SEMIPARAMETRIC ESTIMATION AND INFERENCE WITH MIS-MEASURED, CORRELATED OR MIXED OBSERVATIONS, AND THE APPLICATION IN ECOLOGY, MEDICINE AND NEUROLOGY

A Dissertation

by

KUN XU

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Yanyuan Ma
Committee Members,	Raymond J. Carroll
	Samiran Sinha
	J. Maurice Rojas
Department Head,	Simon J. Sheather

December 2013

Major Subject: Statistics

ABSTRACT

The dissertation considers semiparametric regression models inspired by statistical problems in ecological, medical and neurological studies. In those models, the interest is usually on the estimation of a set of finite parameters with difficulties of handling some unknown distribution functions or some other unknown structures. Developing novel semiparametric treatments and deriving a class of consistent and efficient estimators can not only provide us with better inferences, but also a general framework in those studies.

In capture-recapture models for closed populations, the goal is to estimate the abundance of population. When multiple error-prone measurements of a covariate are available, we discover that no suitable complete and sufficient statistic exists due to the identity between the number of captures and the number of measurements. Hence the existing treatment utilizing such statistic no longer apply. Our investigation indicates that the familiar strategy of generalized method of moments can only resolve the issue with high capture probabilities. Further complexity includes the loss of the surrogacy assumption, commonly assumed in most measurement error problems. We devise a novel semiparametric treatment to overcome those difficulties. Simulation studies and real data analysis show good performance of our method.

In HIV research, we study errors-in-variables problems when the response is binary and instrumental variables are available. We construct consistent estimators through taking advantage of the prediction relation between the unobservable variables and the instruments. The asymptotic properties of the new estimator are established, and illustrated through simulation studies. We also demonstrate that the method can be readily generalized to generalized linear models and beyond. The

usefulness of the method is illustrated through a real data example.

Lastly, we nonparametrically estimate distribution functions for multiple populations in kin-cohort studies. The data is mixed and known to belong to a specific population with certain probabilities. Some of the observations can be further correlated, and are subject to censoring. We estimate the distributions in an optimal way through using the optimal base estimators and then combine the estimators optimally as well. The optimality implies both estimation consistency and minimum estimation variability. One obvious advantage is that our estimator does not assume any parametric forms of the distributions, and does not require to know or to model the potential correlation structure. Analysis on the Huntington's disease data is performed to illustrate the effectiveness of the method.

To my parents and fiancee, for their endless love, support and encouragement.

ACKNOWLEDGMENTS

I am very lucky to have Dr. Yanyuan Ma as my advisor during the period of doctoral training. She sparked my interest in research with her wisdom and enthusiasm. She supervised me with her smile, patience and deep understanding of statistical methodology. I am very grateful to her for her guidance through every stage of research and memorable discussions by varieties of forms: words, emails and revision suggestions in manuscripts. Thank you! My advisor and best friend!

I thank my committee members, Dr. Raymond J. Carroll for generous advice on research, paper, presentation and many more; Dr. Samiran Sinha for encouragement and support, especially when I feel desperate about research; and Dr. J. Maurice Rojas for his willingness to serve on my committee.

The faculties in Department of Statistics, Texas A&M University taught me a lot inside and outside classroom. I am thankful to Dr. Subba Rao, Dr. Pourahmadi, Dr. Marc Genton, Dr. Huang Jianhua, Dr. Liang Faming, Dr. Hart, Dr. Jun Mikyoung, Dr. Lahiri, Dr. Mueller-Harknett Ursula, Dr. Sherman, Dr. Wang Suojin, Dr. Dahm Fred, Dr. Jones Edward and Dr. Wehrly. They are kind and willing to offer me their best help. I obtain enormous good suggestions on life, job and research. I also want to thank Dr. Longnecker for his suggestions on career and help of administrative issues.

I owe thanks to my friends, Yang Xu, Xingheng, Ranye, Ganggang, Qifan, Xiaoqing, Rubin, Bohai, Tanya, Nan Zhang, Abhra, Bryce, Stephanie, Debkumar and Robyn. You make my life colorful in College Station. Thank you so much for accompanying me all the way along! Thank you for giving me so much happiness!

