \sum_{i=1}^{n} \rho_i^T(\theta) W_i \rho_i(\theta),$$

where $\rho_i(\theta) = (Y_i - m(X_i;\beta), Y_i^2 - m^2(X_i;\beta) - \sigma^2)^T$ and $W_i = W(X_i)$ is a 2×2 nonnegative definite matrix. Under the assumptions they made in Wang and Leblanc

(2008), they demonstrated consistency of SLSE and calculated the optimal weighting matrix W corresponding to the most efficient SLSE.

The asymptotic variance of the most efficient SLSE of each parameter is given by

$$\begin{aligned} V(\widehat{\beta}_{\mathrm{SLS}}) &= \left(\sigma_0^2 - \frac{\mu_3^2}{\mu_4 - \sigma_0^4}\right)\left(M_2 - \frac{\mu_3^2}{\sigma_0^2(\mu_4 - \sigma_0^4)} M_1 M_1^T\right)^{-1}, \\ V(\widehat{\sigma}^2_{\mathrm{SLS}}) &= \frac{(\mu_4 - \sigma_0^4)\{\sigma_0^2(\mu_4 - \sigma_0^2) - \mu_3^2\}}{\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2 M_1^T M_2^{-1} M_1}, \end{aligned}$$

where μ_3 and μ_4 are the third and the fourth moment of ϵ respectively and

$$M_1 = E\left[\frac{\partial m(X;\beta_0)}{\partial \beta}\right], \quad M_2 = E\left[\frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\partial m(X;\beta_0)}{\partial \beta^T}\right].$$

They have proved the SLSE of β is more efficient than the ordinary least squares estimator. Our goal is to verify that SLSE reaches the optimal efficiency bound concerning the estimation variability. We approach the model using a semiparametric method and propose a semiparametric efficient estimator in Chapter 2. By comparing variances of SLSE and our estimator, we investigate whether SLSE is efficient or not. Furthermore, we extend the model to the heteroscedastic error model.

C. Sample Selection Bias

Estimating the population center is an important issue in sampling. If a random sample precisely mirrors the data in the population, it is not a difficult problem to estimate the population of interest using a sample. However, in reality, a biased sample might be obtained even if the sampling and measuring are conducted with accuracy because it is not a census but a sample. Selection bias can be produced if a sample is obtained from a selected part of the population. Selection bias often occurs when a sample is truncated, censored or includes missing data. A classic example of

a selection bias is telephone surveys conducted in an age where telephones were not prevalent to most ordinary people.

We consider the following model which reflects selection bias. Assume biased observations $X_1, \cdots, X_n$ are independent and identically distributed with density

$$g(x; \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(x; \boldsymbol{\beta}) \frac{w(x; \boldsymbol{\beta}, \boldsymbol{\alpha})}{E\{w(x; \boldsymbol{\beta}, \boldsymbol{\alpha})\}}, \tag{1.1}$$

where f is the population density and $\boldsymbol{\beta}$ is a q-dimensional vector of unknown parameters. Although the population density is f, a sample has a density g because sampling bias is caused by the sample selection process. A weight function w captures such a selection mechanism and $\boldsymbol{\alpha}$ denotes some r-dimensional vector of additional unknown parameter. This model is called a selection model. For instance, consider two independent random variables X^* and Y, each symmetrically distributed around zero, with X^* having density f and Y having cumulative distribution function H. We observe X if and only if $Y < \beta X^*$, $\beta \in \mathbb{R}$, in which case we set $X = X^*$. Then $\text{pr}(X \leq x) = \text{pr}(X^* \leq x | Y < \beta X^*)$, we have $w(x; \beta) = H(\beta x)$ in the above model. When $f = \phi$ and $H = \Phi$, g is a skew-normal distribution.

Selection bias issues have been acknowledged and modeled extensively, see for example Rao (1985) and more recently Arellano-Valle et al. (2006). When special restrictions are imposed on either the population model f or the selection weights w, it reduces to various special models in the literature. For example, when w satisfies an anti-symmetric property, $w(x; \boldsymbol{\beta}, \boldsymbol{\alpha}) + w(x; \boldsymbol{\beta}, \boldsymbol{\alpha}) = 1$ for all $x \in \mathbb{R}$, Copas and Li (1997), Arnold and Beaver (2002), Azzalini and Capitanio (2003), Ma and Genton (2004), Wang et al. (2004) and many others have described the types of selection mechanisms that lead to (1.1). When f is further assumed to belong to the elliptical family, (4.1) is reduced to the generalized skew-elliptical distribution (Genton and Loperfido, 2005), which includes 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 further details. Other special cases of (1.1) that do not satisfy the anti-symmetric 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).

In Chapter 3 and 4, we assume the population density f is unknown but symmetric and do not impose parametric form on f. We aim at estimating the population center under the following assumptions on the selection function. In Chapter 3, we assume anti-symmetric property, i.e., $w(x;\boldsymbol{\beta},\boldsymbol{\alpha}) + w(-x;\boldsymbol{\beta},\boldsymbol{\alpha}) = 1$. To relax the restriction on w, we consider a more general form of the selection function w in Chapter 4.

CHAPTER II

THE EFFICIENCY OF THE SECOND-ORDER NONLINEAR LEAST SQUARES ESTIMATOR AND ITS EXTENSION*

Wang and Leblanc (2008) considered a regression model with a parametric mean function and a constant variance:

$$Y = m(X; \beta) + \epsilon, \tag{2.1}$$

where they assumed that Y is a one-dimensional continuous response variable, and X is a covariate vector that can be continuous, discrete, or mixed. The mean function m is a known function up to the d-dimensional parameter β and the model error ϵ satisfies the usual mean zero assumption $E(\epsilon|X) = 0$. In addition, they also assumed that ϵ has a constant yet unknown variance σ^2, that is, $E(\epsilon^2|X) = \sigma^2$. The observations are denoted $(X_1, Y_1), \ldots, (X_n, Y_n)$, each satisfies (2.1) and the n observations are independent of each other. Without the additional homoscedastic assumption, this is the usual semiparametric regression problem, or sometimes named the restricted moment model, and the consistent estimator family, as well as the efficient estimator for β is known. See, for example, Tsiatis (2006, p 53). With the additional assumption of homoscedasticity, Wang and Leblanc (2008) proposed a second-order least squares (SLS) type estimation procedure where they take advantage of the knowledge of the

---

* Reprinted with permission from "The efficiency of the second-order nonlinear least squares estimator and its extension" by Kim, M. and Ma, Y., 2011, *Annals of the Institute of Statistical Mathematics*, DOI: 10.1007/s10463-011-0332-y, Copyright by the Institute of Statistical Mathematics.

second moment of $Y - m(X;\beta)$. They showed that this SLS estimation of β indeed yields improvement over classical least squares estimation.

This naturally motivates us to ask: Can further improvement be obtained? In other words, we are curious to find out whether or not the Wang and Leblanc (2008) estimator reaches the optimal efficiency bound in the sense of Bickel et al. (1993). Studying the semiparametric efficiency bound is important in understanding a model. It provides an ultimate conclusion when searching for estimators or trying to improve existing procedures. Only when an efficient estimator is obtained, the procedure of estimation can be considered to have reached certain optimality. Researchers have been searching for optimal estimators in various problems, the most familiar example being the ordinary and weighted least square estimators in the regression setting. Efficiency issues are also considered in more complex problems such as the Cox model (Tsiatis, 2006, Chapter 5.2, p 113), a class of general survival models (Zeng and Lin, 2008), problems in case control designs (Rabinowitz, 2000, Ma, 2010) or involving auxiliary information (Chen et al, 2008), the partially linear models (Chamberlaine, 1992, Ma et al., 2006), the latent variable models (Ma and Genton, 2010), the functional estimation in semiparametric models (Maity et al., 2007, Müller, 2009), the regression with missing covariates (Robins et al., 1994), the skewed distribution families (Ma et al., 2005, Ma and Hart, 2006), the quantile regression models (Newey and Powell, 1990) and the measurement error models (Tsiatis and Ma, 2004, Ma and Carroll, 2006). To answer the question of optimality in our problem, we view the model in (2.1) as a semiparametric problem and take a geometric approach. We construct the locally efficient semiparametric estimators, and proceed to identify the optimal semiparametric efficient (SE) estimator. The SE has the classical root-n convergence rate and is asymptotically normal. We further derive the estimation variance of the SE estimator, which reaches the semiparametric efficiency bound and compare with

the result in Wang and Leblanc (2008). It demonstrates that the SLS estimator by Wang and Leblanc (2008) is semiparametrically efficient as well and is thus optimal asymptotically. The resulting estimator, which is asymptotically optimal, is new in literature.

In order to relax the homoscedasticity assumption on ϵ, we subsequently assume $E(\epsilon^2|X) = \sigma^2(X;\gamma)$, which is a function of X with unknown parameter γ. Note that here σ^2 has a known functional form. This model certainly includes the constant variance model as a special case. Although the model is more complex, we can easily adapt the analysis we performed and derive the optimal efficient estimator. The estimator and its asymptotic optimality is also new in literature.

The rest of the chapter is organized as the following. We introduce the semiparametric method and show the efficiency of the SLS estimator in Section A. In Section B, we adapt our method to the heteroscedastic error models and propose the efficient estimators and the corresponding variance estimation. We also show the asymptotic optimality of the generalized estimator in this section. Numerical experiments are provided in Section C, and we discuss some possible further extensions in Section E. Technical details are provided in an Appendix A.

A. Efficiency Results

For convenience, we denote the augmented parameter $\theta = (\beta^T, \sigma^2)^T$, and aim at finding the class of consistent semiparametric estimators for θ and identifying the most efficient one within this class. The probability density function (pdf) of a single observation (X, Y), ignoring the subscripts, can be written as

$$p_{X,Y}(x,y) = p_X(x)p_{\epsilon|X}\left\{y - m(x;\beta)|X = x\right\} = \eta_1(x)\eta_2\left\{y - m(x;\beta), x\right\}, \qquad (2.2)$$

where $\eta_2(\cdot)$ satisfies $\int \epsilon\eta_2(\epsilon, x)d\epsilon = 0, \int \epsilon^2\eta_2(\epsilon, x)d\epsilon = \sigma^2$, and the third and fourth moments of ϵ conditional on X are constants. Here we use $\eta_1(\cdot), \eta_2(\cdot, \cdot)$ to denote the pdf of X and the conditional pdf of ϵ given X. This emphasizes that these pdfs are infinite-dimensional nuisance parameters. We sometimes write $p_{X,Y}(x, y)$ as $p_{X,Y}(x, y; \theta, \eta_1, \eta_2)$ to emphasize that the pdf contains a finite dimensional parameter θ and infinite dimensional parameters η_1, η_2. We also use ϵ and $Y - m(X; \beta)$ interchangeably, and use a subscript $_0$ to denote the true parameter value.

The geometric approach to semiparametric regression analysis consists of defining a Hilbert space $\mathcal{H}$ and finding two subspaces of $\mathcal{H}$, namely the nuisance tangent space Λ and its orthogonal complement $\Lambda^\perp$. Here, the Hilbert space is the space of all mean zero, length d, finite variance functions of (X, Y). Here and after, all the expectations are calculated under the true distribution. The subspace Λ is a space spanned by the nuisance score functions (the score function obtained through taking derivative of the logarithm of the pdf with respect to the nuisance parameter) of all the parametric submodels of (2.2) and their limiting points. The subspace $\Lambda^\perp$ is a space consisting of all the functions that are orthogonal to all the functions in Λ. See Tsiatis (2006, Chapter 4) for an elaborated explanation of these concepts. Once Λ and $\Lambda^\perp$ are obtained, we then project the score function $S_\theta = \partial \log p_{X,Y}/\partial\theta$ onto $\Lambda^\perp$ to obtain S_{eff}. The orthogonal projection of the score function onto $\Lambda^\perp$, i.e. S_{eff}, is usually referred to as the efficient score function. If it can be constructed, $\sum_{i=1}^n S_{\text{eff}}(X_i, Y_i; \beta, \eta_{10}, \eta_{20}) = 0$ will then be the estimating equation that will yield the optimal estimator of θ.

For model (2.1), a careful analysis yields $\Lambda = \Lambda_{\eta_1} \oplus \Lambda_{\eta_2}$, where

$$\begin{aligned} \Lambda_{\eta_1} &= \{f(X) : E(f) = 0\}, \\ \Lambda_{\eta_2} &= \{g(\epsilon, X) : E(g|X) = 0, E(\epsilon g|X) = 0, E(\epsilon^2 g|X) = 0\}, \\ \text{and} \quad \Lambda^{\perp} &= \{h(\epsilon, X) : h = a(X)\epsilon + b(X)(\epsilon^2 - \sigma^2)\}, \end{aligned}$$

where a, b, f are all length d functions of X, and g, h are length d functions of ϵ, X. The derivation details are in Appendix A1. Taking derivative of the logarithm of the pdf with respect to θ gives us the score function

$$S_\theta(X, Y) = (S_\beta^T, S_{\sigma^2})^T = \left\{ -\frac{\partial \eta_2(\epsilon, X)}{\eta_2(\epsilon, X)\partial \epsilon} \frac{\partial m(X;\beta)}{\partial \beta^T}, \frac{\partial \eta_2(\epsilon, X)}{\eta_2(\epsilon, X)\partial \sigma^2} \right\}^T. \tag{2.3}$$

As pointed out in Appendix A2, the projection of an arbitrary function $h(X, Y) \in \mathcal{H}$ onto $\Lambda^{\perp}$ can be calculated as

$$\Pi(h|\Lambda^{\perp}) = \frac{E(\epsilon h|X)}{\sigma^2}\epsilon + \frac{E(Ch|X)}{E(C^2|X)}C, \tag{2.4}$$

where $C = \epsilon^2 - \sigma^2 - E(\epsilon^3 - \epsilon\sigma^2|X)\epsilon/\sigma^2$. Letting $h(X, Y) = S_\theta(X, Y)$, we can obtain $S_{\text{eff}} = (S_{\beta,\text{eff}}^T, S_{\sigma^2,\text{eff}})^T$. Further solving $\sum_{i=1}^n S_{\text{eff}}(X_i, Y_i) = 0$ gives the SE estimator, and the asymptotic covariance matrix of $\sqrt{n}\widehat{\theta}$ is

$$n\text{cov}(\widehat{\theta}) = \{E(S_{\text{eff}} S_{\text{eff}}^T)\}^{-1}.$$

More specifically, we have

Theorem 1. *The efficient score functions for β and σ^2 have the form*

$$\begin{aligned} S_{\beta,eff}(X, Y) &= \frac{\partial m(X;\beta)}{\partial \beta}\left\{ \frac{\epsilon}{\sigma^2} - \frac{E(\epsilon^3|X)C}{\sigma^2 E(C^2|X)} \right\}, \\ S_{\sigma^2,eff}(X, Y) &= \frac{C}{E(C^2|X)}. \end{aligned} \tag{2.5}$$

The estimation covariance matrix is

$$E(S_{\mathit{eff}}S_{\mathit{eff}}^T|_{\theta=\theta_0})^{-1}$$
$$= \begin{bmatrix} \left(\sigma_0^2 - \frac{\mu_3^2}{\mu_4-\sigma_0^4}\right)\left\{B - \frac{\mu_3^2}{\sigma_0^2(\mu_4-\sigma_0^4)}AA^T\right\}^{-1} & \frac{\mu_3\{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2\}B^{-1}A}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2A^TB^{-1}A} \\ \frac{\mu_3\{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2\}A^TB^{-1}}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2A^TB^{-1}A} & \frac{(\mu_4-\sigma_0^4)\left\{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2\right\}}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2A^TB^{-1}A} \end{bmatrix}, \quad (2.6)$$

where $\mu_3 = E(\epsilon^3|X)$, $\mu_4 = E(\epsilon^4|X)$, *and*

$$A = E\left\{\frac{\partial m(X;\beta_0)}{\partial\beta}\right\}, \quad B = E\left\{\frac{\partial m(X;\beta_0)}{\partial\beta}\frac{\partial m(X;\beta_0)}{\partial\beta^T}\right\}.$$

The proof of Theorem 1 consists the derivation of the efficient score (2.5) , given in Appendix A3, and the derivation of (2.6), given in Appendix A4. Comparing the variances in (2.6), and (7), (8) in Wang and Leblanc (2008), we obtain that their SLS estimator is indeed efficient.

We would like to point out that for convenience, we have assumed both μ_3 and μ_4 are constants, and we estimate them using the residuals from an initial OLS estimator. The same assumptions are made in Wang and Leblanc (2008). If however, these assumptions are not valid, we still have $E(C|X) = E(\epsilon^2|X) - \sigma^2 - \mu_3^* E(\epsilon|X)/\sigma^2 = 0$, where we use μ_3^* to denote $E(\epsilon^3 - \epsilon\sigma^2|X)$ calculated under the wrong model. From the efficient score in (2.5), it is easily verified that we still have $E(S_{\text{eff}}|X) = 0$, hence our estimator remains consistent. On the other hand, if these assumptions are indeed valid, then our procedure achieves the optimal efficiency.

B. Extension

In Section A, we developed the efficient semiparametric estimator and its estimation variance in the model (2.1) with the homoscedasticity error. In this Section, we extend the results to the heteroscedastic error case. We assume $E(\epsilon^2|X) = \sigma^2(X;\gamma)$, where the variance function σ^2 is known up to an unknown parameter γ. We denote

the parameter $\theta = (\beta^T, \gamma^T)^T$. In this case, similar derivations show that the nuisance tangent space, now denoted Ω, can still be expressed as $\Omega = \Omega_{\eta_1} + \Omega_{\eta_2}$, where $\Omega_{\eta_1} = \Lambda_{\eta_1}$ is unchanged from the homoscedastic case, and

$$\Omega_{\eta_2} = \left\{ g(\epsilon, X) : E(g|X) = 0, E(\epsilon g|X) = 0, E\left[\{\epsilon^2 - \sigma^2(X;\gamma)\}g|X\right] = 0 \right\}.$$

Consequently,

$$\Omega^{\perp} = \left\{ h(\epsilon, X) : h = a(X)\epsilon + b(X)\{\epsilon^2 - \sigma^2(X;\gamma)\} \right\}.$$

Here, a, b, g, h are all length d functions.

The score function

$$S_\theta(X, Y) = (S_\beta^T, S_\gamma^T)^T = \left\{ -\frac{\partial \eta_2(\epsilon, X)}{\eta_2(\epsilon, X)\partial \epsilon} \frac{\partial m(X;\beta)}{\partial \beta^T}, \frac{\partial \eta_2(\epsilon, X)}{\eta_2(\epsilon, X)\partial \sigma^2} \frac{\partial \sigma^2(X;\gamma)}{\partial \gamma^T} \right\}^T$$

can be similarly calculated by differentiating the logarithm of the pdf with respect to θ. The projection of an arbitrary function $h(X, Y) \in \mathcal{H}$ onto $\Omega^{\perp}$ also has a form similar to (2.4), that is

$$\Pi(h|\Omega^{\perp}) = \frac{E(\epsilon h|X)}{\sigma^2(X;\gamma)}\epsilon + \frac{E(Ch|X)}{E(C^2|X)}C,$$

where $C = \epsilon^2 - \sigma^2(X;\gamma) - E\{\epsilon^3 - \epsilon\sigma^2(X;\gamma)|X\}\epsilon/\sigma^2(X;\gamma)$. Projecting S_θ onto $\Omega^{\perp}$, we can obtain $S_{\text{eff}} = (S_{\beta,\text{eff}}^T, S_{\gamma,\text{eff}}^T)^T$. The SE estimator can therefore be obtained through solving $\sum_{i=1}^n S_{\text{eff}}(X_i, Y_i) = 0$, and the asymptotic covariance matrix of the resulting estimator satisfies $n\text{cov}(\widehat{\theta}) = \{E(S_{\text{eff}} S_{\text{eff}}^T)\}^{-1}$ evaluated at $\theta = \theta_0$. We summarize the results parallel to Theorem 1 in Theorem 2, but omit the detailed proofs.

Theorem 2. *The efficient score functions for β and γ have the form*

$$\begin{aligned} S_{\beta,\mathit{eff}}(X,Y) &= \frac{\partial m(X;\beta)}{\partial \beta}\left\{\frac{\epsilon}{\sigma^2(X;\gamma)} - \frac{E(\epsilon^3|X)C}{\sigma^2(X;\gamma)E(C^2|X)}\right\}, \\ S_{\gamma,\mathit{eff}}(X,Y) &= \frac{C}{E(C^2|X)}\frac{\partial \sigma^2(X;\gamma)}{\partial \gamma}. \end{aligned}$$

The estimation covariance matrix is

$$\begin{aligned} & E\left(S_{\mathit{eff}}S_{\mathit{eff}}^T\right) \\ = \; & E\begin{bmatrix} \frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\partial m(X;\beta_0)}{\partial \beta^T}\frac{1}{\sigma^2(X;\gamma_0)}\left\{1+\frac{\mu_3^2}{\sigma^2(X;\gamma_0)E(C^2|X)}\right\} & -\frac{\mu_3}{\sigma^2(X;\gamma_0)E(C^2|X)}\frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\partial \sigma^2(X;\gamma_0)}{\partial \gamma^T} \\ -\frac{\mu_3}{\sigma^2(X;\gamma_0)E(C^2|X)}\frac{\partial \sigma^2(X;\gamma_0)}{\partial \gamma}\frac{\partial m(X;\beta_0)}{\partial \beta^T} & \frac{1}{E(C^2|X)}\frac{\partial \sigma^2(X;\gamma_0)}{\partial \gamma}\frac{\partial \sigma^2(X;\gamma_0)}{\partial \gamma^T} \end{bmatrix}, \end{aligned}$$

where μ_3 is defined after (2.6).