One special thank you should go to my mom. She gives me her love and support

for so many years and never give up on me. Another thank you should go to my father. I still remember the days he took me to school on his old bicycle. He did his best to earn study opportunities for me. I cannot obtain the doctoral degree without them. I would also like to thank my darling fiancee, Yingyuan, for her confidence in my ability.

Lastly, I want to thank all my friends and teachers who have ever helped me.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Semiparametric regression models naturally arise in statistical problems where data are generated through a class of distributions containing both parameters of interest and nuisance parameters. An intuitive example is a logistic regression where covariates are measured with error. The coefficients of regressors are the parameters to be estimated, while the unknown distributions of mis-measured covariates are the the nuisance parameters. The dissertation is then dedicated to design novel statistical models to efficiently and consistently estimate parameters of interest with the presence of nuisance parameters, which in most occasions, are from an infinite-dimensional space. Since the nuisance parameter space is left completely unspecified, the estimators from semiparametric regression models will be more general and robust.

The development of semiparametric regression models dates back to Newey (1990). The modern treatment begins with the discussion in Bickel, Klaassen, Ritov & Wellner (1993) and Tsiatis (2006). Later Tsiatis & Ma (2004) introduces semiparametric regression models into measurement error models. The semiparametric theory considers independently, identically distributed random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a class of distributions

$$\{p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

where the parameter can be written as $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$. Suppose $\boldsymbol{\beta}$ is the parameter of interest and $\boldsymbol{\eta}$ is the nuisance parameter. Denote the true parameters by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\eta}_0^{\mathrm{T}})^{\mathrm{T}}$. In order to find a consistent and hopefully more efficient estimator $\widehat{\boldsymbol{\beta}}_n$, we

seek for influence function $\varphi(\mathbf{X})$ such that $E\{\varphi(\mathbf{X})\} = \mathbf{0}$, $E\{\varphi(\mathbf{X})\varphi(\mathbf{X})^{\mathrm{T}}\}$ is finite and positive-definite, and that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \varphi(\mathbf{X}_i) + o_p(1). \tag{1.1}$$

If we further put regular condition on $\widehat{\boldsymbol{\beta}}_n$, we call it regular and asymptotically linear (RAL) estimator, see Tsiatis (2006).

Due to equation (1.1), the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ can be determined by its corresponding influence function $\varphi(\mathbf{X})$. Considering a Hilbert space $\mathcal{H}$ with the covariance inner product, the semiparametric method demonstrates that the most efficient influence function is an element of the orthogonal complement of the nuisance tangent space from the geometric view. Here the covariance inner product is represented as $E(h_1^{\mathrm{T}} h_2)$ for any two elements $h_1, h_2 \in \mathcal{H}$. To this end, we must identify the nuisance tangent space, construct its orthogonal complement in $\mathcal{H}$ and obtain efficient influence functions for semiparametric regression models.

This semiparametric method yields optimal estimators for a wide range of statistical models. However, other methodologies can also be developed. For example, Generalized Method of Moments (GMM) (Hansen (1982)) offers an optimal way to combine estimating equations. Bootstrap (Efron (1979)) can nonparametrically estimate statistics. Those applications are good supplements to semiparametric regression models.

In my dissertation, I investigate three different kinds of semiparametric regression models in the literature of a capture-recapture model in ecological study, an instrumental variable regression model in medical study and a censored mixed correlated data in neurological study. I apply the semiparametric methodology to the first two problems and derive a class of consistent and efficient semiparametric estimators. I

design a resampled and bootstrapped method and obtain optimal estimators for the neurological study.

In section 2, I develop two methods, namely GMM and semiparametric method, to estimate the abundance of animals in capture-recapture experiments. Until now there are no consistent estimators in the literature to take full advantage of multiple measurements of covariates, subject to measurement errors. Besides, the distributions of covariates are left completely unspecified. Both methods yield consistent and robust estimators and bring on efficiency gain. In addition, the advance of GMM method is in scenarios where biologists believe a large portion of animals can be captured. The semiparametric method offers a guideline to solve a class of measurement error models where surrogacy assumption breaks down. Two simulations are conducted to validate the proposed methods and compared to a conditional score method (Hwang, Huang & Wang (2007)) which ignore additional information from multiple measurements of covariates. In real data analysis, the semiparametric method outperforms the conditional score method.

In section 3, I derive a class of consistent, asymptotically normal estimators for generalized regression models where there are errors in variables. So far, there are limited discussions of the use of instrumental variables in the literature. My semiparametric method is the first to solve the problem without any distribution assumptions on both unobserved covariates and measurement errors. In addition, the proposed method is robust and general. The estimation efficiency will not be lost because of a smart configuration of the prediction relationship for unobserved covariates using instruments. Two simulations and a real data analysis are used to show the satisfactory performance.

In section 4, I investigate a problem that requires immediate treatment in kin-cohort studies for diseases like Huntington's disease. The data are mixed with differ-

ent probabilities calculated by Mendalian law for the populations. The facts that data are subject to censoring and correlated increase the complexity of the problem. The parameters of interest are the distribution functions for multiple populations, while the nuisance parameter space could be the space of all family correlation matrices. I nonparametrically estimate the distribution functions for this type of semiparametric models. In particular, I use only one member per family to form base estimator. Afterwards, I devise an optimal way to synthesize a new estimator to take advantage of multiple members in families. In the section, I demonstrate two methods in detail with the same modeling strategy and show their equivalence in asymptotic aspect. These methods are novel, straightforward and flexible. Simulation studies are performed and a real data is illustrated.

In section 5, I summarize the research work of semiparametric regression models in ecological, medical and neurological studies with discoveries, limitations and future applications. All the technical proofs are in section 6, 7 and 8.

2. EFFECTIVE USE OF MULTIPLE ERROR-PRONE COVARIATE MEASUREMENTS IN CAPTURE-RECAPTURE MODELS

2.1 Introduction

Capture-recapture models are widely used to describe the abundance of a species of interest. Through modeling the probability of different numbers of captures of a single animal as a function of the associated covariates, it enables to use the observed covariate and capture information to infer the total population size. Here we work in the closed population framework. This means that there is no population flow such as mortality and immigration occurring during the experiment, hence the population size does not vary. Due to various reasons associated with the capture activity, covariate measurements are almost always prone to error. Hwang, Huang & Wang (2007) studied the effect of covariate measurement error on estimating the population size in capture-recapture models, and established that ignoring the covariate measurement error leads to estimation bias. They further proposed an effective method to correct this bias through accounting for the measurement error structure.

One main feature of a capture-recapture model is that it incorporates the situation that an individual animal is captured multiple times. It is not uncommon that at each capture event, the same covariates are measured, especially if these covariates are difficult to measure precisely or their values could fluctuate. If we believe these are different measurements of one underlying true covariate that affects the capture probability, it is then quite natural to use the average of these measurements in the capture-recapture model. If the measures are performed systematically, then one would think that the average of more measurements is more precise than

that of fewer measurements. This implies different measurement error variabilities for animals that are captured different times. However the multiple measurement information is ignored in Hwang, Huang & Wang (2007), where they assume a single covariate measurement is available, and this covariate measurement has a common error variance across different animals regardless of their different capture times. Huggins & Hwang (2009) realized the advantage of multiple measurements and successfully utilized it. However, to circumvent the inherent difficulty, they made an additional normality assumption on the unobservable covariates. Hence the problem becomes completely parametric and it is no longer in the functional measurement error model framework.

Intuitively, if one uses the average measurements but still assumes a common error variability, it is a model mis-specification issue and bias will occur in the final estimation. Conceptually, the reason underneath the estimation bias is not different from that of the estimation bias if we ignore the measurement error completely. If one ignores the multiple measurements and incorporates only one of these measurements, then not all the information is taken into account hence it implies a potential estimation efficiency loss. It is our goal to take into account the multiple measurements in the estimation procedure and retain consistency, while at the same time improve estimation efficiency.