The upper-left block of the inverse of $E(S_{\mathrm{eff}}S_{\mathrm{eff}}^T)$ gives the covariance matrix of the efficient estimator $n\widehat{\beta}$. Specifically, denote

$$\begin{aligned} U_1 &= m'_\beta(X;\beta_0)\mu_3\sigma^{-2}(X;\gamma_0)/\sqrt{E(C^2|X)}, \\ U_2 &= \frac{\partial \sigma^2(X;\gamma_0)}{\partial \gamma}/\sqrt{E(C^2|X)}, \end{aligned}$$

where $m'_\beta(X;\beta_0)$ denotes $\partial m(X;\beta)/\partial \beta^T$ evaluated at $\beta=\beta_0$, we have

$$n\mathrm{cov}(\widehat{\beta}) = \left[E\left\{m'_\beta(X;\beta_0)m'_\beta(X;\beta_0)^T\sigma^{-2}(X;\gamma_0)\right\} + E(U_1U_1^T) - E(U_1U_2^T)\left\{E(U_2U_2^T)\right\}^{-1}E(U_1U_2)^T\right]^{-1}.$$

In contrast to the efficient semiparametric estimator, we inspect the usual weighted least square (WLS) estimator where $\sum_{i=1}^n w_i\{Y_i - m(X_i;\beta)\}^2$ is minimized to obtain the WLS estimator $\widetilde{\beta}$. In this case, it is well known that the optimal weights are $w_i = 1/\sigma^2(X_i;\gamma_0)$, and the corresponding optimal estimator in the WLS family has asymptotic estimation covariance matrix

$$n\mathrm{cov}(\widetilde{\beta}) = E\left\{\sigma^{-2}(X_i;\gamma_0)m'_\beta(X_i;\beta_0)m'_\beta(X_i;\beta_0)^T\right\}^{-1}.$$

The covariance matrices of the two estimators are obviously different. In fact, it

is easy to verify that

$$\begin{aligned}\left\{n\mathrm{cov}(\widehat{\beta})\right\}^{-1} - \left\{n\mathrm{cov}(\widetilde{\beta})\right\}^{-1} &= E(U_1U_1^T) - E(U_1U_2^T)\left\{E(U_2U_2^T)\right\}^{-1}E(U_1U_2)^T \\ &= \mathrm{cov}\left[U_1 - E(U_1U_2^T)\left\{E(U_2U_2^T)\right\}^{-1}U_2\right],\end{aligned}$$

hence $n\mathrm{cov}(\widetilde{\beta}) - n\mathrm{cov}(\widehat{\beta})$ is nonnegative-definite. This shows that although the optimal WLS is the most efficient among the WLS estimator family, it is in general not as efficient as the SE estimator that we have derived. The practical improvement of the estimation variance will be demonstrated in the simulation studies in Section 1.

Unlike in the homoscedastic error case, it is no longer reasonable to assume μ_3 and μ_4 to be constants. Similarly to σ^2, they are usually functions of the covariate X, say $\mu_3(X)$ and $\mu_4(X)$. Implementing our efficient estimator requires plugging in $\mu_3(X)$ and $\mu_4(X)$, which are generally unknown. In practice, we could first obtain the residual r_i's of the model from an initial OLS estimator, then fit parametric or nonparametric models for (X_i, r_i^3) and (X_i, r_i^4) to obtain $\widehat{\mu}_3(X)$, $\widehat{\mu}_4(X)$. We then proceed with the estimation on β, γ. Similarly to the homoscedastic case, even if the parametric models are misspecified, or the estimation of $\mu_3(X)$ and $\mu_4(X)$ are completely wrong, our estimator remains consistent. This is because $E(S_{\mathrm{eff}}|X) = 0$ is guaranteed by $E(\epsilon|X) = 0$ and $E(\epsilon^2|X) = \sigma^2(X;\gamma)$, it does not rely on the correctness of $\mu_3(X)$ and $\mu_4(X)$. However, when the model is correct, our estimator is efficient, even when $\mu_3(X)$, $\mu_4(X)$ are estimated rather crudely.

C. Numerical Results

We carry out simulations to study the finite sample performance of the various estimators. The first two simulations focus on homoscedastic error models, as studied in Section A, and the last two simulations on heteroscedastic error models as studied in

Section B.

The mean in simulation one has an exponential form, and the model is

$$Y = \beta_1 \exp(\beta_2 X) + \epsilon, \tag{2.7}$$

with the true parameter values $\beta_1 = 10$, $\beta_2 = -0.6$ and a constant error variance $\sigma^2 = 2$. In simulation two, we considered a growth model

$$Y = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 X)} + \epsilon, \tag{2.8}$$

with the true parameter values $\beta_1 = 10$, $\beta_2 = 1.5$, $\beta_3 = -0.8$ and $\sigma^2 = 2$. Models (2.7) and (2.8) are identical to the simulation settings in Wang and Leblanc (2008). In both models, x_i's are generated from a uniform distribution in $(0, 20)$ and $\epsilon_i = (e_i - 3)/\sqrt{3}$, where e_i's are generated from a $\chi^2(3)$ distribution. Thus, ϵ_i's have mean zero and variance 2 but are asymmetrically distributed. The asymmetry insures that the third moment $E(\epsilon^3|X)$ does not vanish, hence the SE or SLS does not degenerate to the WLS estimator. We implemented the OLS, the SLS and the SE estimators, and report the sample mean and sample variance of these estimates. For the SE estimator, we also calculated the estimated variance.

We used a sample size $n = 200$, and generated 1000 data sets. The simulation results are presented in Table I. These results clearly indicate that all three estimators are consistent, while both the SLS and the SE estimators outperform the OLS in terms of estimation variance. The estimation variances of the SLS and the SE estimator are very close, which supports our claim that both are efficient. Finally, our variance estimation is reasonably precise, in that the sample variance, calculated using the 1000 estimates via standard sample variance calculation, and the estimated variance, calculated as the mean of the 1000 estimated variances, are very close. To further demonstrate the impact of the sample size n, we increased n to 500. The numerical

Table I. Simulation results for exponential and growth mean model with homoscedastic error. The average and sample variance (VAR) of 1000 OLS, SLS and SE estimators, as well as the average of the 1000 estimated variances (VAR1) for the SE estimator are presented. Results are based on a sample size of 200.

	OLS	VAR	SLS	VAR	SE	VAR	VAR1
Exponential model							
$\beta_1 = 10$	10.0768	0.5572	10.0938	0.3418	10.0889	0.3379	0.3721
$\beta_2 = -0.6$	-0.6068	0.0038	-0.6070	0.0024	-0.6068	0.0024	0.0024
$\sigma^2 = 2$	1.9597	0.1024	1.9418	0.0714	1.9319	0.0726	0.0641
Growth model							
$\beta_1 = 10$	9.9981	0.0146	9.9872	0.0125	9.9849	0.0107	0.0119
$\beta_2 = 1.5$	1.5236	0.0489	1.5173	0.0259	1.5166	0.0254	0.0259
$\beta_3 = -0.8$	-0.8104	0.0097	-0.8059	0.0047	-0.8060	0.0047	0.0045
$\sigma^2 = 2$	1.9469	0.1112	1.9179	0.1081	1.8920	0.0921	0.1092

outcome in Table II further suggests the relevancy of our asymptotic results.

In section B, we have seen how the variance of the error can be allowed to depend on X. To experiment with the heteroscedastic error situation, we modified the error structure in the first two simulations to have a variance function $\sigma^2(X;\gamma) = \gamma_1 + \gamma_2 X^2$, while keeping the same mean functions and β values. For the exponential model (2.7), a true value $\gamma = (1, 0.1)$ was used and x_i's are generated from a uniform distribution $(0, 5)$. For the growth model (2.8), we use $\gamma = (2, 0.05)$ and generated x_i's from a uniform distribution $(0, 7)$. In both models, we set $\epsilon_i = e_i - k_i$, where $k_i = \sigma^2(x_i;\gamma)/2$ and e_i's are generated from a $\chi^2(k_i)$ distribution. Thus, the errors ϵ_i's in both (2.7) and (2.8) have mean zero, variance $\sigma^2(X;\gamma)$ and have an asymmetric distribution.

In implementing the WLS estimators, we used the ideal weights $1/\sigma^2(X_i;\gamma_0)$, hence the WLS performance is optimal among all WLS estimators. Implementing the SE estimator requires plugging in the third and fourth conditional moment functions of the error. To test the optimality and robustness of our proposed estimator, we experimented with two different scenarios. In the first case, we calculated the true

Table II. Simulation results for exponential and growth mean model with homoscedastic error. The average and sample variance (VAR) of 1000 OLS, SLS and SE estimators, as well as the average of the 1000 estimated variances (VAR1) for the SE estimator are presented. Results are based on a sample size of 500.

	OLS	VAR	SLS	VAR	SE	VAR	VAR1
Exponential model							
$\beta_1 = 10$	10.0185	0.2224	10.0221	0.1187	10.0228	0.1193	0.1193
$\beta_2 = -0.6$	-0.6025	0.0015	-0.6021	$8.73e^{-4}$	-0.6023	$8.76e^{-4}$	$8.74e^{-4}$
$\sigma^2 = 2$	1.9900	0.0483	1.9789	0.0303	1.9730	0.0303	0.0295
Growth model							
$\beta_1 = 10$	10.0005	0.0061	9.9933	0.0051	9.9953	0.0050	0.0051
$\beta_2 = 1.5$	1.5148	0.0181	1.5084	0.0107	1.5077	0.0105	0.0104
$\beta_3 = -0.8$	-0.8039	0.0032	-0.8029	0.0018	-0.8028	0.0018	0.0019
$\sigma^2 = 2$	1.9843	0.0466	1.9728	0.0463	1.9785	0.0468	0.0482

moment functions and plugged them into the SE estimator (SE1). In the second case, we adopted drastically different functions, and plugged them into the SE estimator as if they were the truth (SE2). To be specific, the true third and fourth conditional moment functions can be calculated to be $a_2X^2+a_1$ and $a_5X^4+a_4X^2+a_3$ respectively, where $a_1 = 4\gamma_1$, $a_2 = 4\gamma_2$, $a_3 = 3\gamma_1^2 + 24\gamma_1$, $a_4 = 24\gamma_2 + 6\gamma_1\gamma_2$ and $a_5 = 3\gamma_2^2$. However, we used the wrong models $a_2X + a_1$ and $a_5X^2 + a_4X + a_3$ instead. Both results along with the optimal WLS results were reported in Table III. These results are based on 1000 simulations with a sample size $n = 400$. The results of Table III reflects the fact that all three estimators are consistent. Compared with the two SE estimators, the WLS estimator, although already optimal in its family, is much less efficient in that the sample variances in estimating β's are much larger than both SEs. We had expected to see SE1 to outperform SE2 substantially. However, to our surprise, the performance of two estimators are rather similar. This is a pleasant surprise, since modeling and estimating the third and fourth conditional moments usually needs very

Table III. Simulation results for exponential and growth mean model with heteroscedastic error. The average and sample variance (VAR) of 1000 WLS, SE1 and SE2 estimators are presented. SE1 is the SE estimator with the true moment models, and SE2 the wrong moment models. Median of the 1000 estimated variances (VAR1) for SE1 and SE2 are calculated. Results are based on a sample size of 400.

	WLS	VAR	SE1	VAR	VAR1	SE2	VAR	VAR1
Exponential model								
$\beta_1 = 10$	9.9931	0.0358	9.9980	0.0217	0.0255	10.0032	0.0229	0.0215
$\beta_2 = -0.6$	-0.5998	$3.51e^{-4}$	-0.5996	$2.27e^{-4}$	$2.32e^{-4}$	-0.5994	$2.98e^{-4}$	$2.14e^{-4}$
$\gamma_1 = 1$			1.0305	0.1328	0.1635	1.0232	0.1436	0.1474
$\gamma_2 = 0.1$			0.0967	0.0016	0.0019	0.0998	0.0019	0.0013
Growth model								
$\beta_1 = 10$	10.0086	0.0668	10.0018	0.0517	0.0497	9.9989	0.0512	0.0490
$\beta_2 = 1.5$	1.5081	0.0093	1.5023	0.0061	0.0061	1.5010	0.0068	0.0059
$\beta_3 = -0.8$	-0.8042	0.0039	-0.8020	0.0020	0.0018	-0.8028	0.0024	0.0018
$\gamma_1 = 2$			1.9768	0.3192	0.2523	1.9993	0.3386	0.2827
$\gamma_2 = 0.05$			0.0498	$9.43e^{-4}$	$7.87e^{-4}$	0.0498	0.0012	$6.68e^{-4}$

large sample size and can be numerically unstable. Finally, the sample variance and estimated variance for both SEs match reasonably well, indicating the validity of our inference. We also increased the sample size to 500 and 1000, and find the two get closer when the sample size increases, numerical results for $n = 500$ and $n = 1000$ are given in Table II and Table IV.

D. Discussion

We have derived a semiparametric efficient estimator in a regression model, where the regression error has conditional mean zero and conditional variance a constant. We have shown that this estimator achieves the optimal semiparametric efficiency bound and is equivalent to the second order least square estimator proposed in Wang and Leblanc (2008), hence revealing an unknown optimality of their estimator. We

Table IV. Simulation results for exponential and growth mean model with heteroscedastic error. The average and sample variance (VAR) of 1000 WLS, SE1 and SE2 estimators are presented. SE1 is the SE estimator with the true moment models, and SE2 the wrong moment models. Median of the 1000 estimated variances (VAR1) for SE1 and SE2 are calculated. Results are based on a sample size of 1000.

	WLS	VAR	SE1	VAR	VAR1	SE2	VAR	VAR1
Exponential model								
$\beta_1 = 10$	10.0073	0.0151	10.0068	0.0076	0.0080	10.0107	0.0075	0.0076
$\beta_2 = -0.6$	-0.5999	$1.42e^{-4}$	-0.5999	$7.96e^{-5}$	$8.09e^{-5}$	-0.6001	$8.73e^{-5}$	$7.85e^{-5}$
$\gamma_1 = 1$			1.0379	0.0578	0.0555	1.0395	0.0632	0.0618
$\gamma_2 = 0.1$			0.0959	$6.36e^{-4}$	$5.77e^{-4}$	0.0963	$7.39e^{-4}$	$5.15e^{-4}$
Growth model								
$\beta_1 = 10$	10.0040	0.0258	10.0062	0.0192	0.0194	10.0057	0.0191	0.0192
$\beta_2 = 1.5$	1.5064	0.0034	1.5019	0.0023	0.0024	1.5006	0.0024	0.0023
$\beta_3 = -0.8$	-0.8036	0.0014	-0.8007	$6.92e^{-4}$	$6.94e^{-4}$	-0.8003	$6.93e^{-4}$	$6.81e^{-4}$
$\gamma_1 = 2$			1.9976	0.1207	0.1127	2.0088	0.1260	0.1242
$\gamma_2 = 0.05$			0.0496	$3.77e^{-4}$	$3.51e^{-4}$	0.0489	$3.78e^{-4}$	$3.26e^{-4}$

further extended the model to the case where the second moment can be an arbitrary function of the covariates, and derived the semiparametric efficient estimator in this general case. The same kind of extension can also be made on the second order least square estimator to handle heteroscedasticity. Simulation results demonstrated the significant improvement of the estimation variance in comparison to the classical WLS estimators and supported the inference procedure.

We have adopted fixed models for the third and fourth conditional moment functions of the error distribution, and demonstrated the consistency of the proposed estimator whether or not these higher moment models are misspecified. However, in reality, these moment functions need to be estimated. We caution that the estimation of the higher moments can be rather unstable, usually requiring a large sample size. Although the need to estimate higher order moments will not affect

the estimation variance of the parameter of interest in the asymptotic sense, in finite samples, it is very likely to inflate the variance. Thus, we propose to adopt simple models for these higher moments. Finally, the same line of analysis can be extended to higher moments, although both the theoretical analysis and the implementation of the estimators will become increasingly complex.

CHAPTER III

SEMIPARAMETRIC ESTIMATION OF THE CENTER OF AN UNKNOWN SYMMETRIC POPULATION UNDER SELECTION BIAS

Suppose that in a general population, a certain trait X is symmetrically distributed with center μ, and we are interested in estimating the population center. We denote the probability density function of X in the population of interest as $f(x-\mu)$, $x \in \mathbb{R}$, where f is an even function. The most common practice is to assume f to be a normal density, however, here we do not impose any other assumption on f except that it is a symmetric density. Nevertheless, because of certain mechanisms involved in the data collection process, only a biased sample from the symmetric population is obtained. Taking the selection bias into account, observations $X_1, \ldots, X_n$ are independent and identically distributed with density

$$2f(x-\mu)\pi(x-\mu;\beta), \quad x \in \mathbb{R}, \tag{3.1}$$

where π is decided by the selection mechanism. Here $\pi(x;\beta) \geq 0$ is usually named a skewing function and it satisfies $\pi(x;\beta) + \pi(-x;\beta) = 1$ for any x. To allow additional flexibility, we allow π to contain an unknown parameter vector β. The skewing function captures the selection bias and we assume its functional form known. However, because no parametric form is assumed on the symmetric function f, (3.1) is a semiparametric model.

The special case of (3.1) where the density f and the skewing function π have parametric forms has been coined a skew-symmetric distribution by Wang et al. (2004). Furthermore, if $f = \phi$, the normal density, and $\pi(x;\beta) = \Phi(\beta x)$, $\beta \in \mathbb{R}$,

with Φ the normal cumulative distribution function, then (3.1) reduces to the skew-normal distribution introduced by Azzalini (1985); see the book edited by Genton (2004) for further discussions of this distribution and related families.

A practical example where a distribution of type (3.1) arises is illustrated by an ambulatory expenditure data from the 2001 Medical Expenditure Panel Survey analysed by Cameron and Trivedi (2010). The decision to spend is assumed to be related to the spending amount, hence the observations form a biased sample. Cameron and Trivedi (2010) considered a sample-selection model based on the assumption of normality, hence leading to a parametric skew-normal distribution, which corresponds to assume f to be normal in (3.1). We, instead, suggest to eliminate the normal or any other distributional assumption on the symmetric density f. Hence we only require f to be symmetric but otherwise completely unspecified. In this relaxed model setting, we estimate its center μ, which represents the mean of ambulatory expenditures for the general population had there been no expenditure decision to be made.

The rest of the chapter is organized as follows. In Section A, we adopt a semi-parametric approach to construct a class of consistent estimators of μ that are robust to model mis-specification. We further illustrate how to construct the most efficient estimator via a modified kernel density estimation. We also establish the asymptotic properties of these estimators in this section. Simulation experiments are conducted in Section B to illustrate the finite sample performance of these estimators. We implement the proposed estimators to analyse a real data example in Section C, and give a discussion in Section E. Technical details are provided in an Appendix B.

A. Estimation

1. Semiparametric Derivation

Although the central interest in (3.1) is to estimate μ, because β is also unknown, we estimate β together with μ. To this end, we treat $\theta = (\mu, \beta^T)^T$ as the parameter of interest, and treat the unknown symmetric density function f as an infinite dimensional nuisance parameter.

A rich class of root-n consistent estimators for θ in the semiparametric model (3.1) is the locally efficient semiparametric estimators. Following Bickel et al. (1993) we view the space of all the mean zero, finite variance functions as a Hilbert space $\mathcal{H}$. We begin by finding two subspaces of $\mathcal{H}$, namely the nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$; see the Appendix B for the description of Λ and $\Lambda^{\perp}$ and the locally efficient estimators. Tsiatis (2006) provides more elaborative explanations of these concepts. From here on, we use a subindex $_0$ to denote the true values of the parameters or the true functions, and write the projection of h onto a space A as $\Pi(h|A)$.

For model (3.1), we establish in the Appendix B that the nuisance tangent space Λ, corresponding to the unspecified symmetric probability density function f, and its orthogonal complement $\Lambda^{\perp}$, are respectively

$$\begin{aligned} \Lambda &= \left\{ u(x-\mu) : u(t) = u(-t), \int_0^{\infty} u(t) f_0(t) dt = 0 \right\}, \\ \Lambda^{\perp} &= \{ v(x-\mu) : v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta) = 0 \}. \end{aligned}$$

Any function in $\Lambda^{\perp}$ can be normalized to an influence function hence provides an estimation function. However, within this large class of estimation functions, the efficient score function is the most attractive because the corresponding estimator has the smallest estimation variance. The efficient score is defined as the orthogonal

projection of the score function to $\Lambda^{\perp}$. Denote $g(x;\theta) = 2f(x-\mu)\pi(x-\mu;\beta)$. Calculating $\partial \log g(x;\theta)/\partial\theta$, we obtain the score function

$$S_\theta = (S_\mu, S_\beta^T)^T = \left\{ -\frac{f_0'(x-\mu)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)}, \frac{\pi_\beta'(x-\mu;\beta)^T}{\pi(x-\mu;\beta)} \right\}^T,$$

where we use the notation $\pi_x'(x-\mu;\beta) = \partial\pi(x-\mu;\beta)/\partial x$ and $\pi_\beta'(x-\mu;\beta) = \partial\pi(x-\mu;\beta)/\partial\beta$. We decompose S_μ into

$$\begin{aligned} S_\mu &= \left[-\frac{f_0'(x-\mu)}{f_0(x-\mu)} \{\pi(x-\mu;\beta) - \pi(-x+\mu;\beta)\} - 2\pi_x'(x-\mu;\beta) \right] \\ &\quad + \left\{ -\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta) \right\}. \end{aligned}$$

We verify in the Appendix B that

$$-\frac{f_0'(x-\mu)}{f_0(x-\mu)} \{\pi(x-\mu;\beta) - \pi(-x+\mu;\beta)\} - 2\pi_x'(x-\mu;\beta) \in \Lambda$$

and

$$-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta) \in \Lambda^{\perp}.$$

In addition, since $\pi(t;\beta) + \pi(-t;\beta) = 1$, $\pi_\beta'(t;\beta) + \pi_\beta'(-t;\beta) = 0$, it indicates that $S_\beta \in \Lambda^{\perp}$. We thus obtain the efficient score vector for θ as

$$\begin{aligned} & S_{\theta,\text{eff}}(x;\theta,f_0) \\ = & \left\{ -\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta), \frac{\pi_\beta'(x-\mu;\beta)^T}{\pi(x-\mu;\beta)} \right\}^T. \end{aligned}$$

2. Robust Estimation Family

The form of the efficient score depends on the true yet unknown population density f_0. Thus, it cannot be directly used to construct an estimating equation. However, a useful compromise is to take advantage of the efficient score function form to con-

struct consistent estimators. Note that only the first component of the efficient score function relies on the unknown f_0 function. We find that for any symmetric density f^*, even if $f^* \neq f_0$, we would still have

$$\begin{aligned}
& \left\{ -\frac{f^{*'}(t)2\pi(-t;\beta)}{f^*(t)} - \frac{\pi'_t(t;\beta)}{\pi(t;\beta)} + 2\pi'_t(t;\beta) \right\} \pi(t;\beta) \\
& + \left\{ -\frac{f^{*'}(-t)2\pi(t;\beta)}{f^*(-t)} - \frac{\pi'_t(-t;\beta)}{\pi(-t;\beta)} + 2\pi'_t(-t;\beta) \right\} \pi(-t;\beta) \\
= & -\frac{f^{*'}(t)2\pi(-t;\beta)\pi(t;\beta)}{f^*(t)} - \pi'_t(t;\beta) + 2\pi'_t(t;\beta)\pi(t;\beta) \\
& + \frac{f^{*'}(t)2\pi(t;\beta)\pi(-t;\beta)}{f^*(t)} - \pi'_t(t;\beta) + 2\pi'_t(t;\beta)\pi(-t;\beta) \\
= & \ 0.
\end{aligned}$$

In the above we used $\pi(t;\beta) + \pi(-t;\beta) = 1$ and $\pi'_t(t;\beta) = \pi'_t(-t;\beta)$. Thus, according to the description of $\Lambda^{\perp}$, $S_{\theta,\text{eff}}(x;\theta,f^*)$ is still an element of $\Lambda^{\perp}$.