In the capture-recapture literature, the capture probability is typically assumed to relate to covariates through a linear logistic model (Huggins (1989)), see Pollock (2002) for a comprehensive review of this topic. To increase model flexibility, extension to partially linear structure was studied in Huggins (2006), while Hwang & Huggins (2007) further incorporated categorical variables. When covariates are measured with error, Wang (2000) proposed a refined regression calibration estimator while Hwang & Huang (2003) proposed a conditional likelihood based method to

estimate the population size. A conditional score based method was later proposed by Hwang, Huang & Wang (2007), and Huggins & Hwang (2009) further extended the method to handle unknown measurement error variance. Our contribution in this work is to allow unknown, unequal measurement error variance that depends on the capture times, and construct consistent and efficient estimators that benefit from the special error properties.

The rest of the section is organized as the following. In section 2.2, we investigate a generalized method of moments (GMM) procedure. In section 2.3, we propose an effective way of using multiple measurements based on semiparametrics. Numerical experiments are conducted on both simulated examples in section 2.4 and on a capture-recapture data of the bird population in section 2.5. We conclude the section with a discussion in section 2.6 and relegate all the technical derivations and proofs to appendix.

2.2 Generalized Method of Moments Procedure

In the capture-recapture model we consider here, we use N to denote the total population size of the animals under study. The study interest lies in estimating and making inference about N. Let J be the total number of capture occasions. Using i to index the distinct animals and j to index the capture occasions, we denote the random event of the ith animal being captured on the jth occasion as Y_{ij}, with $Y_{ij} = 1$ for capture and $Y_{ij} = 0$ otherwise. Here $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Assume there are a total of D distinct animals captured at least once. For convenience, we label these animals from 1 to D and the animals never captured from $D + 1$ to N. A widely used model to describe the probability mass function of a binary outcome is

the logistic regression model

$$\text{logit}\{\text{pr}(Y_{ij} = 1 \mid \mathbf{X}_i = \mathbf{x}_i)\} = \alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

which relates the capture probability of an animal on one occasion to its covariates. This model is often used to describe the relation between capture probability and an animal's characteristic, see for example Pollock, Hines & Nichols (1984), Huggins (1989), Alho (1990), Huggins (1991) and Pledger (2000). Let Y_i be the total number of captures of the ith animal. Obviously, $Y_i = \sum_{j=1}^{J} Y_{ij}$. Under the assumptions that for each animal, conditional on its covariates, the different captures are independent of each other, we have

$$\text{pr}(Y_i = y_i \mid \mathbf{X}_i = \mathbf{x}_i) = \binom{J}{y_i} \frac{\exp\{(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})y_i\}}{\{1 + \exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^J}.$$

The above model is the logistic based capture-recapture model.

Under the situation that the covariates $\mathbf{X}_i$'s are not directly observed, alternative information is usually collected. We consider the practical situation that at each capture of an animal, its covariate is measured, subject to measurement error. Let $\mathbf{W}_{ij}$ be the measurement of $\mathbf{X}_i$ at the jth occasion if $Y_{ij} = 1$. We assume $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, where $\mathbf{U}_{ij}$ is a random measurement error, and is typically assumed to have a normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$. Our goal is to estimate N and make inference using the observed data $(Y_{ij}, Y_{ij}\mathbf{W}_{ij}), i = 1, \ldots, N, j = 1, \ldots, J$.

We would like to point out that if only a single measurement $\mathbf{W}_i$ is available in place of $\mathbf{X}_i$, where $\mathbf{W}_i$ has the same relation to $\mathbf{X}_i$ as $\mathbf{W}_{ij}$'s above, elegant results have been established in the literature, see Hwang, Huang & Wang (2007). The cen-

tral consideration there is that if we treat $(Y_i, \mathbf{W}_i)$ to be fixed and the parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ to be known in the joint probability density function of $(Y_i, \mathbf{W}_i)$ given $\mathbf{X}_i$, then $\Delta_i = \mathbf{W}_i + Y_i \boldsymbol{\Sigma} \boldsymbol{\beta}$ is a complete sufficient statistic for $\mathbf{X}_i$. Thus, conditional on Δ_i, because of the sufficiency, $\mathbf{Y}_i$ does not depend on $\mathbf{X}_i$. Further taking advantage of the completeness, one can construct an estimating equation based on $(\mathbf{Y}_i, \Delta_i)$ alone. See Ma & Tsiatis (2006a) for the details on how the sufficiency and completeness contribute to the construction of the estimator. It is thus tempting to form the average $\overline{\mathbf{W}}_i = (\sum_{j=1}^{J} Y_{ij} \mathbf{W}_{i,j})/Y_i$ and use it in the place of $\mathbf{W}_i$. Unfortunately, this is no longer a valid practice. The difficulty is not only caused by the different measurement error variances across different animals, which certainly needs attention. A more fundamental difficulty arises in this approach because the method by Hwang, Huang & Wang (2007) critically relies on the existence of a sufficient and complete "statistic" $\Delta_i = \mathbf{W}_i + Y_i \boldsymbol{\Sigma} \boldsymbol{\beta}$, while under the replacement, the same quantity would equal to $\overline{\mathbf{W}}_i + \boldsymbol{\Sigma} \boldsymbol{\beta}$ and it is no longer a sufficient or complete statistic.

An alternative obvious attempt of taking advantage of the situation is to combine the procedures in Hwang, Huang & Wang (2007) performed on each individual $\mathbf{W}_{ij}$. To this end, we resort to the GMM (Hansen (1982)) approach. Our consideration is the following. We first consider making use of the first measurement of each animal that is captured at least once, forming the complete sufficient statistic with the first measurement. This provides the first set of estimating equations. We then consider making use of the second measurement of each animal that is captured at least twice, forming the complete sufficient statistic with the second measurement. This provides the second set of estimating equations. We continue this process and obtain a maximum of J sets of estimating equations in total. We then use GMM to take advantage of all these equations. Specifically, let $\mathcal{C}_{ik}$ denote the event $Y_i \geq k$ for $k = 1, \ldots, J$. Thus, $I(\mathcal{C}_{ik}) = 1$ if the ith animal is captured at least k times,

and $I(\mathcal{C}_{ik}) = 0$ otherwise. For the ith animal with Y_i total captures, we denote its Y_i available measurements $\mathbf{W}_{i(l)}$, $l = 1, \ldots, Y_i$. Thus, the lth complete sufficient statistic is defined by $\Delta_{i(l)} = \mathbf{W}_{i(l)} + Y_i \boldsymbol{\Sigma\beta}$, $l = 1, \ldots, Y_i$.

Using the above notation, the kth set of estimating equations can be written as

$$\begin{aligned} & \sum_{i=1}^{N} \mathbf{g}_k(Y_i, \Delta_{i(k)}, \alpha, \boldsymbol{\beta}) \\ = & \sum_{i=1}^{N} I(\mathcal{C}_{ik}) \begin{bmatrix} Y_i - E(Y_i \mid \Delta_{i(k)}, \mathcal{C}_{ik}) \\ \left\{\Delta_{i(k)} - E(Y_i \mid \Delta_{i(k)}, \mathcal{C}_{ik})\boldsymbol{\Sigma\beta}\right\} \left\{Y_i - E(Y_i \mid \Delta_{i(k)}, \mathcal{C}_{ik})\right\} \end{bmatrix} = 0, \end{aligned}$$

where $k = 1, \ldots, K \leq J$. We emphasize here that when $k = 1$, the estimating equation is identical to the proposal in Hwang, Huang & Wang (2007). Here we use K to denote the maximum k value where there are still data available to form the estimating equation, i.e. the largest possible k such that $\max_i I(\mathcal{C}_{ik}) = 1$. We now combine these K sets of equations via GMM. Specifically, Let $\mathbf{O}_i = (Y_{i1}, Y_{i1}\mathbf{W}_{i1}^{\mathrm{T}}, \ldots, Y_{iJ}, Y_{iJ}\mathbf{W}_{iJ}^{\mathrm{T}})^{\mathrm{T}}$ be the observations related to the ith animal, let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, write

$$\mathbf{g}(\mathbf{O}_i, \boldsymbol{\theta}) = \begin{pmatrix} \mathbf{g}_1(Y_i, \Delta_{i(1)}, \boldsymbol{\theta}) \\ \mathbf{g}_2(Y_i, \Delta_{i(2)}, \boldsymbol{\theta}) \\ \vdots \\ \mathbf{g}_K(Y_i, \Delta_{i(K)}, \boldsymbol{\theta}) \end{pmatrix},$$