Based on the above observation, we propose to construct a simple consistent and robust estimator for θ as follows. We first postulate a symmetric density f^*. We then calculate the corresponding efficient score function

$$\begin{aligned}
& S_{\theta,\text{eff}}(x;\theta,f^*) \\
= & \left\{ \begin{array}{c} -\dfrac{f^{*\prime}(x-\mu)2\pi(-x+\mu;\beta)}{f^*(x-\mu)} - \dfrac{\pi'_x(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi'_x(x-\mu;\beta) \\ \dfrac{\pi'_\beta(x-\mu;\beta)}{\pi(x-\mu;\beta)} \end{array} \right\}. \qquad (3.2)
\end{aligned}$$

We form the estimating equation

$$\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f^*) = 0$$

to solve for $\widehat{\mu}$ and $\widehat{\beta}$. In practice, a normal density or a Laplace density model for f^* are the obvious choices. If the postulated model f^* happens to be the same as f_0, we indeed obtain the efficient estimator for θ from this procedure. However, even if f^*

is not the same as f_0, the construction still guarantees consistency. This means the estimator has a robustness property against the mis-specification of f_0.

In postulating a model f^* for f, the only constraint we have is $f^*(x) = f^*(-x)$. In other words, we can choose the variance of the density model arbitrarily. Intuitively, a variance choice that is close to the true variance of f may yield a more stable estimation while a drastically different variance choice could cause some loss on computational stability as well as affect the estimation variability of the final estimates for θ. Thus, a very natural alternative is to postulate a parametric model for f_0, instead of one particular density function. We denote the postulated density family $f^*(x;\gamma)$, where γ is a vector of additional parameters, such as the variance parameter of f^*. In terms of determining γ, we can simply augment the estimating function (3.2) with the score function concerning γ or plug in an estimated γ value to (3.2).

To be specific about estimating θ through augmenting or plugging-in when a more general model $f^*(x;\gamma)$ is postulated, we describe how to estimate γ. We write the model as $f^*(x;\gamma)$ whether or not it contains the truth f_0. Calculating $\partial \log g(x;\theta_0,\gamma,f^*)/\partial\gamma$ yields the nuisance score vector

$$S_\gamma(x;\theta,\gamma,f^*) = \frac{\partial f^*(x-\mu;\gamma)/\partial\gamma}{f^*(x-\mu;\gamma)}.$$

We can augment (3.2) with the above estimating function to solve for $\widehat{\gamma}$ and $\widehat{\theta}$ jointly. Alternatively, we can also iteratively use (3.2) with γ fixed at the current value and use $S_\gamma(x;\theta,\gamma,f^*)$ with θ fixed at the current value to obtain $\widehat{\gamma}$ and $\widehat{\theta}$.

In terms of the robustness and efficiency of this more general strategy, we find that if the posited model $f^*(\cdot;\gamma)$ contains the true f_0, then we obtain the efficient estimator. However, if the posited model $f^*(\cdot;\gamma)$ does not contain the true f_0, we still obtain a consistent estimator. Thus, this more general postulation strategy retains the

robust and local efficient property of the simple postulation strategy. In addition, we find that the estimation of the additional parameter γ does not affect the estimation variance of θ. To make a distinction for the two postulation strategies, we write

$$S_{\theta,\text{eff}}(x;\theta,\gamma,f^*) = \begin{Bmatrix} -\dfrac{f^{*'}(x-\mu;\gamma)2\pi(-x+\mu;\beta)}{f^*(x-\mu;\gamma)} - \dfrac{\pi'_x(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi'_x(x-\mu;\beta) \\ \dfrac{\pi'_\beta(x-\mu;\beta)}{\pi(x-\mu;\beta)} \end{Bmatrix}, \quad (3.3)$$

where $f^{*'}(t;\gamma) = \partial f^*(t;\gamma)/\partial t$. We summarize our discovery stated above in Theorem 3, after stating a useful Lemma. The proofs of both Lemma 1 and Theorem 3 are provided in the Appendix B.

Lemma 1. *Assume $n^{1/2}(\widehat{\gamma}-\gamma^*)$ is bounded in probability. Then the two estimators obtained from solving the two estimating equation $\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\theta,\gamma^*,f^*)=0$ and $\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\theta,\widehat{\gamma},f^*)=0$ are asymptotically equivalent; that is, if the estimator $\widehat{\theta}_1$ is the solution of the first equation and $\widehat{\theta}_2$ is the solution of the second equation, then $n^{1/2}(\widehat{\theta}_1-\widehat{\theta}_2)\to 0$ in probability.*

Theorem 3. *i) If the candidate family $f^*(x-\mu;\gamma)$ contains the truth f_0, i.e. there exists γ_0 such that $f^*(x-\mu;\gamma_0)=f_0(x-\mu)$, then $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V_{\text{eff}})$ in distribution when $n\to\infty$, where $V_{\text{eff}}=E\{S_{\text{eff}}(X;\theta_0,f_0)S_{\text{eff}}(X;\theta_0,f_0)^T\}^{-1}$ and $\widehat{\theta}$ solves the estimating equation $\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\theta,\widehat{\gamma},f^*)=0$. Here, $\widehat{\gamma}$ is a root-n consistent estimator for γ_0.*

ii) If the candidate family $f^*(x-\mu;\gamma)$ does not contain the truth, i.e. $f^*(x-\mu;\gamma)\neq f_0(x-\mu)$ for any γ. Then $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V)$ in distribution when $n\to\infty$, where $V=A^{-1}E\{S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)S^T_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)\}(A^{-1})^T$, and

$$A = E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)}{\partial\theta^T}\right\}.$$

Here $\widehat{\theta}$ *solves the estimating equation* $\sum_{i=1}^{n} S_{\theta,eff}(X_i; \theta, \widehat{\gamma}, f^*) = 0$*, where* $\widehat{\gamma} = \gamma^*$ *or is a root-n consistent estimator of* γ^**.*

3. Efficient Estimation

The efficiency of an estimator depends on how close the posited model f^* is to the true f_0. Although the various robust estimators proposed in Section 2 guarantee consistency, they only provide a possibility of achieving efficiency. That is, the estimation variability relies on the specific postulated model or the family of models. Only if the model happens to be true or the family happens to contain the true density f_0, the optimal estimation variance is achieved, otherwise, all one can obtain is consistency.

To overcome this potential loss of efficiency, we propose to perform a nonparametric estimation of f using a modified procedure of the kernel density estimation. The explicit form of the modified kernel estimator for f is

$$\widehat{f}(t) = \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{h} \left[K\left\{ \frac{(X_i - \mu) - t}{h} \right\} + K\left\{ \frac{(X_i - \mu) + t}{h} \right\} \right],$$

where K is a symmetric kernel function and h is a bandwidth. To see the rationale behind this estimation, we first ignore the semiparametric model (3.1). Then we can use the usual kernel density estimator at a given point x to obtain $(nh)^{-1} \sum_{i=1}^{n} K(X_i - x)$. Taking into account (3.1), we write the estimate as

$$2\widehat{f}(x-\mu)\pi(x-\mu;\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h} \right) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left\{ \frac{(X_i - \mu) - (x - \mu)}{h} \right\}.$$

Since f is an even function, we require its estimator $\widehat{f}$ to be also even. Thus we have

$$\begin{aligned} 2\widehat{f}(x-\mu)\pi(-x+\mu;\beta) &= 2\widehat{f}(-x+\mu)\pi(-x+\mu;\beta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left\{\frac{(X_i-\mu)-(-x+\mu)}{h}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i+x-2\mu}{h}\right). \end{aligned}$$

Combining the above two equalities, we obtain

$$\widehat{f}(x-\mu) = \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h}\left\{K\left(\frac{X_i-x}{h}\right)+K\left(\frac{X_i+x-2\mu}{h}\right)\right\}, \tag{3.4}$$

which yields the modified kernel density estimation for f.

The robust estimation described in Section 2 can be combined with the modified nonparametric kernel density estimation to yield a procedure that achieves the optimal semiparametric efficient bound for θ. The estimation procedure is the following:

- Step 1. Choose any even density function f^*. From f^* and the known function π, obtain $\tilde{\theta} = (\tilde{\mu},\, \tilde{\beta}^T)^T$ through solving $\sum_{i=1}^{n} S_{\text{eff}}(X_i;\theta,\, f^*) = 0$.

- Step 2. Choose a kernel function K and a bandwidth h, plug K, h and $\tilde{\mu}$ into (3.4) to obtain $\widehat{f}$.

- Step 3. Using the estimated $\widehat{f}$, obtain $\widehat{\theta} = (\widehat{\mu}, \widehat{\beta}^T)^T$ through solving $\sum_{i=1}^{n} S_{\text{eff}}(X_i;\theta,\widehat{f}) = 0$.

It is worth pointing out that in the above procedure, it is not necessary to perform any iteration. The first order optimal asymptotic property is achieved via the simple one-step procedure. However, in practice, iterating between Steps 2 and 3, while using the previous obtained $\widehat{\mu}, \widehat{\beta}$ to replace $\widetilde{\mu}, \widetilde{\beta}$ in Step 2 is often recommended, especially when the sample size is moderate or small. A bandwidth h is needed in Step 2. Interestingly, there is no need to perform any under-smoothing in this step and the

procedure is very insensitive to the bandwidth. Thus, a standard cross-validation procedure can be used on an initial kernel density estimation to obtain a bandwidth h, and one can then use this bandwidth throughout the estimation procedure. This practice is both theoretically justified and practically well behaved.

We assume the following regularity conditions are satisfied in the rest of the Theorems in this Chapter.

Regularity Conditions 1. *C1 The true distribution of $X_1, \ldots, X_n$ has a compact support. That is, $-\infty < C_1 < X_{(1)}$ and $X_{(n)} < C_2 < \infty$, where $X_{(j)}$ is the jth smallest observation and C_1, C_2 are two constants.*

C2 The symmetric function f_0 is twice differentiable. Its second derivative satisfies the Lipschitz condition. f_0 and f_0' are bounded away from zero and ∞. f_0 has bounded second derivative with respect to its center μ.

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 h is such that $C_3^{-1} n_2^{-1/8} < h < C_3 n_2^{-1/2}$ for all n_2, where $C_3 > 1$ is a constant and can be arbitrarily large.

Proposition 1. *Let $X_1, \cdots, X_n$ be independent and identically distributed with common density $2f(x-\mu)\pi(x-\mu, \beta)$. 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. Suppose that $\widetilde{\mu}$ and $\widetilde{\beta}$ are estimators constructed from the observations $X_{n_1+n_2+1}, \cdots, X_n$ and satisfies $\widetilde{\mu} - \mu = O_p(n_3^{-1/2})$ and $\widetilde{\beta} - \beta = O_p(n_3^{-1/2})$. Let*

$$\widehat{f}(t; \widetilde{\mu}) = \frac{1}{2n_2} \sum_{i=n_1+1}^{n_1+n_2} \frac{1}{h} \left\{ K\left(\frac{X_i - \widetilde{\mu} - t}{h}\right) + K\left(\frac{X_i - \widetilde{\mu} + t}{h}\right) \right\}$$

or equivalently

$$\widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) = \frac{1}{2n_2}\sum_{i=n_1+1}^{n_1+n_2}\frac{1}{h}\left\{K\left(\frac{X_i-x}{h}\right)+K\left(\frac{X_i+x-2\widetilde{\mu}}{h}\right)\right\}.$$

It then follows that $||\widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0)|| = o_p(n_2^{-1/4})$*, under the Regularity Conditions 1.*

Note that unknown function f_0 is estimated from m observations, while the initial estimates $\widetilde{\mu}$, $\widetilde{\beta}$ of unknown parameters μ_0, β_0 are computed from $n-m$ observations. We can calculate efficient estimator $\widehat{\mu}$ and $\widehat{\beta}$ using $\widehat{f}(x;\widetilde{\mu})$ by solving efficient score $S^*_{\text{eff}}(X,\mu,\beta)$.

Corollary 1. *Under the Regularity Conditions 1,*

$$\widehat{f'}(x-\widetilde{\mu};\widetilde{\mu}) = \frac{1}{2n_2}\sum_{i=n_1+1}^{n_1+n_2}\frac{1}{h^2}\left\{K'\left(\frac{x-X_i}{h}\right)+K'\left(\frac{X_i+x-2\widetilde{\mu}}{h}\right)\right\},$$

or equivalently,

$$\widehat{f'}(t;\widetilde{\mu}) = \frac{1}{2n_2}\sum_{i=n_1+1}^{n_1+n_2}\frac{1}{h^2}\left\{K'\left(\frac{t-X_i+\widetilde{\mu}}{h}\right)+K'\left(\frac{X_i+t-\widetilde{\mu}}{h}\right)\right\}.$$

then $||\widehat{f'}(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0)|| = o_p(n^{-1/4})$.

Corollary 2. *Under the Regularity Conditions 1, if we approximate* f_0' *using numerical differentiation* $\widehat{f'}(t;\widetilde{\mu}) \equiv \{\widehat{f}(t+n^{-1/4};\widetilde{\mu}) - \widehat{f}(t-n^{-1/4};\widetilde{\mu})\}/(2n^{-1/4})$*, then* $||\widehat{f'}(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0)|| = o_p(1)$.

In Theorem 4, we state the optimal property of the modified nonparametric kernel estimation. We use the notation $a^{\otimes 2}$ to denote aa^T.

Theorem 4. *Let* $X_1,\ldots,X_n$ *be independent and identically distributed with common*

density $2f_0(x-\mu_0)\pi(x-\mu_0;\beta_0)$. *For any* t, *let*

$$\begin{aligned}\widehat{f}(t;\widetilde{\mu}) &= \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h}\left\{K\left(\frac{X_i-\widetilde{\mu}-t}{h}\right)+K\left(\frac{X_i+t-\widetilde{\mu}}{h}\right)\right\},\\ \widehat{f}'(t;\widetilde{\mu}) &= \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h^2}\left\{K'\left(\frac{t-X_i+\widetilde{\mu}}{h}\right)+K'\left(\frac{X_i+t-\widetilde{\mu}}{h}\right)\right\},\end{aligned}$$

where $\widetilde{\mu}$ *is estimated from Step 1. Assume* $\widehat{\theta}=(\widehat{\mu},\widehat{\beta}^T)^T$ *satisfies*

$$\sum_{i=1}^{n} S_{\theta,\mathit{eff}}\{X_i;\widehat{\theta},\widehat{f}(\cdot;\widetilde{\mu})\}=0.$$

It then follows that when $n\to\infty$, *under the Regularity Conditions 1,* $\widehat{\theta}$ *is the semi-parametric efficient estimator and it satisfies* $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V_{\mathit{eff}})$ *in distribution when* $n\to\infty$. *Here* $V_{\mathit{eff}}=[E\{S^{\otimes 2}_{\theta,\mathit{eff}}(X;\theta_0,f_0)\}]^{-1}$.

4. Population Density Estimation

A by-product of the efficient estimation described in Section B is the nonparametric estimation of the population density f itself. Because the only apriori information we have about f is its symmetry, hence it is not a surprise that f has the typical nonparametric bias and variance properties. We point out here that the fact we do not know the center μ and had to estimate it does not affect the first order asymptotic property of estimating f. In other words, the first order asymptotic convergence properties of $\widehat{f}$ remain the same whether we know μ or not. We summarize the theoretical results in Theorem 5 and provide the necessary proofs in the Appendix B under the Regularity Condition 1.

Theorem 5. *Let* $c_2=\int_{-1}^{1}s^2K(s)ds$, $v_2=\int_{-1}^{1}K^2(s)ds$. *Under the Regularity Conditions 1, the nonparametric estimation* $\widehat{f}$ *obtained from Step 2 satisfies the usual*

nonparametric bias and variance property

$$
\begin{aligned}
bias\{\widehat{f}(t;\widetilde{\mu})\} &= \frac{h^2}{2} f_0''(t) c_2 + o(h^2), \\
var\{\widehat{f}(t;\widetilde{\mu})\} &= \frac{1}{nh}\left\{ \frac{f_0(t) v_2}{2} + I(|t|<h) \int_{-1+\frac{|t|}{h}}^{1-\frac{|t|}{h}} K(s-t/h) K(s+t/h) w(hs) ds \right\} \\
&\quad + o\{(nh)^{-1}\}.
\end{aligned}
$$

Because the bias and variance properties have a similar form as in the usual nonparametric estimation, the subsequent MSE and MISE results also remain in the standard form. Once a nonparametric estimation of f is obtained, it is straightforward to assemble $\widehat{f}$ and $\widehat{\theta}$ together to reconstruct an estimation of the density of the biased samples. This can provide a visual verification of the estimation in practice, see Section C for an illustration.

Estimating the population density function curve $\widehat{f}$ and the density $\widehat{g}$ of the biased samples is a nonparametric density estimation problem. Hence the bandwidth selection is important for the final performance. Here, the usual bandwidth selection procedure such as cross-validation and plugin methods are applicable. However, we recommend a more refined indirect cross-validation procedure, which allows us to use two different kernels: one is suitable for cross-validation purpose while the other is suitable for estimation purpose. The rationale of the indirect cross-validation method and suitable kernels are studied in Savchuk et al. (2010).

B. Simulations

We performed a set of simulation studies to investigate the finite sample performance of the various estimators we proposed. In our first simulation study, the data sets were generated from model (3.1) with true $f_0(x)$ being a normal density with mean 4 and standard deviation 3. The corresponding skewing function π was a logit function

with skewing parameter $\beta = 2$, i.e. $\pi(x-\mu;\beta) = 1 - \{1 + e^{\beta(x-\mu)}\}^{-1}$. In the second simulation study, the true $f_0(x)$ was a Laplace density with the same location and scale parameters. We also experimented with the more common situation where the skewing function π was a probit function, i.e. $\pi(x-\mu;\beta) = \Phi\{\beta(x-\mu)\}$, where Φ is the standard normal cumulative distribution function. This skewing function model in combination with normal or Laplace $f_0(x)$ is respectively studied in simulations 3 and 4. In each simulation, a sample size $n = 500$ was used and 1000 data sets were generated.

We implemented five different estimators to illustrate the relative performance. The first estimator can be considered as an oracle estimator, where we solve $\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f_0) = 0$ to obtain $\widehat{\theta}$. Here we plug in the true density f_0 to the estimating equation, as if we would know the true f_0, hence the name oracle. The second estimator is very similar to the first, except that we plug in a wrong density. The estimating equation is explicitly $\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f^*) = 0$, where $f^* \neq f_0$. In particular, in the first simulation study, where the true f_0 is normal, we plugged in a Laplace f^*, while in the second simulation study, where the true f_0 is Laplace, we plugged in a normal f^*. Our third and fourth estimators involve an additional parameter γ, which is the scale parameter of f^* in both simulations. To be precise, in the third estimator, we augment $S_{\theta,\text{eff}}(X_i;\theta,f^*)$ with S_γ, where f^* is a correct model (i.e. normal in simulations 1, 3 and Laplace in simulations 2, 4) with location parameter μ and scale parameter γ. While in the fourth estimator, we also augment $S_{\theta,\text{eff}}(X_i;\theta,f^*)$ with S_γ. However, f^* is now a mis-specified model, where we used a Laplace model in simulations 1, 3 and a normal model in simulations 2, 4, with location parameter μ and scale parameter γ. Finally, the efficient estimator described in Section B is implemented for all simulation studies as the fifth estimator.

The simulation results of the five estimators are summarized in Table V. It can

Table V. Median of 1000 estimates of μ and β and their sample standard deviation (sd) and average of the estimated standard deviation ($\widehat{\text{sd}}$) for simulations 1-4. True values are $\mu_0 = 4$, $\beta_0 = 2$. Results based on sample size $n = 500$.

	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\beta}$	sd	$\widehat{\text{sd}}$	95% cvg
	simulation 1 (f=normal, π=Logit)							
est1	3.8326	0.6632	0.6376	94.9%	2.2957	1.5414	1.3880	93.3%
est2	3.9588	0.6593	0.7015	96.3%	2.0575	0.9004	0.9625	97.2%
est3	3.8352	0.5708	0.5655	93.2%	2.2872	1.4147	1.3769	94.9%
est4	3.9731	0.9169	0.8313	95.2%	2.0289	1.1832	1.1863	96.9%
est5	4.0000	0.3062	0.2662	98.8%	2.0180	0.4264	0.4390	98.4%
	simulation 2 (f=Laplace, π=Logit)							
est1	3.9955	0.4334	0.5204	97.9%	2.0143	0.7296	0.8058	98.6%
est2	3.8913	0.5238	0.4986	94.4%	2.2428	1.3228	1.2594	95.4%
est3	3.8298	1.1201	1.3139	94.7%	2.3498	1.3943	1.6159	96.3%
est4	3.8585	0.5745	0.5258	91.5%	2.3085	1.4766	1.4221	93.7%
est5	3.9989	0.2314	0.2693	98.4%	2.0156	0.4951	0.4733	98.1%
	simulation 3 (f=Normal, π=Probit)							
est1	3.9686	0.3762	0.4022	97.7%	2.0898	0.9771	1.1044	97.7%
est2	3.9834	0.5253	0.5232	98.1%	1.9985	1.2394	1.0299	93.3%
est3	3.9666	0.3119	0.3463	97.9%	2.0824	0.9636	1.0971	98.0%
est4	3.9902	0.4909	0.5342	98.2%	1.9682	0.9763	0.9365	95.7%
est5	4.0005	0.2641	0.2812	99.0%	2.0001	0.5920	0.7828	99.0%
	simulation 4 (f=Laplace, π=Probit)							
est1	3.9930	0.3671	0.3816	98.6%	2.0035	0.9482	0.9228	96.1%
est2	3.9495	0.2750	0.2814	96.5%	2.1894	1.1037	1.1254	97.7%
est3	4.0122	0.9234	1.0231	95.3%	1.8684	1.4034	1.4804	95.8%
est4	3.9347	0.2517	0.3129	97.5%	2.2416	1.1966	1.2427	96.8%
est5	4.0006	0.3245	0.2774	98.1%	2.0197	0.7223	0.7920	98.2%

be seen that all the five estimators in all simulations exhibit very small bias, indicating the evidence that regardless whether the f^* function or f^* model is correctly specified or incorrectly specified, regardless whether f^* is fully specified or partially specified or completely decided through data, the estimators remain consistent. In addition, the average of the estimated variability is very close to the sample version, hence the inference is reasonably reliable. Here, we point out that the reported estimated standard deviation is obtained via a bootstrap procedure, because in our experiment,

we find that the asymptotic properties require much larger sample size than $n = 500$. This difficulty is also reflected in the estimation in the extreme tail region, as we can see that the 95% coverage is not very close to the nominal level. Finally, concerning the estimation efficiency, although we expect estimators 1, 3 and 5 to be asymptotically equivalent to their first order approximation, and estimators 2 and 4 to be less efficient than 1, 3, 5, this clearly is not always the case. Based on the fact that the asymptotic variability estimation is not sufficiently accurate for $n = 500$, it is not difficult to see that this is also caused by the moderate sample size, which masks out the first order performance. The encouraging news is that the efficient estimator (estimator 5) does not perform less favorably in comparison to alternative estimators for this moderate sample size. In fact, it some times performs competitively with respect to the oracle estimator. Because it is a painless procedure, we highly recommend implementing it. On the other hand, if a quick assessment of the parameters are needed, then one should feel comfortable to postulate a model and perform a Step 1 simple estimation. The simulation evidence strongly supports the consistency of such procedure.