and obtain the estimator of $\boldsymbol{\theta}$ through minimizing

$$\left\{\sum_{i=1}^{N} \mathbf{g}(\mathbf{O}_i, \boldsymbol{\theta})\right\}^{\mathrm{T}} \left\{\sum_{i=1}^{N} \mathbf{g}(\mathbf{O}_i, \boldsymbol{\theta})\mathbf{g}^{\mathrm{T}}(\mathbf{O}_i, \boldsymbol{\theta})\right\}^{-1} \left\{\sum_{i=1}^{N} \mathbf{g}(\mathbf{O}_i, \boldsymbol{\theta})\right\}.$$

It is well known that the GMM estimator provides the optimal combination of the es-

timating equations in terms of the estimation efficiency, and the resulting estimator has the usual root-n consistency and asymptotic normality. Here, estimation efficiency is measured by the inverse of the variance of an estimator. A smaller variance results in larger efficiency, hence indicates a more efficient estimator. Specifically, the above estimation procedure yields $\widehat{\boldsymbol{\theta}}$ that is consistent, and has the asymptotic variance

$$N^{-1}\left(E\left\{\frac{\partial \mathbf{g}^{\mathrm{T}}(\mathbf{O}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}\left[E\left\{\mathbf{g}(\mathbf{O}_i,\boldsymbol{\theta})\mathbf{g}^{\mathrm{T}}(\mathbf{O}_i,\boldsymbol{\theta})\right\}\right]^{-1}E\left\{\frac{\partial \mathbf{g}(\mathbf{O}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}\right)^{-1}.$$

The above GMM uses each measurement $\mathbf{W}_{i(k)}$ separately. Alternatively, as suggest by a referee, one can also consider the averaged measurement $k^{-1}\sum_{l=1}^{k}\mathbf{W}_{i(l)}$ in the construction, and form estimating equation using all animals captured at least k times. Since the averaged quantity does not depend on Y_i, it will result in a sufficient complete statistic. One obvious advantage of such a configuration is that it can stabilize the estimation procedure.

To take advantage of the multiple measurements fully, it is tempting to use the average of all the Y_i measurements. However, this operation will lead to dependence between the response variable Y_i and the averaged measurement $Y_i^{-1}\sum_{l=1}^{Y_i}\mathbf{W}_{i(l)}$. In other words, the measurement error problem is now differential hence it requires a more careful investigation, as is discussed in the next section.

2.3 Semiparametric Method

We consider the observed animals, which are the ones captured at least once. Modeling the probability of the observed animal being captured y times given that

it is captured at least once, we have

$$\begin{aligned} & \mathrm{pr}(Y_i = y_i \mid \mathbf{X}_i = \mathbf{x}_i, \mathcal{C}_{i1}) \\ = \; & \binom{J}{y_i} \frac{\exp\{(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})y_i\}}{\{1+\exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^J} \left[1 - \left\{\frac{1}{1+\exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right\}^J\right]^{-1}, \end{aligned} \tag{2.1}$$

where $y_i = 1, \ldots J$. In addition, The averaged measurement $\overline{\mathbf{W}}_i$ satisfies $\overline{\mathbf{W}}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is normally distributed with mean zero and variance-covariance $\boldsymbol{\Sigma}/Y_i$, i.e. $\mathbf{U}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}/Y_i)$. Of course, $\overline{\mathbf{W}}_i$ is well defined only if $y_i \neq 0$. When $y_i = 0$, we can set $\overline{\mathbf{W}}_i$ to 0 and we will see that it has no effect on our estimation. In the situation that $\boldsymbol{\Sigma}$ is not known, we can easily estimate it by forming differences of the repeated observations, say $\mathbf{W}_{i(1)} - \mathbf{W}_{i(2)}$, and calculate the sample variance-covariance matrix, see Hall & Ma (2007) for details. Thus, for the rest of the article, unless we specifically point out otherwise, we assume $\boldsymbol{\Sigma}$ is known.

The variance expression of $\mathbf{U}_i$ indicates that Y_i and $\overline{\mathbf{W}}_i$ are no longer independent conditional on $\mathbf{X}_i$. This means that the standard surrogacy assumption in the measurement error literature is violated in this context. Let p be the dimension of $\mathbf{X}_i$, then the joint distribution of Y_i and $\overline{\mathbf{W}}_i$ conditional on $(\mathbf{X}_i, \mathcal{C}_{i1})$ is

$$\begin{aligned} & f_{Y_i, \overline{\mathbf{W}}_i | \mathbf{X}_i, \mathcal{C}_{i1}}(y_i, \overline{\mathbf{w}}_i \mid \mathbf{x}_i, \mathcal{C}_{i1}) \\ = \; & \mathrm{pr}(Y_i = y_i \mid \mathbf{X}_i = \mathbf{x}_i, \mathcal{C}_{i1}) f_{\overline{\mathbf{W}}_i | Y_i, \mathbf{X}_i, \mathcal{C}_{i1}}(\overline{\mathbf{w}}_i \mid y_i, \mathbf{x}_i, \mathcal{C}_{i1}) \\ = \; & \binom{J}{y_i} \frac{\exp\{(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})y_i\}}{\{1+\exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^J} \left[1 - \left\{\frac{1}{1+\exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right\}^J\right]^{-1} \\ & (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}/y_i|^{-\frac{1}{2}} \exp\left\{-\frac{y_i}{2}(\overline{\mathbf{w}}_i - \mathbf{x}_i)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{w}}_i - \mathbf{x}_i)\right\} \\ = \; & \binom{J}{y_i} (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}/y_i|^{-\frac{1}{2}} \frac{\exp(y_i\alpha - y_i \overline{\mathbf{w}}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \overline{\mathbf{w}}_i/2)}{\{1+\exp(\alpha + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^J - 1} \\ & \exp\left\{\mathbf{x}_i^{\mathrm{T}}(y_i\boldsymbol{\beta} + y_i \boldsymbol{\Sigma}^{-1}\overline{\mathbf{w}}_i) - y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i/2\right\}. \end{aligned}$$

Because it is in the form of the exponential family, the minimal complete statistic is $(Y_i\boldsymbol{\beta} + Y_i\boldsymbol{\Sigma}^{-1}\overline{\mathbf{W}}_i, Y_i)$, or equivalently $(\overline{\mathbf{W}}_i, Y_i)$. This statistic certainly does not help simplifying the problem.

The lack of both the surrogacy property and a suitable sufficient and complete statistic requires a new way of treating the problem. Our way is through casting the problem in a semiparametric framework. In the semiparametric derivation, the distribution of $\mathbf{X}_i$ has to be taken into account and we treat it as a nuisance parameter with infinite dimension. However, we avoid estimating the distribution of $\mathbf{X}_i$. Instead, we calculate its corresponding tangent space formed by the mean squared closure of the set of all score functions of its parametric submodels. The orthogonal complement of the tangent space subsequently contains the elements for building consistent estimating equations. This type of approach originates from Bickel, Klaassen, Ritov & Wellner (1993), and a nice explanation and more elaborated discussions about such calculations can be found in Tsiatis (2006).

To this end, the joint distribution of the observed $(Y_i, \overline{\mathbf{W}}_i)$ is

$$\begin{aligned} & f_{Y_i,\overline{\mathbf{W}}_i|\mathcal{C}_{i1}}(y_i, \overline{\mathbf{w}}_i \mid \mathcal{C}_{i1}) \\ = & \int f_{\overline{\mathbf{W}}_i|Y_i,\mathbf{X}_i,\mathcal{C}_{i1}}(\overline{\mathbf{w}}_i \mid y_i, \mathbf{x}_i, \mathcal{C}_{i1}) f_{Y_i|\mathbf{X}_i,\mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i, \mathcal{C}_{i1}) f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1}) d\mu(\mathbf{x}_i), \end{aligned}$$

where $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$ is the unknown probability density function of $\mathbf{X}_i$ conditional on $\mathcal{C}_{i1}$, while the conditional distribution $f_{Y_i|\mathbf{X}_i,\mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i, \mathcal{C}_{i1})$ is completely determined by $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$.