To further examine the performance of the additional nonparametric estimation procedure, we also plotted the estimated density curves of both the underlying true population and the population reflected by the biased selected sample in Figures 1-4. All the results are based on the Quartic kernel function and the bandwidth is selected via the indirect cross-validation procedure. As we can see, the estimation is satisfying when the true population is normal. However, when the true population is extremely heavy-tailed as in the Laplace case, the performance deteriorated. This is not a surprise since in this case, even with a non-biased sample from the population, the nonparametric estimation is a difficult problem.

C. Data Example

We now analyse the ambulatory expenditures data mentioned in the introduction. The data consists of $n = 2802$ observations and because the distribution of expenditures is highly skewed, the logarithmic scale was used. Following Cameron and Trivedi (2010), we fit model (3.1) with a normal skewing function $\pi(x-\mu;\beta) = \Phi\{\beta(x-\mu)\}$. We computed the five estimators of the center μ described in the previous section. Specifically, the estimator 1 posited a normal model for f with a fixed standard deviation of 1.4107, which is the sample standard deviation. The estimator 2 posited a Laplace model for f with again a fixed standard deviation of 1.4107. The estimator 3 posited normal model for f with an unknown standard deviation, whereas the estimator 4 posited Laplace model for f with an unknown standard deviation. Finally, the estimator 5 estimated f nonparametrically.

The results for our five estimators are listed in Table VI, as well as their estimated variance and standard deviations. Our estimate of the center is $\widehat{\mu} = 5.93$, whereas other more stringent assumptions on model (3.1) lead to different estimates. In contrast, the sample mean, an estimator of μ that does not correct for the sample selection bias, is 6.56, significantly different from 5.93 at the 95% level according to Table VI.

The estimated densities of the population distribution $\widehat{f}$ and the selected sample distribution $\widehat{g}$ are plotted in Figure 5. 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 non-normal shape, hence confirming that it is wise to leave f completely unspecified.

Table VI. Five estimates of μ and β and their estimated variance and standard deviation for the ambulatory expenditures data.

	$\widehat{\mu}$	$\widehat{\beta}$	$\widehat{\mathrm{var}}(\widehat{\mu})$	$\widehat{\mathrm{var}}(\widehat{\beta})$	$\widehat{\mathrm{sd}}(\widehat{\mu})$	$\widehat{\mathrm{sd}}(\widehat{\beta})$
est1	6.5507	0.0028	0.0008	0.0000	0.0275	0.0010
est2	6.0124	0.3450	0.0049	0.0020	0.0699	0.0445
est3	6.5507	0.0027	0.0008	0.0000	0.0279	0.0040
est4	5.8606	0.4296	0.0025	0.0012	0.0504	0.0342
est5	5.9299	0.3919	0.0116	0.0040	0.1076	0.0635

D. Discussion

We have focused on a rather special selection process, which naturally yields a selection function π that satisfies $\pi(x) + \pi(-x) = 1$. This property has enabled us to derive consistent estimators that are robust to mis-specification of the symmetric part of the model f. Without this property, a consistent estimation of the population center generally requires estimating the population density f itself, and we will no longer be able to construct a robust estimator. In other words, if we still postulate a wrong density or wrong family of models for the density, then the subsequent estimation for the population center may no longer be consistent.

However, as long as we are willing to perform nonparametric estimation procedures, possibly taking into account the additional symmetry property of the population distribution f and any characteristics of the selection procedure reflected in π, consistent and even efficient estimates for the population center may be still possible. How to best treat various selection mechanisms is something worth further investigation.

We have treated the case of model (3.1) where the density f is completely unspecified and the skewing function π is assumed to have a known parametric form due to a specific selection procedure. An alternative setting is when the density f

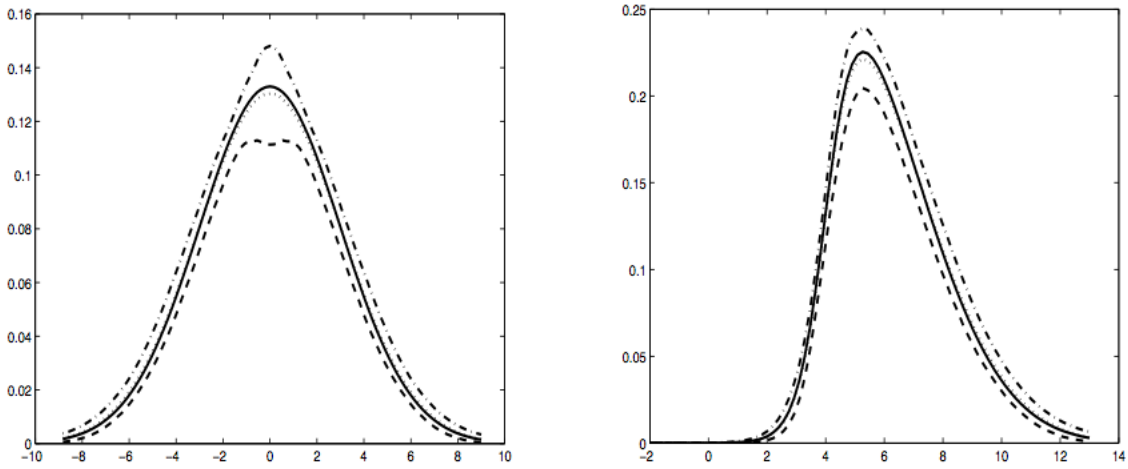

Fig. 1. Pointwise quantile curves from simulation 1. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively.

has a known parametric form, whereas the selection mechanism is somewhat hidden, hence the skewing function π is unknown. These models have been investigated by Ma et al. (2005) and Ma and Hart (2007).

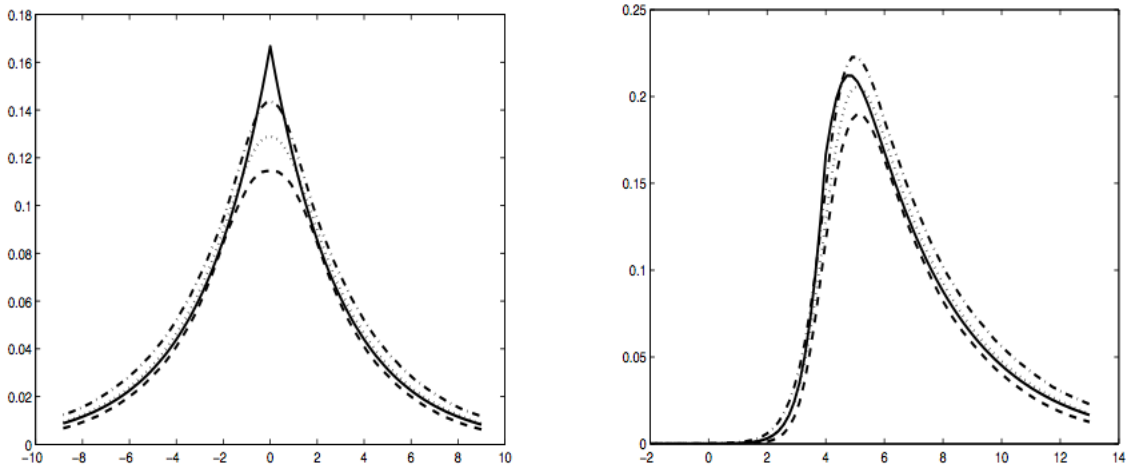

Fig. 2. Pointwise quantile curves from simulation 2. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 2. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively.

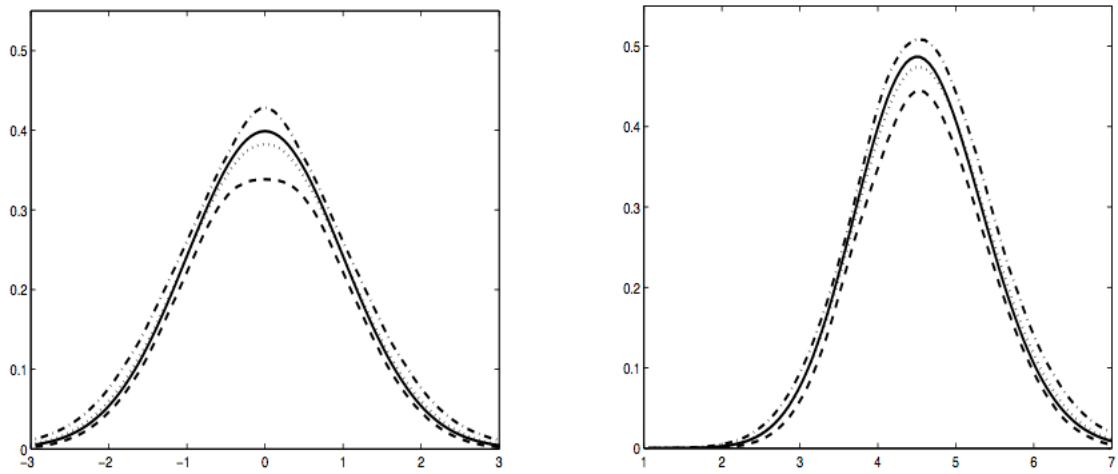

Fig. 3. Pointwise quantile curves from simulation 3. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively.

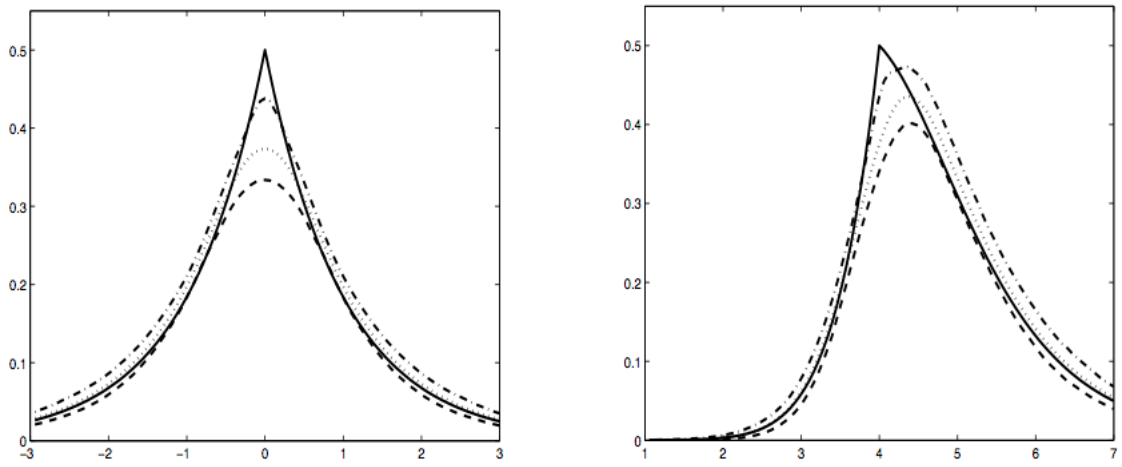

Fig. 4. Pointwise quantile curves from simulation 4. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 4. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected subsample density ($g(x)$), respectively.

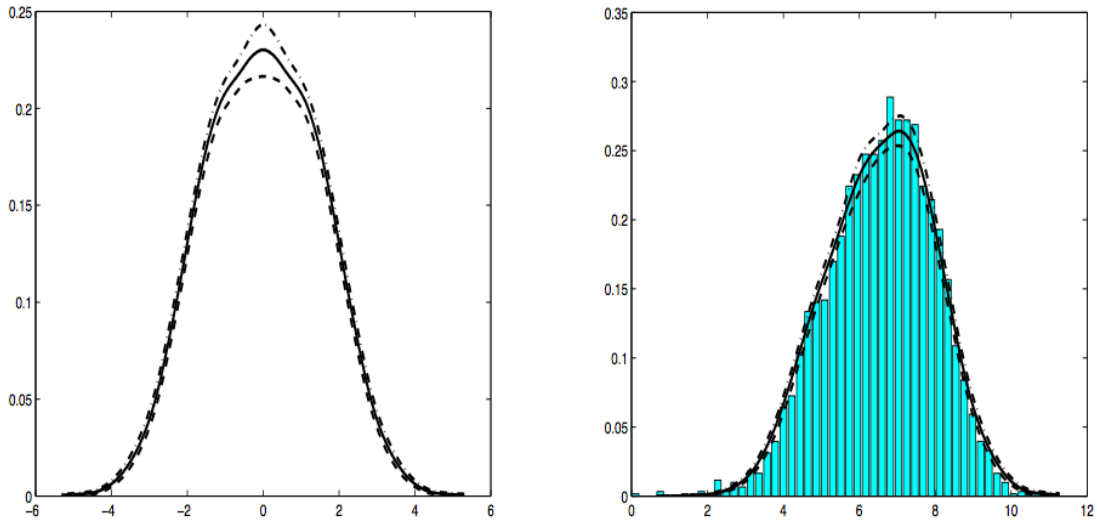

Fig. 5. The estimated densities of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ for the ambulatory expenditures data. The estimated sample density curve is overlayed on the histogram of the observations.

CHAPTER IV

EFFICIENT AND ROBUST ESTIMATION USING BIASED SAMPLES

In Chapter III, we have studied estimation of a population center when the population is symmetric but the sample does not reflect the population due to the selection bias. We adopted a skewing function which captured the selection mechanism and imposed a rather strict assumption. That is, the skewing function π satisfies an anti-symmetric property, i.e., $\pi(t;\beta) + \pi(-t,\beta) = 1$. In this Chapter, we relax the restriction on a skewing function and consider a more general selection model.

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 in general write the probability density function (pdf) of one observation as

$$g(x, \mu, \boldsymbol{\beta}, f) = c(\boldsymbol{\beta}) f(x-\mu) w(x-\mu, \boldsymbol{\beta}) = \frac{f(x-\mu) w(x-\mu, \boldsymbol{\beta})}{\int f(t) w(t, \boldsymbol{\beta}) dt}, \tag{4.1}$$

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 (4.1), other than being even, the specific form of f is not known. 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 (4.1). 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 the weight function.

A familiar example of samples subject to selection bias is given by Cameron and Trivedi (2010), where they consider a data set of ambulatory expenditures from the 2001 Medical Expenditure Panel Survey. 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 (4.1) 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 (4.1). 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, $b_Y(Y) \leq b_X(X-\mu)$, where b_X, b_Y denote the corresponding benefit functions associated with the two expenditures. This can be described by $w(X-\mu) = \mathrm{pr}\{Y \leq a+b(X-\mu)\} = H\{a+b(X-\mu)\}$ 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 $w(X-\mu) = \mathrm{pr}\{b_Y(Y) \leq b_X(X-\mu)\}$. 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 (4.1). Estimating and studying the corresponding inference on μ using the biased sample is the main purpose of this chapter.

We organize the rest of the chapter as follows. In Section A, 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 in Section B. The asymptotic properties of both the consistent and efficient estimators are derived in Section C. We conduct numerical experiments via simulations and the ambulatory expenditure data analysis in Section D. We finish the chapter with a discussion in Section E. Technical details are collected in an Appendix C.

A. Consistent Estimation Under Misspecified f

1. The Estimator Family

Model (4.1) 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 (4.1) 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 by-product.

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 (4.1). Thus, semiparametric methods described in Bickel et al. (1993) and Tsiatis (2006) become applicable. The central result of the semiparameric 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 its matching member in $\Lambda^{\perp}$. For model (4.1), we explicitly derived in the Appendix C 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 C and throughout the rest of the chapter, 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 which 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, $\Lambda^{\perp} = \{v(X-\mu) : v(z) + v(-z) = 0\ a.s.\}$. 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 tiltered version of the anti-symmetric requirement of $v(z) + v(-z) = 0$.

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, 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 have the potential of reaching asymptotic efficiency, yet are robust against possible model misspecifications 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 $\mathbf{S}_\theta$, is defined as $\partial \log g(x, \boldsymbol{\theta}, f)/\partial \boldsymbol{\theta}$, which has the explicit form

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

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

$$\mathbf{S}_{\text{eff}}(x, \boldsymbol{\theta}, f_0) = \left\{ \begin{array}{c} \frac{-2f_0'(x-\mu)w(-x+\mu,\boldsymbol{\beta})}{f_0(x-\mu)\{w(x-\mu,\boldsymbol{\beta})+w(-x+\mu,\boldsymbol{\beta})\}} + \frac{w'(x-\mu,\boldsymbol{\beta})+w'(-x+\mu,\boldsymbol{\beta})}{w(x-\mu,\boldsymbol{\beta})+w(-x+\mu,\boldsymbol{\beta})} - \frac{w'(x-\mu,\boldsymbol{\beta})}{w(x-\mu,\boldsymbol{\beta})} \\ -\frac{\mathbf{w}_\beta(x-\mu,\boldsymbol{\beta})+\mathbf{w}_\beta(-x+\mu,\boldsymbol{\beta})}{w(x-\mu,\boldsymbol{\beta})+w(-x+\mu,\boldsymbol{\beta})} + \frac{\mathbf{w}_\beta(x-\mu,\boldsymbol{\beta})}{w(x-\mu,\boldsymbol{\beta})} \end{array} \right\}.$$

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

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(X_i, \boldsymbol{\theta}, f^*) = \mathbf{0}. \tag{4.2}$$

Obviously, if we postulate a correct model, i.e. 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, i.e. if $f^*(t) \neq f_0(t)$, we find that the last $p-1$ components of the difference $\mathbf{S}_{\text{eff}}(X, \boldsymbol{\theta}, f^*) - \mathbf{S}_{\text{eff}}(X, \boldsymbol{\theta}, f_0)$ is zero, and the first component satisfies

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

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 $E\{\mathbf{S}_{\text{eff}}(X, \boldsymbol{\theta}, f^*)\} = \mathbf{0}$ regardless of the choice of f^*. In other words, the estimator obtained from (4.2) has an additional robustness property, in that even if the model for f is misspecified, the resulting estimator is still consistent.

B. Efficiency Considerations

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 un-specified 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

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

We can then augment (4.2) with $\sum_{i=1}^n \mathbf{S}_\gamma(X_i, \boldsymbol{\theta}, \boldsymbol{\gamma}, f^*) = \mathbf{0}$ 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 $f_0(x-\mu)$ in order 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 $\widehat{\boldsymbol{\theta}}_{\gamma_0}$, 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 $\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}}$, the estimation variabilities of $\widehat{\boldsymbol{\theta}}_{\gamma_0}$ and $\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}}$ are the same asymptotically. This property will be studied more carefully in Section C.

2. Efficient Estimation of μ

When we postulate a family $f^*(x-\mu,\boldsymbol{\gamma})$ instead of one single $f^*(x-\mu)$, 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 estimatior which takes advantage of the symmetry of $f(t)$.

To derive the nonparametric kernel density estimator for $f(t)$, we begin by using the usual kernel density estimation at a given point x from the density g. That is, $\widehat{g}(x,\boldsymbol{\theta},f)=\dfrac{1}{n}\sum_{i=1}^{n}K_h(X_i-x)$, where $K_h(t)=K(t/h)/h$, K is a kernel function and h is a bandwidth. We propose $\widetilde{f}(t,\boldsymbol{\theta})$ for the kernel density estimator of $c(\boldsymbol{\beta})f(t)$. The explicit form of the refined kernel estimator we propose is

$$\widetilde{f}(t,\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{K_h\left(X_i-\mu-t\right)+K_h\left(X_i-\mu+t\right)}{w(t,\boldsymbol{\beta})+w(-t,\boldsymbol{\beta})}, \tag{4.3}$$

where $K_h(t)=K(t/h)/h$, K is a kernel function and h is a bandwidth. The form of the estimator in (4.3) 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. However, 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 (4.2);
- Step 2: Obtain $\widetilde{f}(t, \widetilde{\boldsymbol{\theta}})$ from (4.3);
- Step 3: Obtain $\widehat{\boldsymbol{\theta}}$ through solving (4.2) 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 (4.2) in Step 3, it is completely equivalent to the estimating equation solving procedure in Step 1. Thus, the above 3-step 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 property of $\widehat{\boldsymbol{\theta}}$ is 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 (4.3) 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 (4.3) 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 $\widehat{c}_f^{-1} = \int \widetilde{f}(t, \widehat{\boldsymbol{\theta}})dt$, and form $\widehat{f}(t) = \widehat{c}_f \widetilde{f}(t, \widehat{\boldsymbol{\theta}})$.

C. 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 C.

Theorem 6. *Assume $f^*(t)$ is a symmetric density function and $E\{\mathbf{S}_{\mathit{eff}}(X, \boldsymbol{\theta}, f^*)\} = \mathbf{0}$ has a unique root. Let*

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

be bounded non-singular matrices. Then the estimator $\widetilde{\boldsymbol{\theta}}$, obtained by solving (4.2) 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 6 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 7.

Theorem 7. *Assume* $f^*(t, \boldsymbol{\gamma})$ *is a family of symmetric density functions and*

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

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

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

be bounded non-singular matrices. Then the estimator $\widetilde{\boldsymbol{\theta}}$*, obtained through solving*

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

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 6 and in Theorem 7, 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 which 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 phi-

losophy behind the refined nonparametric estimator proposed in Section 2. We summarize the asymptotic properties of the estimator in Theorem 8 under the following regularity conditions. For notational brevity, we write $w_1(t, \boldsymbol{\beta}) = w(t, \boldsymbol{\beta}) + w(-t, \boldsymbol{\beta})$, $w_1'(t, \boldsymbol{\beta}) = \partial w_1(t, \boldsymbol{\beta})/\partial t$, $w_1''(t, \boldsymbol{\beta}) = \partial^2 w_1(t, \boldsymbol{\beta})/\partial t^2$.

Regularity Conditions 2. *C1 The symmetric function f_0 is twice differentiable with a compact support. f_0 and f_0' are bounded away from zero 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 $0 < w(t, \boldsymbol{\beta}_0) \leq 1$ 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 $w(t, \boldsymbol{\beta}_0) \leq 1$.

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 $n^{-1/5}$.

Theorem 8. *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 (4.2). Under the Regularity Conditions 2, the nonparametric estimator $\widetilde{f}(t, \widetilde{\boldsymbol{\theta}})$ given in (4.3) satisfies*

$$
\begin{aligned}
& bias\{\widetilde{f}(t,\widetilde{\boldsymbol{\theta}})\} \equiv E\{\widetilde{f}(t,\widetilde{\boldsymbol{\theta}})\} - c(\boldsymbol{\beta}_0) f_0(t) \\
= \;& \frac{h^2 c(\boldsymbol{\beta}_0) c_2}{2} \left\{ f_0{}''(t) + \frac{2 f_0'(t) w_1'(t,\boldsymbol{\beta}_0)}{w_1(t,\boldsymbol{\beta}_0)} + \frac{f_0(t) w_1''(t,\boldsymbol{\beta}_0)}{w_1(t,\boldsymbol{\beta}_0)} \right\} + o(h^2) \\
& var\{\widetilde{f}(t,\widetilde{\boldsymbol{\theta}})\} \\
= \;& \frac{c(\boldsymbol{\beta}_0)}{n_1 h w_1(t,\boldsymbol{\beta}_0)} \left\{ v_2 f_0(t) + \frac{2I(|t|<h)}{w_1(t,\boldsymbol{\beta}_0)} \int_0^{1-\frac{|t|}{h}} K(s-t/h) K(s+t/h) f_0(hs) w_1(hs,\boldsymbol{\beta}_0) ds \right\} \\
& + o\{(nh)^{-1}\} \\
\leq \;& \frac{2 c(\boldsymbol{\beta}_0) v_2 f_0(t)}{n_1 h w_1(t,\boldsymbol{\beta}_0)} + o\{(n_1 h)^{-1}\}.
\end{aligned}
$$

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 8 is a bounded quantity under the Regularity Conditions 2, hence the nonparametric estimator $\widetilde{f}(t,\widetilde{\boldsymbol{\theta}})$ has the classical bias and variance properties. Because the only apriori 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 8 prepares the results in Theorem 9.

Theorem 9. *Let $X_1, \ldots, X_n$ be iid with density (4.1) and let $\widetilde{\boldsymbol{\theta}}$ be an initial estimator obtained from solving (4.2). Let $\widetilde{f}(t,\boldsymbol{\theta})$ be given by (4.3) for any t and any $\boldsymbol{\theta}$. Assume $E\{\mathbf{S}_{\mathit{eff}}(X,\boldsymbol{\theta},f_0)\} = \mathbf{0}$ has a unique root and $\widehat{\boldsymbol{\theta}}$ satisfies*

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

It then follows that when $n \to \infty$, under the Regularity Conditions 2, $\widehat{\boldsymbol{\theta}}$ is the semi-

parametric efficient estimator and it satisfies

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

in distribution when $n \to \infty$.

In terms of estimating $\boldsymbol{\theta}$, Theorem 9 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 (4.1).

D. Numerical Performance

1. Simulations

We illustrate the finite sample performance of the estimators proposed in Sections A and B through a series of extensive simulation studies.

In the first simulation, we generated 1000 data sets, each with sample size $n = 500$ from model (4.1), where the true $f_0(x)$ function is a normal density with mean $\mu = 4$ and standard deviation $\sigma = 1.5$, and the selection weight function is $w(x, \beta) = e^{-\beta x}/(1 + e^{-\beta x})^2$, with the true β value equal to 1. We implemented five different estimators on each of the data sets to illustrate their respective performances. 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 (4.2), we adopt $f^* = f_0$, and solve it to obtain $\widehat{\boldsymbol{\theta}}$. 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 Laplace density function with standard deviation 1.5 as f^* and plugged it into the estimating equation (4.2) 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 estimator three, we used a normal model for f^*, and in estimator four, we used a Laplace 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 A. Finally, we also implemented the fully nonparametric estimator described in Section B, which reaches the optimal semiparametric efficiency as our fifth estimator. The results of the first set of simulations are given in the first block of Table VII. It is quite noticeable that although the proposed model f^* is completely wrong in estimators 2 and 4, the corresponding estimation biases are not much larger than their comparable ones where the proposed models are correct. This is especially clear in comparing estimators 1 and 2, where estimator 2 even shows smaller finite sample biases than estimator 1. It is also obvious that when additional nuisance parameters are included in estimators 3 and 4, the resulting estimation variability did not increase in comparison with their corresponding estimators 1 and 2. Finally, the nonparametrically estimated $\widehat{f}$ in estimator 5 also yields a comparable finite sample bias for both μ and β in comparison with the other estimators. Despite the asymptotic optimality of estimator 5, with sample size $n = 500$, the efficiency is not manifested. We also point out that the reported estimated standard deviation is obtained via a bootstrap procedure because the asymptotic properties require much larger sample size than $n = 500$.

To further study the properties, we conducted a second simulation study, where we changed the true f_0 function in simulation 1 from normal to a Laplace model, where we kept the same mean $\mu_0 = 4$ and the same scale $\sigma = 1.5$. The weighting function w is also kept unchanged. We implemented the same five estimators, with

the corresponding results in the second block of Table VII. The order of the five estimators is kept the same, in the sense that the first estimator uses the correct f_0 function, the second uses an incorrect model, which is normal in this case. The third and fourth estimators are again the correct and incorrect f^* function which contains additional nuisance parameters respectively. The fifth estimator is the nonparametric based estimator. Similar claims can be made regarding the finite sample bias, while in this simulation study, we observe the estimators 1 and 3 having smaller finite sample variability than 2 and 4, which indicate the advantage of proposing a correct f^* function or correct f^* model.

To further inspect the performance of the nonparametric based estimator, we also experimented a third set of simulations, where the true f_0 function is kept the same as in simulation 1, while we used $e^{-e^{-\beta x}}$, the cdf of Gumbel distribution as a weight function. The results are in the third block of Table VII. While the performance of the first four estimators are similar to simulation 1, we now observe a significant gain of the estimation efficiency for estimator 5. The same phenomenon is observed when we change the true f_0 function to a Laplace model as in simulation 2 (results in the fourth block of Table VII). Simulations 3 and 4 indicate the efficiency of the fifth estimator which contains nonparametric estimation of the unknown f_0 function. Together with simulations 1 and 2, the simulations suggest that the asymptotic results stated in Section C could require various sample sizes in order to be evident. The sample size requirement depends largely on the weighting function, but it also varies with different f_0 functions and different parameter values. Fortunately, even though the semiparametric efficiency may require large sample size, the robustness of these estimators against a misspecified f^* function seems relevant in moderate sample sizes in all the situations we experimented.

To provide a visual inspection of the various simulation settings and the results of

the function estimation, we provided the plots of both f (left panel) and g (right panel) in Figures 6-9. We would like to point out that in all these cases, the f and g curves are visually rather different, hence the selection bias should not be ignored. Obviously, when the true f_0 function is normal, the nonparametric estimation performs much better than when f_0 is Laplace. This is caused by the nonsmoothness of the Laplace which makes the nonparametric density estimation a very difficult problem even when no selection bias is involved.

2. Ambulatory Expenditures Data

The ambulatory expenditures data mentioned in the introduction consists of $n = 2802$ observations. To take into the possible selection bias, we fit model (4.1) with the two weighting functions in Section 1. The two weighting functions here are chosen with the intention of capturing the possible behavior patterns when the decision of using an ambulance is made. Considering that people are less willing to spend when the medical cost is high, we implemented the weighting function $w(x, \beta) = e^{-e^{-\beta x}}$, which is a monotone function of x. Further taking into account that if the associated medical cost is very low, it could indicate a minor medical situation and people could also be inclined to not using ambulance service when they do not think the situation is sufficiently grave, we also used the weighting function $w(x, \beta) = \frac{e^{-\beta x}}{(1+e^{-\beta x})^2}$, which consists of two monotone pieces.

The analysis is performed on the logarithm of the data and using five estimators of the center μ, respectively with a posited normal model for f with a fixed standard deviation 1.4107, a posited Laplace model for f with standard deviation 1.4107, a posited normal model for f with an unknown standard deviation, a posited Laplace model for f with an unknown standard deviation, and a nonparametrically estimated f.

The results for the five estimators as well as the estimated standard deviations, in conjunction with the two different weighting functions, are listed in Table VIII. Out of the 10 estimates, 8 of them resulted in the population center to be larger than the sample average, and all of them yielded a negative estimates of β. The negative value of β in the first weighting function indicates a monotonically decreasing weighting function as the expenditure increases, indicating the increasing unwillingness of using the ambulance with the increase of the associated cost. The negative value of β in the second weighting function indicates that medical events that will incur very high or very low expenses are under represented in the data set. This is an indication that patients or their family are less likely to call an ambulance when the situations are either minor or tend to incur very large costs.

The estimated densities of the population distribution $\widehat{f}$ and the selected sample distribution $\widehat{g}$ are plotted in Figure 10. 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 non-normal shape, hence confirming that it is wise to leave f completely unspecified.

E. 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

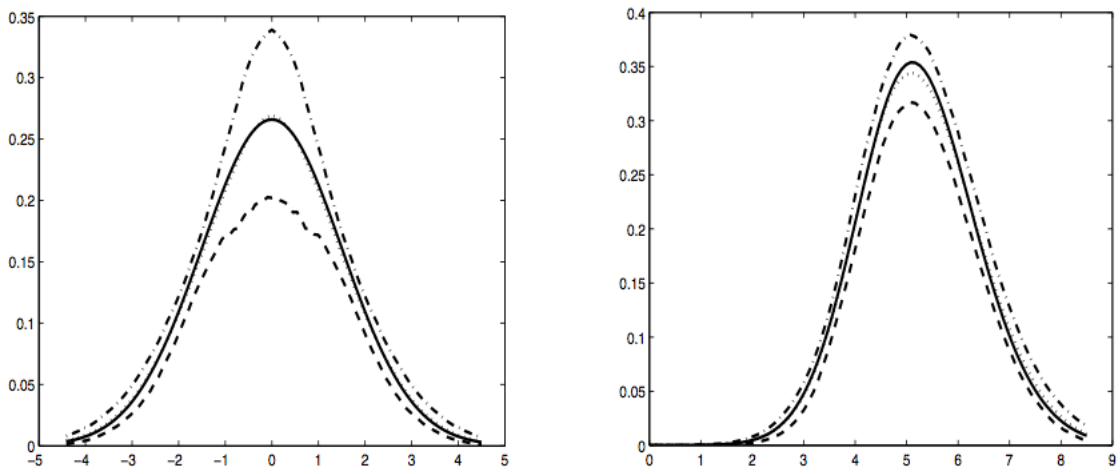

Fig. 6. Pointwise quantile curves from simulation 1. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 1. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively.

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 (4.1) where the pdf f is completely unspecified and the selection weight 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 weight function w is unknown. Such models, with the additional anti-symmetric assumption on w, have been investigated by Ma et al. (2005), Ma and Hart (2007), and Azzalini et al. (2010).

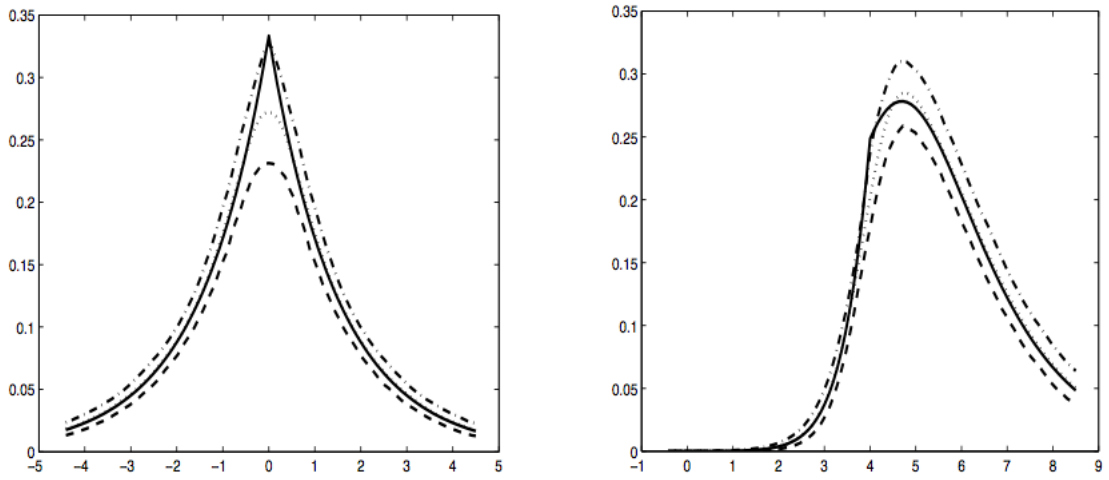

Fig. 7. Pointwise quantile curves from simulation 2. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 2. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively.

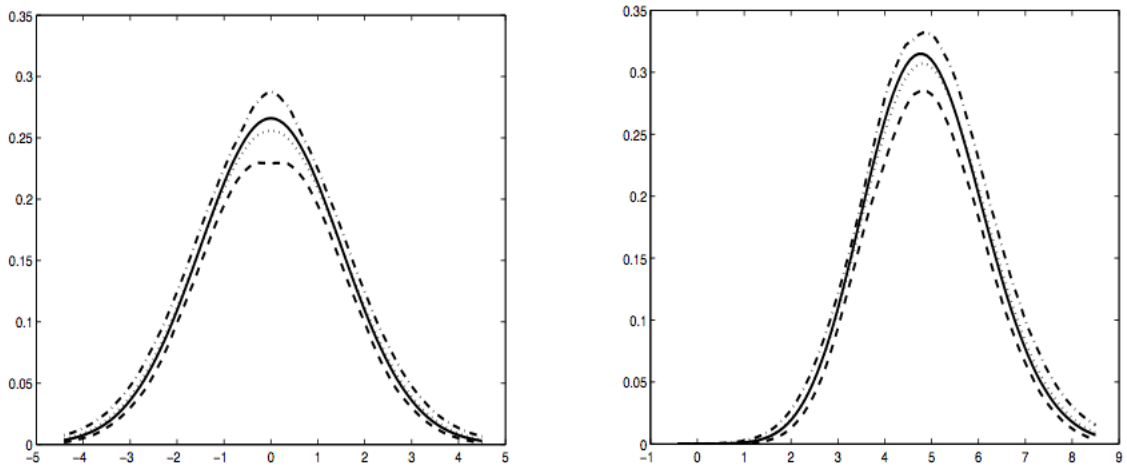

Fig. 8. Pointwise quantile curves from simulation 3. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 3. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively.

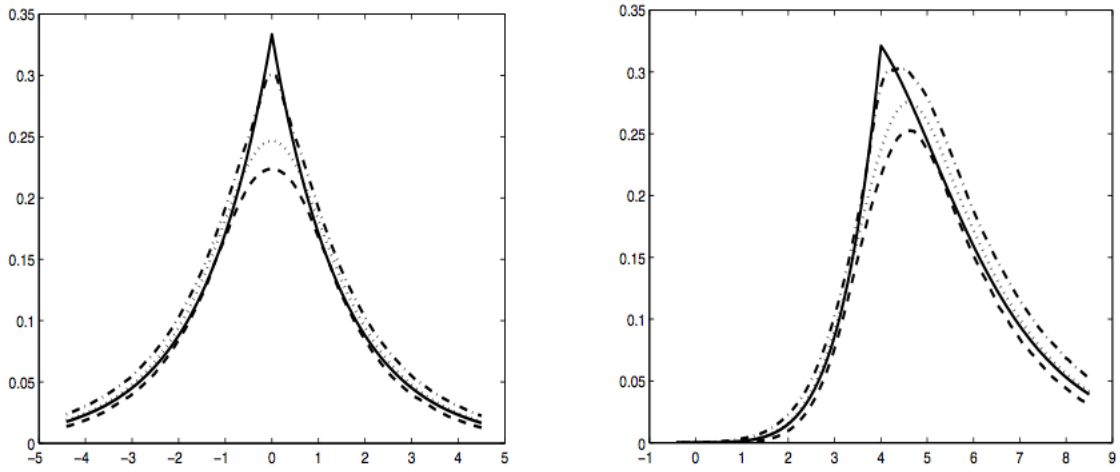

Fig. 9. Pointwise quantile curves from simulation 4. In each plot the solid line is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates in simulation 4. The left and right panels correspond to the underlying population density ($f(x)$) and the observed selected sample density ($g(x)$), respectively.

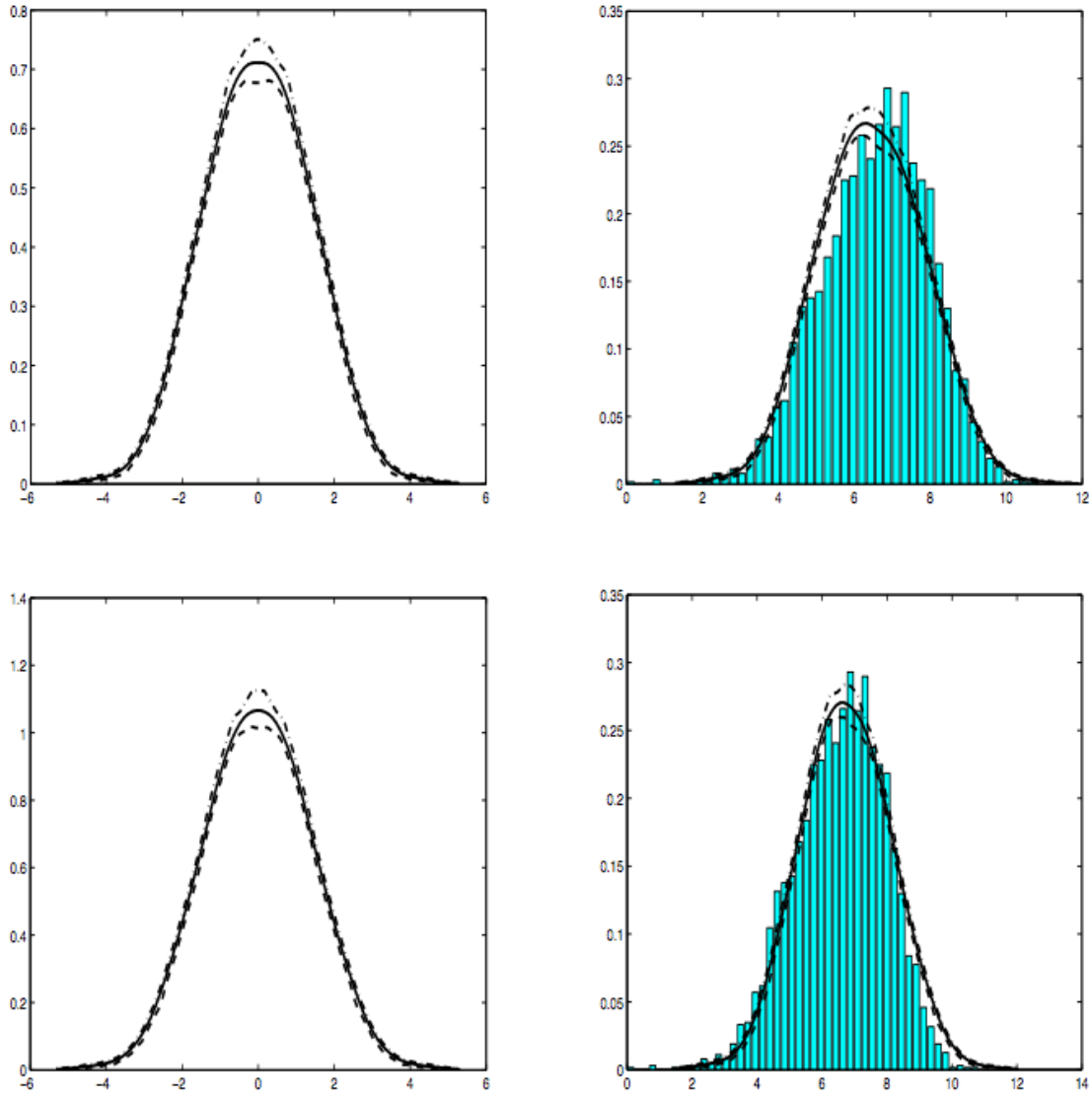

Fig. 10. The estimated densities of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ for the ambulatory expenditures data (right), under the first (upper) and second (lower) weighting functions. The estimated sample density curve is overlayed on the histogram of the observations.

Table VII. Results of the four simulation studies. Mean, sample standard deviation (sd), average of the estimated standard deviation ($\widehat{\text{sd}}$), and the 95% coverage probabilities of μ and β are reported. Results are obtained with sample size $n = 500$ and 1000 simulations.

	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	95% cvg	$\widehat{\beta}$	sd	$\widehat{\text{sd}}$	95% cvg
simulation 1	$f_0(x-\mu_0) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-\mu_0)^2}{4.5}}, w(x,\beta_0) = \frac{e^{-\beta_0 x}}{(1+e^{-\beta_0 x})^2}$, $\mu_0 = 4, \beta_0 = 1$							
est1	4.1537	0.6496	0.6537	97.7%	0.8879	0.5577	0.5880	98.0%
est2	3.9556	0.2796	0.3532	98.8%	1.0442	0.2508	0.2635	98.5%
est3	4.0494	0.5582	0.6122	98.3%	0.9866	0.5541	0.5978	98.1%
est4	3.8482	0.2628	0.3342	95.7%	1.1413	0.2647	0.3196	97.2%
est5	3.9640	0.4751	0.3022	96.9%	1.0403	0.3764	0.2568	96.7%
simulation 2	$f_0(x-\mu_0) = \frac{1}{3} e^{-\frac{\lvert x-\mu_0 \rvert}{1.5}}, w(x,\beta_0) = \frac{e^{-\beta_0 x}}{(1+e^{-\beta_0 x})^2}$, $\mu_0 = 4, \beta_0 = 1$							
est1	3.9431	0.3294	0.3088	98.6%	1.0660	0.2900	0.2838	98.6%
est2	3.8982	0.3754	0.4314	97.7%	1.1612	0.4451	0.4654	97.6%
est3	3.9116	0.2487	0.2758	96.6%	1.1050	0.2823	0.2901	96.5%
est4	3.7093	0.5895	0.5509	93.8%	1.3404	0.6264	0.6186	95.6%
est5	3.9604	0.4717	0.3430	98.4%	0.9993	0.3687	0.2839	98.1%
simulation 3	$f_0(x-\mu_0) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-\mu_0)^2}{4.5}}, w(x,\beta_0) = e^{-e^{-\beta_0 x}}$, $\mu_0 = 4, \beta_0 = 0.5$							
est1	4.0024	0.4364	0.4374	97.6%	0.5008	0.2397	0.2422	97.5%
est2	3.9758	0.2215	0.2284	97.3%	0.5130	0.1237	0.1259	96.6%
est3	3.9453	0.4073	0.4078	97.3%	0.5335	0.2251	0.2281	97.3%
est4	3.9339	0.3504	0.3488	94.2%	0.5321	0.1920	0.1881	93.8%
est5	4.0010	0.0824	0.0896	97.4%	0.5008	0.0551	0.0611	97.4%
simulation 4	$f_0(x-\mu_0) = \frac{1}{3} e^{-\frac{\lvert x-\mu_0 \rvert}{1.5}}, w(x,\beta_0) = e^{-e^{-\beta_0 x}}$, $\mu_0 = 4, \beta_0 = 0.5$							
est1	3.9871	0.1031	0.1238	98.3%	0.5153	0.0714	0.0786	96.0%
est2	3.9385	0.2215	0.2433	96.4%	0.5436	0.1144	0.1259	96.5%
est3	3.9481	0.1924	0.2151	97.8%	0.5356	0.1080	0.1195	96.3%
est4	3.9525	0.1689	0.1713	94.1%	0.5392	0.1100	0.1090	94.5%
est5	4.0030	0.0729	0.0852	97.4%	0.5048	0.0549	0.0611	97.6%

Table VIII. Five estimates of μ and β and their estimated standard deviation for the ambulatory expenditures data, under two weighting function models. The first weighting function is $w(x,\beta) = e^{-e^{-\beta x}}$, the second weighting function is $w(x,\beta) = \frac{e^{-\beta x}}{(1+e^{-\beta x})^2}$.

	First weighting function				Second weighting function			
	$\widehat{\mu}$	$\widehat{\beta}$	$\widehat{\text{sd}}(\widehat{\mu})$	$\widehat{\text{sd}}(\widehat{\beta})$	$\widehat{\mu}$	$\widehat{\beta}$	$\widehat{\text{sd}}(\widehat{\mu})$	$\widehat{\text{sd}}(\widehat{\beta})$
est1	6.5551	-0.0004	0.0270	0.0003	6.6075	-0.0264	0.6647	1.0334
est2	6.6248	-0.0544	0.0327	0.0271	8.0762	-0.9001	0.0753	0.0708
est3	6.2360	-0.0573	0.1552	0.0079	6.6678	-0.0567	0.0545	0.0187
est4	6.6523	-0.0332	0.0366	0.0113	6.7166	-0.0741	0.4023	0.2307
est5	6.5962	-0.0898	0.0584	0.0299	6.7377	-0.0921	0.0521	0.0313

CHAPTER V

CONCLUSION

In this dissertation, we have presented consistent, robust and efficient estimators in a regression model and a model under sample selection bias using semiparametric approach. We have demonstrated the theoretical properties of our estimators through asymptotic analysis and supported our theory by the numerical performance.

In a regression model, we have derived a semiparametric efficient estimator, where the regression error has conditional mean zero and conditional variance a constant. We have verified that our semiparmetric efficient (SE) estimator reaches the optimal efficiency bound in the semiparametric point of view. Our estimator has the classical root-n convergence rate and is asymptotically normal. The SE estimator is also equivalent to the second order least square estimator (SLSE) proposed in Wang and Leblanc (2008). Thus, we have concluded that SLSE is indeed efficient concerning estimation variance. In addition, we extended the model to the heterscedastic error case and derived the semiparametric efficient estimator. For the heteroscedastic model, we have adopted fixed models for the third and fourth conditional moment functions of the error distribution, and verified the consistency of our estimator even if these higher moments are misspecified. Simulation results advocated SLSE and SE estimators have the significant improvement of the estimation variance over the classical WLS estimators. Inference procedures are also supported by the simulation results.

Next, 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. To begin with, we have focused on a rather special selection process, which naturally yields a selection function π that satisfies $\pi(x) + \pi(-x) = 1$. Under this property, we have derived consistent estimators that are robust to mis-specification of the symmetric part of the model f through semiparametric method. In order to improve the efficiency, we have performed nonparametric estimation procedures, taking into account the symmetry property of the population distribution f and the characteristics of the selection procedure reflected in π. To relax the assumption on a selection function, we have allowed an arbitrary sample selection mechanism determined by the data collection procedure. Under this general framework, we have constructed a family of consistent estimators that is robust to population model mis-specification, and identified the efficient member that reaches the minimum possible estimation variance. We have demonstrated the theoretical properties of our estimators through asymptotic analysis and assess their finite sample performance through simulations. We also have implemented a real data example of ambulatory expenditures to illustrate the applicability of the methods in practice.

REFERENCES

[1] Arellano-Valle, R. B., Branco, M. D. and Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *The Canadian Journal of Statistics*, *34*, 581-601.

[2] Arellano-Valle, R. B. and Genton, M. G. (2010a). Multivariate unified skew-elliptical distributions. *Chilean Journal of Statistics*, *1*, 17-33.

[3] Arellano-Valle, R. B. and Genton, M. G. (2010b). Multivariate extended skew-t distributions and related families. *Metron*, *68*, 201-234.

[4] Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, *11*, 7-54.

[5] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171-178.

[6] Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal of Statistics*, *32*, 159-200.

[7] 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, Ser. B*, *65*, 367-389.

[8] Azzalini, A., Genton, M. G. and Scarpa, B. (2010). Invariance-based estimating equations for skew-symmetric distributions. *Metron*, *68*, 275-298.

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

[10] Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using Stata*. College Station, TX: Stata Press (Revised ed).

[11] Chamberlaine, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, *60*, 567-596.

[12] Chen, X. H., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics*, *36*, 808-843.

[13] Copas, J. B. and Li, H. G. (1997). Inference from non-random samples (with discussion). *Journal of the Royal Statistical Society, Ser. B*, *59*, 55-95.

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

[15] 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.

[16] Ma, Y. (2010). A semiparametric efficient estimator in case-control studies. *Bernoulli*, *16*, 585-603.

[17] Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, *101*, 1465-1474.

[18] Ma, Y. and Genton, M. G. (2004). A flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, *31*, 459-468.

[19] Ma, Y. and Genton, M. G. (2010). Explicit Semiparametric Estimators for Generalized Linear Latent Variable Models. *Journal of Royal Statistical Society, Series B* (in press).

[20] 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.

[21] Ma, Y. and Hart, J. (2007). Constrained local likelihood estimators for semiparametric skew-normal distributions. *Biometrika*, *94*, 119-134.

[22] Ma, Y., Chiou, J. M. and Wang, N. (2006). Efficient semiparametric estimator for heteroscedastic partially-linear models. *Biometrika*, *93*, 75-84.

[23] Maity, A., Ma, Y. and Carroll, R. J. (2007). Efficient estimation of population-level summaries in general semiparametric regression models. *Journal of the American Statistical Association*, *102*, 123-139.

[24] Marchenko, Y. V. and Genton, M. G. (2012). A Heckman selection-t model. *Journal of the American Statistical Association* (in press).

[25] Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Annals of Statistics*, *37*, 2245-2277.

[26] Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of the Applied Econometrics*, *5*, 99-135.

[27] Newey, W. and Powell, J. L. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory*, *6*, 295-317.

[28] Rabinowitz, D. (2000). Computing the efficient score in semi-parametric problems. *Statistica Sinica*, *10*, 265-280.

[29] Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: what populations does a sample represent? *A Celebration of Statistics: The ISI Centenary Volume, Ed. A. C. Atkinson & S. E. Fienberg* (pp. 543-569). New York: Springer-Verlag.

[30] Robins J. M, Rotnitzky, A. and Zhao, L. P. (1994), Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846-866.

[31] 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.

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

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

[34] Tsiatis, A. A. and Ma, Y. (2004) Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, *91*, 835-848.

[35] Wang, J., Boyer, J. and Genton, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica*, *14*, 1259-1270.

[36] Wang, L. and Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics*, *60*, 883-900.

[37] Zeng, D. and Lin, D. Y. (2007). Maximum Likelihood Estimation in Semiparametric Models with Censored Data (with discussion). *Journal of Royal Statistical Society, Series B*, *69*, 507-564.

APPENDIX A

A1. Derivation of Λ and $\Lambda^{\perp}$

We consider Λ first. From Tsiatis (2006, Section 4.5), $\Lambda_{\eta_1} =$ (all length d mean zero functions of X).

We now derive Λ_{η_2}. As a model for $p_{\epsilon|X}(\epsilon|x)$, the pdf $\eta_2(\epsilon, x)$ satisfies the following conditions:

$$\int \eta_2(\epsilon, x)d\epsilon = 1, \quad \int \epsilon\eta_2(\epsilon, x)d\epsilon = 0, \quad \int \epsilon^2\eta_2(\epsilon, x)d\epsilon = \sigma^2,$$

which can be equivalently written as

$$\int \eta_2(\epsilon, x)d\epsilon = 1, \quad \int \epsilon\eta_2(\epsilon, x)d\epsilon = 0, \quad \int (\epsilon^2 - \sigma^2)\eta_2(\epsilon, x)d\epsilon = 0.$$

Following Tsiatis (2006, Section 4.5), the first constraint implies that any function $g(\epsilon, X)$ in Λ_{η_2} has to satisfy $E(g|X) = 0$, and the second constraint implies that g has to satisfy $E(\epsilon g|X) = 0$. Applying similar arguments to the third constraint, we can obtain that g has to also satisfy $E\{(\epsilon^2 - \sigma^2)g|X\} = 0$ and, consequently, $E(\epsilon^2 g|X) = 0$. These three requirements on g yield the desired form of the space Λ_{η_2}.

We point out that the space Λ_{η_1} is orthogonal to Λ_{η_2}, which justifies the notation $\Lambda = \Lambda_{\eta_1} \oplus \Lambda_{\eta_2}$. This is because for an arbitrary element $f_1(X) \in \Lambda_{\eta_1}$ and an arbitrary element $f_2(\epsilon, X) \in \Lambda_{\eta_2}$,

$$E\{f_1(X)f_2(\epsilon, X)\} = E[E\{f_1(X)f_2(\epsilon, X)|X\}] = E[f_1(X)E\{f_2(\epsilon, X)|X\}] = 0.$$

To show the form of $\Lambda^{\perp}$, we first define a space $K = \{a(X)\epsilon + b(X)(\epsilon^2 - \sigma^2)\}$, then show $K \subset \Lambda^{\perp}$ and $\Lambda^{\perp} \subset K$.

For any function $h(\epsilon, X) = a(X)\epsilon + b(X)(\epsilon^2 - \sigma^2) \in K$, we will show that $E(hf) = 0$ for all $f \in \Lambda_{\eta_1}$ and $E(hg) = 0$ for all $g \in \Lambda_{\eta_2}$. This would demonstrate that $h \in \Lambda^{\perp}$. We have

$$\begin{aligned}
E\{h(\epsilon, X)f^T(X)\} &= E(E[\{a(X)\epsilon + b(X)(\epsilon^2 - \sigma^2)\}f^T(X)|X]) \\
&= E\{a(X)f^T(X)E(\epsilon|X)\} + E\{b(X)f^T(X)E(\epsilon^2 - \sigma^2|X)\} \\
&= 0, \\
E\{h(\epsilon, X)g^T(\epsilon, X)\} &= E(E[\{a(X)\epsilon + b(X)(\epsilon^2 - \sigma^2)\}g^T(\epsilon, X)|X]) \\
&= E\{a(X)E(\epsilon g^T|X)\} + E\{b(X)E(\epsilon^2 g^T|X)\} - \sigma^2 E\{b(X)E(g^T|X)\} \\
&= 0.
\end{aligned}$$

Thus, $K \subset \Lambda^{\perp}$.

To show $\Lambda^{\perp} \subset K$, we consider an arbitrary $h \in \Lambda^{\perp}$. Let $\Lambda_{\eta_2} = \Lambda_a \cap \Lambda_b \cap \Lambda_c$, where

$$\Lambda_a = \{g : E(g|X) = 0\}, \ \Lambda_b = \{g : E(\epsilon g|X) = 0\}, \ \Lambda_c = \left\{g : E(\epsilon^2 g|X) = 0\right\}.$$

Lemma 4.3 of Tsiatis (2006) implies that $\Lambda_{\eta_1}^{\perp} = \Lambda_a$. It is then trivial to see that $h \in \Lambda^{\perp}$ implies $h \perp \Lambda_{\eta_1}$, which further implies $h \in \Lambda_a$. Thus $E(h|X) = 0$. Form $r(\epsilon, X) = E(\epsilon h|X)\epsilon/\sigma^2 + E(Ch|X)C/E(C^2|X)$, where C is defined after (2.4) and decompose h as

$$h = \{h - E(\epsilon h|X)\epsilon/\sigma^2 - E(Ch|X)C/E(C^2|X)\} + r.$$

Note $r \in K \subset \Lambda^{\perp}$, hence $h_1 = h - E(\epsilon h|X)\epsilon/\sigma^2 - E(Ch|X)C/E(C^2|X) = h - r \in \Lambda^{\perp}$ as well. However, we can easily verify that $h_1 \in \Lambda_{\eta_2}$ at the same time, by verifying that $E(h_1|X) = 0, E(\epsilon h_1|X) = 0$ and $E(\epsilon^2 h_1|X) = 0$. Hence, $h_1 = 0$. This indicates $h = r \in K$, thus $\Lambda^{\perp} \subset K$.

A2. Proof of (2.4)

Let $r(\epsilon, X) = \dfrac{E(\epsilon h|X)}{\sigma^2}\epsilon + \dfrac{E(Ch|X)}{E(C^2|X)}C$. Obviously $r(\epsilon, X) \in \Lambda^{\perp}$. Decompose $h - r$ as

$$h(\epsilon, X) - r(\epsilon, X) = E(h|X) + \{h(\epsilon, X) - r(\epsilon, X) - E(h|X)\}.$$

Note that $E(h|X) \in \Lambda_{\eta_1}$. We can also verify that $h(\epsilon, X) - r(\epsilon, X) - E(h|X) \in \Lambda_{\eta_2}$, by verifying that $E[\{h - r - E(h|X)\}|X] = 0$, $E[\epsilon\{h - r - E(h|X)\}|X] = 0$ and $E[\epsilon^2\{h - r - E(h|X)\}|X] = 0$.

Hence $h(\epsilon, X) - r(\epsilon, X) \in \Lambda$. Thus, we obtain that $\Pi(h|\Lambda) = h(\epsilon, X) - r(\epsilon, X)$ and $\Pi(h|\Lambda^{\perp}) = r(\epsilon, X)$. □

A3. Calculation of S_{eff} given in (2.5)

$S_{\text{eff}}(X, Y)$ can be written as

$$S_{\text{eff}} = \Pi(S_\theta|\Lambda^{\perp}) = \frac{E(\epsilon S_\theta|X)}{\sigma^2}\epsilon + \frac{E(CS_\theta|X)}{E(C^2|X)}C.$$

Using the form of S_β and S_{σ^2} in (2.3), we can verify that $E(\epsilon S_\beta|X) = \dfrac{\partial m(X;\beta)}{\partial \beta}$ and $E(CS_\beta|X) = -\dfrac{\partial m(X;\beta)}{\partial \beta}\dfrac{\mu_3}{\sigma^2}$, thus

$$S_{\beta,\text{eff}} = \frac{\partial m(X;\beta)}{\partial \beta}\left\{\frac{\epsilon}{\sigma^2} - \frac{\mu_3}{\sigma^2 E(C^2|X)}C\right\}.$$

Similarly, we can verify that $E(\epsilon S_{\sigma^2}|X) = 0$ and $E(CS_{\sigma^2}|X) = 1$, hence $S_{\sigma^2,\text{eff}} = C/E(C^2|X)$.

A4. Derivation of the variances in (2.6)

Using the explicit form of S_θ, we have

$$S_{\text{eff}}S_{\text{eff}}^T = \begin{bmatrix} \frac{\partial m(X;\beta)}{\partial \beta}\frac{\partial m(X;\beta)}{\partial \beta^T}\left\{\frac{\epsilon}{\sigma^2} - \frac{\mu_3 C}{\sigma^2 E(C^2|X)}\right\}^2 & \frac{\partial m(X;\beta)}{\partial \beta}\left\{\frac{\epsilon}{\sigma^2} - \frac{\mu_3 C}{\sigma^2 E(C^2|X)}\right\}\frac{C}{E(C^2|X)} \\ \frac{\partial m(X;\beta)}{\partial \beta^T}\left\{\frac{\epsilon}{\sigma^2} - \frac{\mu_3 C}{\sigma^2 E(C^2|X)}\right\}\frac{C}{E(C^2|X)} & \frac{C^2}{E(C^2|X)^2} \end{bmatrix}.$$

Taking expectation of $S_{\text{eff}}S_{\text{eff}}^T$ evaluated at the true parameter values, we have

$$\begin{aligned} & E(S_{\text{eff}}S_{\text{eff}}^T|_{\theta=\theta_0}) \\ = & \begin{bmatrix} \frac{1}{\sigma_0^2}E\left\{\frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\partial m(X;\beta_0)}{\partial \beta^T}\right\} + \frac{1}{\sigma_0^4}E\left\{\frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\partial m(X;\beta_0)}{\partial \beta^T}\frac{\mu_3^2}{E(C^2|X)}\right\} & -\frac{1}{\sigma_0^2}E\left\{\frac{\partial m(X;\beta_0)}{\partial \beta}\frac{\mu_3}{E(C^2|X)}\right\} \\ -\frac{1}{\sigma_0^2}E\left\{\frac{\partial m(X;\beta_0)}{\partial \beta^T}\frac{\mu_3}{E(C^2|X)}\right\} & E\left\{\frac{1}{E(C^2|X)}\right\} \end{bmatrix} \\ = & \begin{bmatrix} \frac{1}{\sigma_0^2}\left\{1 + \frac{\mu_3^2}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2}\right\}B & -\frac{\mu_3}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2}A \\ -\frac{\mu_3}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2}A^T & \frac{\sigma_0^2}{\sigma_0^2(\mu_4-\sigma_0^4)-\mu_3^2} \end{bmatrix}. \end{aligned}$$

Its inverse can then be calculated using the matrix inversion and is easy to verify to have the form in (2.6).

APPENDIX B

B1. Establishment of Λ

Consider the set of functions $A = \{u(x-\mu) : u(t) = u(-t), \int_0^\infty u(t) f_0(t) dt = 0\}$. We show $A = \Lambda$ through showing both $A \subset \Lambda$ and $\Lambda \subset A$.

We first show $A \subset \Lambda$. Let $u(x-\mu) \in A$. Then $\int_0^\infty u(t) f_0(t) dt = 0$ and u is an even function. We need to show $u(x-\mu) \in \mathcal{H}$. From

$$\begin{aligned} 0 &= \int_0^\infty u(t) 2 f_0(t) dt \\ &= \int_0^\infty u(t) 2 f_0(t) \pi(t;\beta) dt + \int_0^\infty u(t) 2 f_0(t) \pi(-t;\beta) dt \\ &= \int_0^\infty u(t) 2 f_0(t) \pi(t;\beta) dt + \int_{-\infty}^0 u(t) 2 f_0(t) \pi(t;\beta) dt \\ &= \int_{-\infty}^\infty u(t) 2 f_0(t) \pi(t;\beta) dt \\ &= E\{u(X-\mu)\}, \end{aligned}$$

we obtain $u(x-\mu) \in \mathcal{H}$. Now consider

$$g(x;\theta,\gamma) = 2 f(x-\mu;\gamma) \pi(x-\mu;\beta),$$

where

$$f(t;\gamma) = \frac{f_0(t)\{1 + e^{-2\gamma^T u(t)}\}^{-1}}{\int f_0(t)\{1 + e^{-2\gamma^T u(t)}\}^{-1} dt}$$

and γ is a nuisance parameter. Note that $\gamma = 0$ yields the true model. We have

$$\left.\frac{\partial \log g(x;\theta,\gamma)}{\partial \gamma}\right|_{\gamma=0} = \left.\frac{\partial f(x-\mu_0;\gamma)/\partial\gamma}{f(x-\mu_0;\gamma)}\right|_{\gamma=0} = \frac{\partial f(x-\mu_0;\gamma)/\partial\gamma\big|_{\gamma=0}}{f_0(x-\mu_0)} = u(x-\mu).$$

Therefore, $u(x-\mu)$ is a nuisance score vector of a particular submodel, hence $u(x-\mu) \in \Lambda$.

We next show $\Lambda \subset A$. Consider an element of Λ which is the nuisance score of an arbitrary parametric submodel $2f(x-\mu;\gamma)\pi(x-\mu;\beta)$. Then, we can write it as

$$u(x-\mu) = \left.\frac{\partial f(x-\mu;\gamma)/\partial\gamma}{f(x-\mu;\gamma)}\right|_{\gamma=0}.$$

Since $f(t;\gamma) = f(-t;\gamma)$, we have $\partial f(t;\gamma)/\partial\gamma = \partial f(-t;\gamma)/\partial\gamma$ for any γ. It implies

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

for any γ. Thus, we obtain $u(t) = u(-t)$. Furthermore, we have

$$\begin{aligned}
\int_0^\infty u(t)2f_0(t)dt &= \int_0^\infty u(t)2f_0(t)\{\pi(t;\beta)+\pi(-t;\beta)\}dt \\
&= \int_0^\infty u(t)2f_0(t)\pi(t;\beta)dt + \int_0^\infty u(t)2f_0(t)\pi(-t;\beta)dt \\
&= \int_0^\infty u(t)2f_0(t)\pi(t;\beta)dt + \int_0^\infty u(-t)2f_0(-t)\pi(-t;\beta)dt \\
&= \int_0^\infty u(t)2f_0(t)\pi(t;\beta)dt + \int_{-\infty}^0 u(t)2f_0(t)\pi(t;\beta)dt \\
&= \int_{-\infty}^\infty u(t)2f_0(t)\pi(t;\beta)dt \\
&= \int_{-\infty}^\infty u(x-\mu)2f_0(x-\mu)\pi(x-\mu;\beta)dx \\
&= E\{u(X-\mu)\} = 0.
\end{aligned}$$

Because $u(t)$ is symmetric and $\int_0^\infty u(t)f_0(t)dt = 0$, $u(x-\mu) \in A$.

B2. Establishment of $\Lambda^\perp$

To establish the form of $\Lambda^\perp$, we first define a space $K = \{v(x-\mu) : v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta) = 0\}$, then show $K \subset \Lambda^\perp$ and $\Lambda^\perp \subset K$. For any even function u, we have

$$\begin{aligned} E\{u(X-\mu)v(X-\mu)\} &= \int u(x-\mu)v(x-\mu)2f_0(x-\mu)\pi(x-\mu;\beta)dx \\ &= \int u(x)v(x)2f_0(x)\pi(x;\beta)dx \\ &= \int_0^\infty u(t)v(t)2f_0(t)\pi(t;\beta)dt + \int_0^\infty u(t)v(-t)2f_0(t)\pi(-t;\beta)dt \\ &= \int_0^\infty u(t)2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}dt. \end{aligned}$$

We first show $K \subset \Lambda^\perp$. For any function $v(x-\mu) \in K$ and any function $u \in \Lambda$, we have

$$E\{u(X-\mu)v(X-\mu)\} = \int_0^\infty u(t)2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}dt = \int_0^\infty u(t)0dt = 0.$$

Hence $v(x-\mu) \perp \Lambda$. In addition,

$$E\{v(X-\mu)\} = E\{1v(X-\mu)\} = \int_0^\infty 2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}dt = \int_0^\infty 0dt = 0.$$

Hence $v(x-\mu) \in \mathcal{H}$. Using the above two equalities, we have $v(x-\mu) \in \Lambda^\perp$, hence $K \subset \Lambda^\perp$.

Next, we show $\Lambda^\perp \subset K$. Suppose $v(x-\mu) \in \Lambda^\perp$, then $E\{u(X-\mu)v(X-\mu)\} = 0$ for any $u(x-\mu) \in \Lambda$. Denote $w(t) = v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta)$, we have

$$w(t) = v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta) = w(-t),$$

which implies that $w(t)$ is symmetric. Since $v(x-\mu) \in \mathcal{H}$, we have

$$0 = E\{v(X-\mu)\} = \int_0^\infty 2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}dt = \int_0^\infty f_0(t)w(t)dt = 0.$$

Hence $w(x-\mu) \in \Lambda$. Thus, we can let $u(t) = v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta)$, and obtain

$$\begin{aligned} 0 &= E\{u(X-\mu)v(X-\mu)\} \\ &= \int_0^\infty u(t)2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}dt \\ &= \int_0^\infty 2f_0(t)\{v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)\}^2dt. \end{aligned}$$

Hence $v(t)\pi(t;\beta)+v(-t)\pi(-t;\beta)=0$. This indicates $v(x-\mu)\in K$. Hence $\Lambda^{\perp}\subset K$.

B3. Verification of the orthogonal projection of the score function

Obviously,

$$\frac{f_0'(x-\mu)}{f_0(x-\mu)}\{\pi(x-\mu;\beta)-\pi(-x+\mu;\beta)\}+2\pi_x'(x-\mu;\beta)$$

is symmetric. We also have

$$\begin{aligned}&\int_0^{\infty}\left[f_0'(t)\{2\pi(t;\beta)-1\}+2\pi_t'(t;\beta)f_0(t)\right]dt=\int_0^{\infty}\frac{\partial\{2f_0(t)\pi(t;\beta)-f_0(t)\}}{\partial t}dt\\ =\;&-\{2f_0(0)\pi(0;\beta)-f_0(0)\}=0.\end{aligned}$$

Thus,

$$\frac{f_0'(x-\mu)}{f_0(x-\mu)}\{\pi(x-\mu;\beta)-\pi(-x+\mu;\beta)\}+2\pi_x'(x-\mu;\beta)\in\Lambda.$$

On the other hand, we have

$$\begin{aligned}&\left\{\frac{f_0'(t)2\pi(-t;\beta)}{f_0(t)}+\frac{\pi_t'(t;\beta)}{\pi(t;\beta)}-2\pi_t'(t;\beta)\right\}\pi(t;\beta)\\ &+\left\{\frac{f_0'(-t)2\pi(t;\beta)}{f_0(-t)}+\frac{\pi_t'(-t;\beta)}{\pi(-t;\beta)}-2\pi_t'(-t;\beta)\right\}\pi(-t;\beta)\\ =\;&2\pi(t;\beta)\pi(-t;\beta)\left\{\frac{f_0'(t)}{f_0(t)}+\frac{f_0'(-t)}{f_0(-t)}\right\}+\pi_t'(t;\beta)+\pi_t'(-t;\beta)\\ &-2\pi_t'(t;\beta)\pi(t;\beta)-2\pi_t'(-t;\beta)\pi(-t;\beta)\\ =\;&2\pi(t;\beta)\pi(-t;\beta)\cdot 0+2\pi_t'(t;\beta)-2\pi_t'(t;\beta)\pi(t;\beta)-2\pi_t'(t;\beta)\{1-\pi(t;\beta)\}\\ =\;&0.\end{aligned}$$

In the second equality above, we used the fact that f_0' is an odd function, and π_t' is

an even function with respect to the first argument. Hence, we have shown

$$-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)}-\frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)}+2\pi_x'(x-\mu;\beta)\in\Lambda^{\perp}.$$

Combining the above results, we obtain that

$$\Pi(S_\mu|\Lambda^{\perp})=-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)}-\frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)}+2\pi_x'(x-\mu;\beta).$$

B4. Proof of Lemma 1

A Taylor expansion of $\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\widehat{\gamma},f^*)$ at γ^* gives the result

$$\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\widehat{\gamma},f^*)=\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\gamma^*,f^*)+\left\{\sum_{i=1}^n \partial S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\tilde{\gamma},f^*)/\partial\gamma^T\right\}(\widehat{\gamma}-\gamma^*)$$

where $\tilde{\gamma}$ is between γ^* and $\widehat{\gamma}$.

Letting $S_n=\{\sum_{i=1}^n \partial S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\tilde{\gamma},f^*)/\partial\gamma^T\}/n$, we obtain

$$\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\widehat{\gamma},f^*)=nS_n(\widehat{\gamma}-\gamma^*).$$

Since $\widehat{\gamma}$ converges to γ^* as $n\to\infty$, we obtain $S_n\to E\{\partial S_{\theta,\text{eff}}(X_i;\theta_0,\gamma^*,f^*)/\partial\gamma^T\}=0$ in probability.

A Taylor expansion of $\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\widehat{\gamma},f^*)$ at $\widehat{\theta}_2$ yields

$$\begin{aligned}\widehat{\theta}_1-\widehat{\theta}_2 &= \left\{\sum_{i=1}^n\frac{\partial S_{\theta,\text{eff}}(X_i;\tilde{\theta},\widehat{\gamma},f^*)}{\partial\theta^T}\right\}^{-1}\left\{\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta}_1,\widehat{\gamma},f^*)-0\right\}\\ &= \left\{\frac{1}{n}\sum_{i=1}^n\frac{\partial S_{\theta,\text{eff}}(X_i;\tilde{\theta},\widehat{\gamma},f^*)}{\partial\theta^T}\right\}^{-1}S_n(\widehat{\gamma}-\gamma^*)\end{aligned}$$

where $\tilde{\theta}$ is between $\widehat{\theta}_2$ and $\widehat{\theta}_1$.

As $n\to\infty$,

$$J_n=\frac{1}{n}\sum_{i=1}^n\frac{\partial S_{\theta,\text{eff}}(X_i;\tilde{\theta},\widehat{\gamma},f^*)}{\partial\theta^T}\to E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)}{\partial\theta^T}\right\}$$

in probability. We denote $E\{\partial S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)/\partial\theta^T\}$ by J, which is the Fisher information matrix and is nonsingular in general. Combining the above results, we have $n^{1/2}(\widehat{\theta}_1-\widehat{\theta}_2)=n^{1/2}J_n^{-1}S_n(\widehat{\gamma}-\gamma^*)$. Since $n^{1/2}(\widehat{\gamma}-\gamma^*)$ is bounded in probability, $J_n^{-1}\to J^{-1}$ in probability and $S_n\to 0$ in probability, and we thus have $\sqrt{n}(\widehat{\theta}_1-\widehat{\theta}_2)\to 0$ in probability. □

B5. Proof of Theorem 3

Whether or not f^* is a correct model, the nuisance score vector is

$$S_\gamma(x;\theta,\gamma,f^*)=\frac{\partial \log f^*(x;\theta,\gamma)}{\partial\gamma}=\frac{\partial f^*(x-\mu;\gamma)/\partial\gamma}{f^*(x-\mu;\gamma)}$$

and the estimators $\widehat{\theta},\widehat{\gamma}$ satisfy

$$\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta},\widehat{\gamma},f^*)=0 \quad \text{and} \quad \sum_{i=1}^n S_\gamma(X_i;\widehat{\theta},\widehat{\gamma},f^*)=0.$$

We have

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\widehat{\theta},\widehat{\gamma},f^*) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\theta_0,\widehat{\gamma},f^*)+\frac{1}{n}\sum_{i=1}^n \frac{\partial S_{\theta,\text{eff}}(X_i;\theta^*,\widehat{\gamma},f^*)}{\partial\theta^T}\sqrt{n}(\widehat{\theta}-\theta_0) \end{aligned}$$

where θ^* is between θ_0 and $\widehat{\theta}$. Notice that when $n\to\infty$,

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial S_{\theta,\text{eff}}(X_i;\theta^*,\widehat{\gamma},f^*)}{\partial\theta^T}\to E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,\widehat{\gamma},f^*)}{\partial\theta^T}\right\}$$

in probability. It follows that

$$\sqrt{n}(\widehat{\theta}-\theta_0)=-E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,\widehat{\gamma},f^*)}{\partial\theta^T}\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n S_{\theta,\text{eff}}(X_i;\theta_0,\widehat{\gamma},f^*)+o_p(1).$$

Using Lemma 1, we further have

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)}{\partial\theta^T}\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta_0,\gamma^*,f^*)+o_p(1),$$

where $\gamma^*=\gamma_0$ is f^* contains f_0. We thus have

$$\sqrt{n}(\widehat{\theta}-\theta_0)\to N\left[0, A^{-1}E\{S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)S^T_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)\}A^{-T}\right],$$

where A is defined in Theorem 3. It can be easily verified that when f^* contains f_0, $\gamma^*=\gamma_0$ and $A=E\{S_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)S^T_{\theta,\text{eff}}(X;\theta_0,\gamma^*,f^*)\}$, thus this completes the proof of Theorem 3. □

B6. Proof of Proposition 1

The square norm of $\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})-f_0(x-\mu_0)$ is

$$\parallel \widehat{f}(x-\widetilde{\mu};\widetilde{\mu})-f_0(x-\mu_0)\parallel = \left[\int_{-\infty}^{\infty}\left\{\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})-f_0(x-\mu_0)\right\}^2\pi(x-\mu_0,\beta_0)dx\right]^{1/2}.$$

Let $a_{n_2}=\max\{-X_{(1)}+\widetilde{\mu}, X_{(n_2)}-\widetilde{\mu}\}$. Since the kernel K has support $(-1,1)$, $\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})=0$ for $x\le -a_{n_2}+\widetilde{\mu}-h$ or $x\ge a_{n_2}+\widetilde{\mu}+h$.

The above is derived from the following details:

$$\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})=\frac{1}{2n_2}\sum_{i=n_1+1}^{n_1+n_2}\frac{1}{h}\left\{K\left(\frac{X_i-x}{h}\right)+K\left(\frac{X_i+x-2\widetilde{\mu}}{h}\right)\right\}.$$

From the above kernel estimation, $\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})=0$ for x such that

$$\begin{aligned}
\frac{X_{(n_2)}-x}{h}<-1 &\longrightarrow x>X_{(n_2)}+h \text{ and}\\
\frac{X_{(1)}-x}{h}>1 &\longrightarrow x<X_{(1)}-h \text{ and}\\
\frac{X_{(n_2)}+x-2\widetilde{\mu}}{h}<-1 &\longrightarrow x<-X_{(n_2)}+2\widetilde{\mu}-h \text{ and}\\
\frac{X_{(1)}+x-2\widetilde{\mu}}{h}>1 &\longrightarrow x>-X_{(1)}+2\widetilde{\mu}+h
\end{aligned}$$

We thus have

$$\begin{aligned} & \| \widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0) \|^2 \\ = & \int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h} \left\{ \widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0) \right\}^2 \pi(x-\mu_0,\beta_0) dx \\ \leq & \int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h} \left\{ \widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0) \right\}^2 dx \end{aligned} \tag{B.1}$$

The above inequality holds because $\pi(x-\mu_0,\beta_0)$ is a skewing function which is nonnegative and not greater than 1. We need to show that (B.1) is $o_p(n_2^{-1/2})$.

We replace x with $t+\widetilde{\mu}$ in the following.

For any sequence of positive constants C_{n_2}, we have

$$\begin{aligned} & \mathrm{pr}\left[\int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h} \left\{ \widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0) \right\}^2 dx > \frac{\epsilon}{\sqrt{n_2}} \right] \\ = & \mathrm{pr}\left[\int_{-a_{n_2}-h}^{a_{n_2}+h} \left\{ \widehat{f}(t;\widetilde{\mu}) - f_0(t+\widetilde{\mu}-\mu_0) \right\}^2 dt > \frac{\epsilon}{\sqrt{n_2}} \right] \\ \leq & \mathrm{pr}\left[\int_{-C_{n_2}}^{C_{n_2}} \left\{ \widehat{f}(t;\widetilde{\mu}) - f_0(t+\widetilde{\mu}-\mu_0) \right\}^2 dt > \frac{\epsilon}{\sqrt{n_2}} \right] + \mathrm{pr}\left(-a_{n_2}-h \leq -C_{n_2}\right) \\ & + \mathrm{pr}\left(a_{n_2}+h \geq C_{n_2}\right). \end{aligned}$$

From condition (i), we have $\mathrm{pr}(-a_{n_2}-h \leq -C_{n_2}) = o(1)$ and $\mathrm{pr}(a_{n_2}+h \geq C_{n_2}) = o(1)$ if $C_{n_2} = (C\log n)^{1/2}$ for $C > 4$ and $n \to \infty$ because $h - C_{n_2}$ tends to $-\infty$ and $-h + C_{n_2}$ tends to ∞ as n goes to ∞, while both $X_{(1)}$ and $X_{(n_2)}$ are bounded.

Note that $\displaystyle\int_{-C_{n_2}}^{C_{n_2}} \left\{ \widehat{f}(t;\widetilde{\mu}) - f_0(t+\widetilde{\mu}-\mu_0) \right\}^2 dt$ is a function of random variable $\widetilde{\mu}$ and X_i's for $i = n_1+1, \cdots, n_1+n_2$. Let E_0 denote the expectation with respect to the distribution of $\widetilde{\mu}$, while pr^* and E^* denote probability and expectation with respect to the conditional distribution of $X_{n_1+1}, \cdots, X_{n_1+n_2}$ given $\widetilde{\mu}$. Defining

$$F_n = F_n(\epsilon) = \mathrm{pr}^*\left[\int_{-C_{n_2}}^{C_{n_2}} \left\{ \widehat{f}(t;\widetilde{\mu}) - f_0(t+\widetilde{\mu}-\mu_0) \right\}^2 dt > \frac{\epsilon}{\sqrt{n_2}} \right],$$

we have

$$\text{pr}\left[\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t+\widetilde{\mu}-\mu_0)\right\}^2dt>\frac{\epsilon}{\sqrt{n_2}}\right]=E_0F_n.$$

Since dominated convergence in probability implies convergence in mean by the theorm (Serfling, 2002, 1.3.6), it suffices to show that F_n coverges in probability to 0 for each $\epsilon>0$. By Markov's inequality, we have

$$F_n \;\leq\; \frac{\sqrt{n_2}}{\epsilon}E^*\left[\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t+\widetilde{\mu}-\mu_0)\right\}^2dt\right].$$

Since $\widetilde{\mu}-\mu_0=O_p(n^{-\frac{1}{2}})$, Taylor expansion at t yields

$$\begin{aligned}
& E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t+\widetilde{\mu}-\mu_0)\right\}^2dt \\
= \;& E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t)-(\widetilde{\mu}-\mu_0)f_0'(t)-(\widetilde{\mu}-\mu_0)^2f_0''(t^*)/2\right\}^2dt \\
= \;& E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t)\right\}^2dt+O_p(n_2^{-1}),
\end{aligned}$$

where t^* is between t and $t+\widetilde{\mu}-\mu_0$. In the last equality, we used the fact that $\{\widehat{f}(t;\widetilde{\mu})-f_0(t)\}f_0'(t)$ is an odd fucntion of t.

Assuming we can exchange expectation and integration, it is sufficient to show

$$\frac{\sqrt{n_2}}{\epsilon}\int_{-C_{n_2}}^{C_{n_2}}E^*\left\{\widehat{f}(t;\widetilde{\mu})-f_0(t)\right\}^2dt$$

converges 0 in probability.

$$\begin{aligned}
& E^*\{\widehat{f}(t;\widetilde{\mu})-f_0(t)\}^2=\text{MSE}\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\}=\text{Var}\{f(t;\widetilde{\mu})|\widetilde{\mu}\}+\text{Bias}^2\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\} \\
= \;& \frac{1}{n_2h}f_0(t)\int_{-1}^{1}K^2(s_1)ds_1+\frac{h^4}{4}\{f_0''(t)\}^2\left\{\int_{-1}^{1}K(s_2)s_1^2ds_1\right\}^2+O_p(n_2^{-1}h) \\
& +O_p(n_2^{-3/2}h^{-1}).
\end{aligned}$$

It follows that MISE$\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\}$ have the following asymptotic expression,

$$\int_{-C_{n_2}}^{C_{n_2}} E^*\{\widehat{f}(t;\widetilde{\mu}) - f_0(t)\}^2 dt = \int_{-C_{n_2}}^{C_{n_2}} \text{MSE}\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\}dt$$

$$= \frac{1}{n_2 h}\int_{-1}^{1} K^2(s_1)ds_1 + \frac{h^4}{4}\left\{\int_{-1}^{1} K(s_2)s_1^2 ds_1\right\}^2 \int_{-C_{n_2}}^{C_{n_2}} \{f_0''(t)\}^2 dt + O_p(n_2^{-1}h)$$
$$+O_p(n_2^{-3/2}h^{-1}).$$

From condition (iv), $h = O_p(n^{-1/5})$ and we thus have

$$\frac{\sqrt{n_2}}{\epsilon}\int_{-C_{n_2}}^{C_{n_2}} E^*\left\{\widehat{f}(t;\widetilde{\mu}) - f_0(t)\right\}^2 dt = \frac{\sqrt{n_2}}{\epsilon}\text{MISE}\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\} = o_p(1).$$

□

B7. Proof of Corollary 1

The square norm of $\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0)$ is

$$\| \widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0) \| = \left[\int_{-\infty}^{\infty}\left\{\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0)\right\}^2 \pi(x-\mu_0,\beta_0)dx\right]^{1/2}.$$

Let $a_{n_2} = \max\{-X_{(1)} + \widetilde{\mu}, X_{(n_2)} - \widetilde{\mu}\}$. Since the kernel K has support $(-1,1)$, $\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) = 0$ for $x \le -a_{n_2} + \widetilde{\mu} - h$ or $x \ge a_{n_2} + \widetilde{\mu} + h$.

The above is derived from the following details:

$$\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) = \frac{1}{2n_2}\sum_{i=n_1+1}^{n_1+n_2}\frac{1}{h^2}\left\{K'\left(\frac{x-X_i}{h}\right) + K'\left(\frac{X_i + x - 2\widetilde{\mu}}{h}\right)\right\},$$

From the above kernel estimation, $\widehat{f}(x-\widetilde{\mu};\widetilde{\mu})=0$ for x such that

$$\begin{aligned}
\frac{x-X_{(1)}}{h}<-1 &\longrightarrow x<X_{(1)}-h \text{ and}\\
\frac{x-X_{(n_2)}}{h}>1 &\longrightarrow x>X_{(n_2)}+h \text{ and}\\
\frac{X_{(n_2)}+x-2\widetilde{\mu}}{h}<-1 &\longrightarrow x<-X_{(n_2)}+2\widetilde{\mu}-h \text{ and}\\
\frac{X_{(1)}+x-2\widetilde{\mu}}{h}>1 &\longrightarrow x>-X_{(1)}+2\widetilde{\mu}+h
\end{aligned}$$

It follows that

$$\begin{aligned}
&\| \widehat{f}'(x-\widetilde{\mu};\widetilde{\mu})-f_0'(x-\mu_0)\|^2\\
=&\int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h}\left\{\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu})-f_0'(x-\mu_0)\right\}^2\pi(x-\mu_0,\beta_0)dx\\
\leq&\int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h}\left\{\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu})-f_0'(x-\mu_0)\right\}^2 dx.
\end{aligned}$$

Now we need to show the above display is $o_p(n^{-1/2})$. For any sequence of positive constants C_{n_2}, we have

$$\begin{aligned}
&\text{pr}\left[\int_{-a_{n_2}+\widetilde{\mu}-h}^{a_{n_2}+\widetilde{\mu}+h}\left\{\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu})-f_0'(x-\mu_0)\right\}^2 dx>\epsilon\right]\\
=&\text{pr}\left[\int_{-a_{n_2}-h}^{a_{n_2}+h}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t+\widetilde{\mu}-\mu_0)\right\}^2 dt>\epsilon\right]\\
\leq&\text{pr}\left[\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t+\widetilde{\mu}-\mu_0)\right\}^2 dt>\epsilon\right]+\text{pr}\left(-a_{n_2}-h\leq -C_{n_2}\right)\\
&+\text{pr}\left(a_{n_2}+h\geq C_{n_2}\right).
\end{aligned}$$

From condition (i), we have $\text{pr}(-a_{n_2}-h\leq -C_{n_2})=o(1)$ and $\text{pr}(a_{n_2}+h\geq C_{n_2})=o(1)$ if $C_{n_2}=(C^*\log n)^{1/2}$ for $C^*>4$ and $n\to\infty$ because $h-C_{n_2}$ tends to $-\infty$ and $-h+C_{n_2}$ tends to ∞ as n goes to ∞, while both $X_{(1)}$ and $X_{(n_2)}$ are bounded.

Similarly to the proof of Proposition 1, we only need to show

$$\frac{\sqrt{n_2}}{\epsilon}E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t+\widetilde{\mu}-\mu_0)\right\}^2dt\to 0.$$

Since $\widetilde{\mu}-\mu_0=O_p(n^{-\frac{1}{2}})$, Taylor expansion at t yields

$$\begin{aligned}
& E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t+\widetilde{\mu}-\mu_0)\right\}^2dt \\
= \; & E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t)-(\widetilde{\mu}-\mu_0)f_0''(t)+o_p(n^{-1})\right\}^2dt \\
= \; & E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t)\right\}^2dt+O_p(n^{-1}).
\end{aligned}$$

In the last equality, we use the fact that $\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t)\}f_0''(t)$ is an odd fucntion of t. Assuming we can exchange expectation and integration, it suffices to show

$$\frac{\sqrt{n_2}}{\epsilon}E^*\int_{-C_{n_2}}^{C_{n_2}}\left\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t)\right\}^2dt\to 0.$$

We calculate MSE in the following way,

$$\begin{aligned}
\mathrm{MSE}\{\widehat{f}'(t;\widetilde{\mu})|\widetilde{\mu}\} &= E^*\{\widehat{f}'(t;\widetilde{\mu})-f_0'(t)\}^2=\mathrm{Var}\{\widehat{f}'(t;\widetilde{\mu})|\widetilde{\mu}\}+\mathrm{Bias}^2\{\widehat{f}'(t;\widetilde{\mu})|\widetilde{\mu}\} \\
&= \frac{1}{n_2h}\{-f_0(t)\pi(t,\beta_0)+f_0(-t)\pi(-t,\beta_0)\}\int_{-1}^{1}K'^2(s)ds+O_p(h^4).
\end{aligned}$$

It follows that $\mathrm{MISE}\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\}$ have the following asymptotic expression

$$\int_{-C_{n_2}}^{C_{n_2}}E^*\{\widehat{f}'(t;\widetilde{\mu})-f_0(t)\}^2 \;=\; O_p(h^4),$$

since

$$\int_{-C_{n_2}}^{C_{n_2}}\{-f_0(t)\pi(t,\beta_0)+f_0(-t)\pi(-t,\beta_0)\}dt=0.$$

From condition (iv), $h = O_p(n^{-1/5})$ and we thus have

$$\frac{\sqrt{n}}{\epsilon}\int_{-C_{n_2}}^{C_{n_2}} E^*\left\{\widehat{f}(t;\widetilde{\mu}) - f_0(t)\right\}^2 dt = \frac{\sqrt{n}}{\epsilon}\text{MISE}\{\widehat{f}(t;\widetilde{\mu})|\widetilde{\mu}\} = o_p(1).$$

□

B8. Proof of Corollary 2

Approximating the derivative f_0' using numerical differentiation, we have

$$\begin{aligned}
& \lim_{n\to\infty} ||\widehat{f}'(x-\widetilde{\mu};\widetilde{\mu}) - f_0'(x-\mu_0)|| \\
= & \lim_{n\to\infty} ||\frac{\widehat{f}(x-\widetilde{\mu}+n^{-1/4};\widetilde{\mu}) - \widehat{f}(x-\widetilde{\mu}-n^{-1/4};\widetilde{\mu})}{2n^{-1/4}} - \frac{f_0(x-\mu_0+n^{-1/4}) - f_0(x-\mu_0-n^{-1/4})}{2n^{-1/4}}|| \\
= & \lim_{n\to\infty} \frac{||\widehat{f}(x-\widetilde{\mu}+n^{-1/4};\widetilde{\mu}) - \widehat{f}(x-\widetilde{\mu}-n^{-1/4};\widetilde{\mu}) - f_0(x-\mu_0+n^{-1/4}) + f_0(x-\mu_0-n^{-1/4})||}{2n^{-1/4}} \\
\leq & \lim_{n\to\infty} ||\widehat{f}(x-\widetilde{\mu}+n^{-1/4};\widetilde{\mu}) - f_0(x-\mu_0+n^{-1/4})||/(2n^{-1/4}) \\
& + \lim_{n\to\infty} ||\widehat{f}(x-\widetilde{\mu}-n^{-1/4};\widetilde{\mu}) - f_0(x-\mu_0-n^{-1/4})||/(2n^{-1/4}) = 0
\end{aligned}$$

with probability 1. The last eqaulity holds because $||\widehat{f}(x-\widetilde{\mu};\widetilde{\mu}) - f_0(x-\mu_0)|| = o_p(n^{-1/4})$ from Proposition 1. □

B9. Proof of Theorem 4

To avoid the complexity caused by various correlations, 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. Suppose that $\widetilde{\mu}$ and $\widetilde{\beta}$ are estimators constructed from the observations $X_{n_1+n_2+1}, \ldots, X_n$ and satisfies $\widetilde{\mu} - \mu = O_p(n_3^{-1/2})$ and $\widetilde{\beta} - \beta = O_p(n_3^{-1/2})$. Obviously, the estimate obtained from Step 1 satisfies these requirements.

We write the estimating equation as

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} S_{\theta,\text{eff}}\{X_i;\widehat{\theta},\widehat{f}(\cdot;\widetilde{\mu})\} \\ &= \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} S_{\theta,\text{eff}}(X_i;\theta_0,f_0) + \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{\partial S_{\theta,\text{eff}}\{X_i;\theta^*,\widehat{f}(\cdot;\widetilde{\mu})\}}{\partial\theta^T}(\widehat{\theta}-\theta_0) \\ &\quad + \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[S_{\theta,\text{eff}}\{X_i;\theta_0,\widehat{f}(\cdot;\widetilde{\mu})\} - S_{\theta,\text{eff}}(X_i;\theta_0,f_0)\right], \end{aligned}$$

where $\theta^* = \lambda\widehat{\theta} + (1-\lambda)\theta_0$ for $0 \le \lambda \le 1$. It is easy to see that

$$\begin{aligned} &\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{\partial S_{\theta,\text{eff}}\{X_i;\theta^*,\widehat{f}(\cdot;\widetilde{\mu})\}}{\partial\theta^T}(\widehat{\theta}-\theta_0) \\ &= E\left\{\frac{\partial S_{\theta,\text{eff}}(X;\theta_0,f_0)}{\partial\theta^T}\right\}\sqrt{n_1}(\widehat{\theta}-\theta_0) + o_p(1) \\ &= -E\{S_{\theta,\text{eff}}(X;\theta_0,f_0)^{\otimes 2}\}\sqrt{n_1}(\widehat{\theta}-\theta_0) + o_p(1), \end{aligned}$$

where the last equality is because $S_{\theta,\text{eff}}$ is the orthogonal projection of the score function to $\Lambda^{\perp}$.

Thus, to show the desired result, we only need to demonstrate that

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[S_{\theta,\text{eff}}\{X_i;\theta_0,\widehat{f}(\cdot;\widetilde{\mu})\} - S_{\theta,\text{eff}}(X_i;\theta_0,f_0)\right] = o_p(1),$$

or equivalently,

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left\{\frac{\widehat{f}'(X_i-\mu_0;\widetilde{\mu})}{\widehat{f}(X_i-\mu_0;\widetilde{\mu})} - \frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}2\pi_0(-X_i+\mu_0;\beta_0) = o_p(1). \tag{B.2}$$

We first point out that since $\widehat{f}(t;\widetilde{\mu})$ is an even function of t and $\widehat{f}'(t;\widetilde{\mu})$ is an odd function of t, hence

$$\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})} - \frac{f_0'(t)}{f_0(t)}\right\}2\pi_0(-t;\beta_0)2f_0(t)\pi_0(t;\beta_0)$$

is an odd function, hence

$$\begin{aligned} & E\left[\left\{\frac{\widehat{f}'(X_i-\mu_0;\widetilde{\mu})}{\widehat{f}(X_i-\mu_0;\widetilde{\mu})}-\frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}2\pi_0(-X_i+\mu_0;\beta_0)\right] \\ = \; & E\int\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{f_0'(t)}{f_0(t)}\right\}2\pi_0(-t;\beta_0)2f_0(t)\pi_0(t;\beta_0)dt \\ = \; & E(0)=0 \end{aligned}$$

for all $i=1,\ldots,n_1$.

The second moment of the left side of (C.2) is

$$\begin{aligned} & E\left(\left[\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left\{\frac{\widehat{f}'(X_i-\mu_0;\widetilde{\mu})}{\widehat{f}(X_i-\mu_0;\widetilde{\mu})}-\frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}2\pi_0(-X_i+\mu_0;\beta_0)\right]^2\right) \\ = \; & E\left[\left\{\frac{\widehat{f}'(X_i-\mu_0;\widetilde{\mu})}{\widehat{f}(X_i-\mu_0;\widetilde{\mu})}-\frac{f_0'(X_i-\mu_0)}{f_0(X_i-\mu_0)}\right\}^2 4\pi_0^2(-X_i+\mu_0;\beta_0)\right] \\ = \; & E\int\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{f_0'(t)}{f_0(t)}\right\}^2 4\pi_0^2(-t;\beta_0)2f_0(t)\pi_0(t;\beta_0)dt \\ \leq \; & 8E\int\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{f_0'(t)}{f_0(t)}\right\}^2 f_0(t)dt \\ \leq \; & 16E\int\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{\widehat{f}'(t;\mu_0)}{\widehat{f}(t;\mu_0)}\right\}^2 f_0(t)dt+16E\int\left\{\frac{\widehat{f}'(t;\mu_0)}{\widehat{f}(t;\mu_0)}-\frac{f_0'(t)}{f_0(t)}\right\}^2 f_0(t)dt. \end{aligned}$$

Using the delta method, the first term satisfies

$$\begin{aligned} & E\int\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{\widehat{f}'(t;\mu_0)}{\widehat{f}(t;\mu_0)}\right\}^2 f_0(t)dt \\ = \; & E\int E\left[\left\{\frac{\widehat{f}'(t;\widetilde{\mu})}{\widehat{f}(t;\widetilde{\mu})}-\frac{\widehat{f}'(t;\mu_0)}{\widehat{f}(t;\mu_0)}\right\}^2\Big|X_{n_1+1},\ldots,X_{n_1+n_2}\right]f_0(t)dt \\ = \; & E\{O_p(n_3^{-1})\}=o(1). \end{aligned}$$

The second term can be recognized as the mean integrated squared error of nonparametric estimation, hence standard analysis yields that it is of order $O\{h^4+$

$(n_2 h)^{-1}\} = o(1)$ for h satisfied condition (iv). Thus the second moment of the left side of (C.2) converges to zero as $n \to \infty$. From Serfling (2002, 1.2.3), (C.2) is indeed true.

The above result yields

$$\sqrt{n_1}(\widehat{\theta} - \theta_0) \to N\left(0, \left[E\{S_{\theta,\text{eff}}(X; \theta_0, f_0)^{\otimes 2}\}\right]^{-1}\right).$$

Since $n_1 = n - 2n^{1-\epsilon}$,

$$\sqrt{n}(\widehat{\theta} - \theta_0) - \sqrt{n_1}(\widehat{\theta} - \theta_0) = o_p(1).$$

We hence have $\sqrt{n}(\widehat{\theta} - \theta_0) \to N(0, V)$. □

B10. Proof of Theorem 5

To shorten the notation, we denote $w(t) = f_0(t)\pi(t; \beta_0)$. It follows that

$$\begin{aligned} w'(t) &= f_0'(t)\pi(t;\beta_0) + f_0(t)\pi'(t;\beta_0) \\ w''(t) &= f_0''(t)\pi(t;\beta_0) + 2f_0'(t)\pi'(t;\beta_0) + f_0(t)\pi''(t;\beta_0), \end{aligned}$$

It is easy to verify that

$$\begin{aligned} w(t) + w(-t) &= f_0(t) \\ w(t) - w(-t) &= 2f_0(t)\pi(t;\beta_0) - f_0(t), \\ w'(t) + w'(-t) &= 2f_0'(t)\pi(t;\beta_0) + 2f_0(t)\pi'(t;\beta_0) - f_0'(t), \\ w'(t) - w'(-t) &= f_0'(t), \\ w''(t) + w''(-t) &= f_0''(t), \end{aligned}$$

where $w'(-t) = w'(s)|_{s=-t}$ and $w''(-t) = w''(s)|_{s=-t}$. These results will be used repeatedly in the following calculation. To simplify the proof, we split the n obser-

vations 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 $\widetilde{\mu}$ and $\widetilde{\beta}$ are estimators constructed from the observations $X_{n_1+1}, \ldots, X_n$ and satisfy $\widetilde{\mu} - \mu = O_p(n_2^{-1/2})$ and $\widetilde{\beta} - \beta = O_p(n_2^{-1/2})$.

We first analyse the bias of $\widehat{f}$. We have

$$\begin{aligned}
\text{bias}\{\widehat{f}(t;\widetilde{\mu})\} &= E\{\widehat{f}(t;\widetilde{\mu})\} - f_0(t) \\
&= E\{\widehat{f}(t;\mu_0)\} - f_0(t) + O(n_2^{-1/2}) \\
&= E\left[\frac{1}{2n_1}\sum_{i=1}^{n_1}\frac{1}{h}\left\{K\left(\frac{X_i-\mu_0-t}{h}\right) + K\left(\frac{X_i-\mu_0+t}{h}\right)\right\}\right] - f_0(t) \\
&\quad +O(n_2^{-1/2}) \\
&= \frac{1}{2h}E\left\{K\left(\frac{X-\mu_0-t}{h}\right) + K\left(\frac{X-\mu_0+t}{h}\right)\right\} - f_0(t) + O(n_2^{-1/2}) \\
&= \int_{t+\mu_0-h}^{t+\mu_0+h}\frac{1}{h}K\left(\frac{x-\mu_0-t}{h}\right) f_0(x-\mu_0)\pi(x-\mu_0;\beta_0)dx \\
&\quad + \int_{\mu_0-t-h}^{\mu_0-t+h}\frac{1}{h}K\left(\frac{x-\mu_0+t}{h}\right) f_0(x-\mu_0)\pi(x-\mu_0;\beta_0)dx - f_0(t) \\
&\quad +O(n_2^{-1/2}) \\
&= \int_{-1}^{1}K(s)w(t+hs)ds + \int_{-1}^{1}K(s)w(hs-t)ds - f_0(t) + O(n_2^{-1/2}) \\
&= w(t) + \frac{w''(t)h^2}{2}\int_{-1}^{1}K(s)s^2ds + w(-t) + \frac{w''(-t)h^2}{2}\int_{-1}^{1}K(s)sds \\
&\quad -f_0(t) + o(h^2) + O(n_2^{-1/2}) \\
&= \frac{h^2}{2}f_0''(t)c_2 + o(h^2).
\end{aligned}$$

To analyse the variance, we have

$$\begin{aligned}
\text{var}\{\widehat{f}(t;\widetilde{\mu})\} &= \text{var}\{\widehat{f}(t;\mu_0) + \widehat{f}'_\mu(t;\mu_0)(\widetilde{\mu}-\mu_0)\} + O(n_2^{-1}) \\
&= \text{var}\{\widehat{f}(t;\mu_0)\} + 2\text{cov}\{\widehat{f}(t;\mu_0), \widehat{f}'_\mu(t;\mu_0)(\widetilde{\mu}-\mu_0)\} + O(n_2^{-1}).
\end{aligned}$$

The first term

$$\begin{aligned}\mathrm{var}\{\widehat{f}(t;\mu_0)\} &= \mathrm{var}\left[\frac{1}{2n_1}\sum_{i=1}^{n_1}\frac{1}{h}\left\{K\left(\frac{X_i-\mu_0-t}{h}\right)+K\left(\frac{X_i-\mu_0+t}{h}\right)\right\}\right] \\ &= \frac{1}{4n_1h^2}\mathrm{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)+K\left(\frac{X-\mu_0+t}{h}\right)\right\} \\ &= (4n_1h^2)^{-1}\mathrm{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)\right\}+(4n_1h^2)^{-1}\mathrm{var}\left\{K\left(\frac{X-\mu_0+t}{h}\right)\right\} \\ &\quad +(2n_1h^2)^{-1}\mathrm{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right),K\left(\frac{X-\mu_0+t}{h}\right)\right\}.\end{aligned}$$

We can easily obtain

$$\begin{aligned}(4n_1h^2)^{-1}\mathrm{var}\left\{K\left(\frac{X-\mu_0-t}{h}\right)\right\} &= (4n_1h)^{-1}\int K^2(s)2w(t+hs)ds \\ &= (2n_1h)^{-1}w(t)v_2+O(n_1^{-1}).\end{aligned}$$

Similarly, $(4n_1h^2)^{-1}\mathrm{var} = (2n_1h)^{-1}w(-t)v_2+O(n_1^{-1})$. The covariance term vanishes unless t satisfies $-h+|X-\mu_0|<t<h-|X-\mu_0|$. Thus, for $|t|\geq h$, the covariance term is zero. Otherwise, we have

$$\begin{aligned}&(2n_1h^2)^{-1}\mathrm{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right),K\left(\frac{X-\mu_0+t}{h}\right)\right\} \\ =\ &(n_1h)^{-1}\int_{\frac{|t|}{h}-1}^{1-\frac{|t|}{h}}K(s-t/h)K(s+t/h)w(hs)ds+O(n_1^{-1}).\end{aligned}$$

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

$$\begin{aligned}&\mathrm{var}\{\widehat{f}(t;\mu_0)\} \\ =\ &(n_1h)^{-1}\left\{f_0(t)v_2/2+I(|t|<h)\int_{\frac{|t|}{h}-1}^{1-\frac{|t|}{h}}K(s-t/h)K(s+t/h)w(hs)ds\right\} \\ &+O(n_1^{-1}).\end{aligned}$$

On the other hand,

$$
\begin{aligned}
& 2\mathrm{cov}\{\widehat{f}(t;\mu_0), \widehat{f}'_{\mu}(t;\mu_0)\} \\
= & \frac{-1}{2n_1h^3}\mathrm{cov}\left\{K\left(\frac{X-\mu_0-t}{h}\right)+K\left(\frac{X-\mu_0+t}{h}\right), K'\left(\frac{X-\mu_0-t}{h}\right)-K'\left(\frac{X-\mu_0+t}{h}\right)\right\} \\
= & \frac{-1}{2n_1h^3}E\left\{K\left(\frac{X-\mu_0-t}{h}\right)K'\left(\frac{X-\mu_0+t}{h}\right)-K\left(\frac{X-\mu_0+t}{h}\right)K'\left(\frac{X-\mu_0-t}{h}\right)\right\} \\
= & \frac{I(|t|<h)}{-n_1h^2}\int_{\frac{|t|}{h}-1}^{1-\frac{|t|}{h}}\left\{K(s-t/h)K'(s+t/h)-K(s+t/h)K'(s-t/h)\right\}w(hs)ds+O\{(n_1h)^{-1}\} \\
= & O(n_1^{-1}h^2).
\end{aligned}
$$

Because $\widetilde{\mu}-\mu_0$ has order $O_p(n_2^{-1/2})$, thus $\mathrm{var}\{\widehat{f}(t;\widetilde{\mu})\}$ is dominated by $\mathrm{var}\{\widehat{f}(t;\mu_0)\}$. Taking into account the relation between n_1, n_2 and n, we obtain the result in Theorem 5.

□

APPENDIX C

C1. Derivation of $\Lambda^{\perp}$

To prepare for the derivation of $\Lambda^{\perp}$, we first show that the nuisance tangent space of (4.1) 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, i.e. $\mathbf{u}(X-\mu) \in \Lambda$.

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

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

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

$$\mathbf{u}_1(z) = \frac{\mathbf{v}(z)w(z,\boldsymbol{\beta}) + \mathbf{v}(-z)w(-z,\boldsymbol{\beta})}{w(z,\boldsymbol{\beta}) + w(-z,\boldsymbol{\beta})}, \qquad \mathbf{u}(z) = \mathbf{u}_1(z) - E\{\mathbf{u}_1(X-\mu)\},$$

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

$$\begin{aligned} &E\{\mathbf{u}(X-\mu)\mathbf{v}^{\mathrm{T}}(X-\mu)\} \\ &= c(\boldsymbol{\beta}) \int_{0}^{\infty} \frac{\{\mathbf{v}(t)w(t,\boldsymbol{\beta}) + \mathbf{v}(-t)w(-t,\boldsymbol{\beta})\}^{\otimes 2} f_0(t)}{w(t,\boldsymbol{\beta}) + w(-t,\boldsymbol{\beta})}dt - \left[E\{\mathbf{v}(X-\mu)\}\right]^{\otimes 2}, \end{aligned}$$

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 $\Lambda^{\perp} \subset L$.

C2. Derivation of the efficient score $\mathbf{S}_{\text{eff}}$ for model (4.1)

Define

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

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

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

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

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

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

Then we have $\mathbf{S}_\beta = \mathbf{u}_2(x-\mu)+\mathbf{v}_2(x-\mu)$. 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

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

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 $\Pi(\mathbf{S}_{\beta} | \Lambda^{\perp}) = \mathbf{v}_2(x - \mu)$.

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

C3. Proof of Theorem 7

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

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

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

$$n^{1/2}(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = -n^{-1/2}\left[E\left\{\frac{\partial \mathbf{S}(X_i, \boldsymbol{\alpha}_0, f^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right\}\right]^{-1} \sum_{i=1}^{n} \mathbf{S}(X_i, \boldsymbol{\alpha}_0, f^*) + o_p(1). \quad \text{(C.1)}$$

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 7. The remaining upper-right block satisfies

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

where the last equality is because $\mathbf{S}_{\gamma}$ 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 (C.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 7. □

C4. Proof of Theorem 8

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 6, $\widetilde{\boldsymbol{\theta}}$ satisfies

$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n_2^{-1/2})$. To calculate bias, we have

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

Thus the bias is

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

To analyze the variance, we have

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

We can easily obtain

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

Similarly,

$$\frac{1}{n_1 h^2 w_1^2(t,\boldsymbol{\beta}_0)} \mathrm{var}\left\{K\left(\frac{X-\mu_0+t}{h}\right)\right\} = \frac{c(\boldsymbol{\beta}_0) v_2}{n_1 h w_1^2(t,\boldsymbol{\beta}_0)} f_0(t) w(-t,\boldsymbol{\beta}_0) + O(n_1^{-1}).$$

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

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

Obviously,

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

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

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

□

C5. Proof of Theorem 9

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 6, we have $\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 = O_p(n_3^{-1/2})$.

We write the estimating equation as

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

where $\boldsymbol{\theta}^* = \Lambda\widehat{\boldsymbol{\theta}} + (1-\Lambda)\boldsymbol{\theta}_0$ for $0 \leq \Lambda \leq 1$. 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 6 and 8 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

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

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

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

because the integrand is an odd function. Consider the second moment of the left

side of (C.2). We have

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

where we used

$$
\frac{w(t,\boldsymbol{\beta}_0)w(-t,\boldsymbol{\beta}_0)}{w_1(t,\boldsymbol{\beta}_0)}\leq\frac{1}{2}\quad\text{and}\quad\frac{w(-t,\boldsymbol{\beta}_0)}{w_1(t,\boldsymbol{\beta}_0)}\leq 1.
$$

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 8. Thus the second moment of the left side of (C.2) converges to zero as $n\to\infty$. From Serfling (2002, 1.2.3), (C.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

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

□

VITA

Mi Jeong Kim was born in 1981. She double-majored in Statistics and Mathematics with a minor in Economics, and obtained B.S. degrees from Ewha University in 2004. She received her M.S. in Statistics from Ewha University in 2006. She received her Ph.D. in Statistics from Texas A&M University in August 2012. Her research interests include Semiparametrics, Skewed distributions and Time series.

She may be reached at:

Department of Statistics

Texas A&M University

3143 TAMU

College Station, TX 77843-3143.

Her web page URL is http://www.stat.tamu.edu/∼mjkim

and her email address is mjkim@stat.tamu.edu.