Our goal is to construct the estimating equation based on the conditional joint distribution of $(Y_i, \overline{\mathbf{W}}_i)$ through calculating the efficient score function. The process contains three steps. The first step is to calculate the score function with respect to

$\boldsymbol{\theta}$. We have

$$\mathbf{S}_\theta(Y_i, \overline{\mathbf{W}}_i) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f_{Y_i, \overline{\mathbf{W}}_i | \mathcal{C}_{i1}}(y_i, \overline{\mathbf{w}}_i \mid \mathcal{C}_{i1}) \right\} = E\left\{ \mathbf{S}_\theta^F(Y_i, \mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1} \right\},$$

where $\mathbf{S}_\theta^F(Y_i, \mathbf{X}_i) \equiv \partial \log f_{Y_i|\mathbf{X}_i, \mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i, \mathcal{C}_{i1}) / \partial \boldsymbol{\theta}$ and $f_{Y_i|\mathbf{X}_i, \mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i, \mathcal{C}_{i1})$ is given in (2.1). The second step is to find the nuisance tangent space Λ and its orthogonal complement with respect to the infinite dimensional parameter $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$. We start with considering the parametric submodels which lie in the family of the unknown conditional distributions and contain the true distribution. For each of the parametric submodels, the score function with respect to the nuisance parameter can be calculated directly. Then we proceed to take the mean squared closure of all these score functions corresponding to the different submodels to obtain Λ. Detailed calculation in appendix yields

$$\Lambda = \left[E\left\{ \mathbf{h}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1} \right\} : \; E\left\{ \mathbf{h}(\mathbf{X}_i) \mid \mathcal{C}_{i1} \right\} = \mathbf{0} \right].$$

Here $\mathbf{h}(\mathbf{X}_i)$ is a random function in the Hilbert space $\mathcal{H}$. The corresponding orthogonal complement of Λ, denoted $\Lambda^\perp$ is then given by

$$\Lambda^\perp = \left[\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) : E\left\{ \mathbf{g}(Y_i, \overline{\mathbf{W}}_i) \mid \mathbf{X}_i, \mathcal{C}_{i1} \right\} = \mathbf{0} \right].$$

The third step is to project the score function $\mathbf{S}_\theta(Y_i, \overline{\mathbf{W}}_i)$ to $\Lambda^\perp$ to obtain the efficient score $\mathbf{S}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i)$. Any random function $\mathbf{g}(Y_i, \overline{\mathbf{W}}_i)$ in $\Lambda^\perp$ must satisfy that its conditional expectation on $\mathbf{X}_i$ and $\mathcal{C}_{i1}$ is zero. On the other hand, any random function in Λ must be able to be expressed as $E\left\{ \mathbf{a}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1} \right\}$. Thus, if we

identify a function $\mathbf{a}(\mathbf{X}_i)$ such that

$$E\left[\mathbf{S}_\theta(Y_i, \overline{\mathbf{W}}_i) - E\left\{\mathbf{a}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\right\} \mid \mathbf{X}_i, \mathcal{C}_{i1}\right] = \mathbf{0}, \tag{2.2}$$

then we have found the efficient score

$$\begin{aligned} \mathbf{S}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i) &= \mathbf{S}_\theta(Y_i, \overline{\mathbf{W}}_i) - E\big\{\mathbf{a}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\big\} \\ &= E\big\{\mathbf{S}_\theta^F(Y_i, \mathbf{X}_i) - \mathbf{a}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\big\}. \end{aligned}$$

The conditional expectation involved in the calculation of the efficient score relies on the unknown distribution $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$. In practice, we propose a candidate distribution $f^*_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$, and carry out the estimation procedure under $f^*_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$. We denote the resulting efficient score function $\mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i)$. Because our procedure in obtaining $\mathbf{a}(\mathbf{X}_i)$ from (2.2) calculated under $f^*_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$ ensures that $E\{\mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i)\} = \mathbf{0}$ regardless $f^*_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$ equals the true conditional distribution or not, we will still have a consistent estimator even if the candidate model is not the same as the true model. This is usually referred to as a locally efficient estimator. To solve for $\mathbf{a}(\mathbf{X}_i)$, we use the similar computational technique as in Tsiatis & Ma (2004). Although the statistical derivation and problem context is very different, the integral equation (2.2) shares similar mathematical properties as the integral equation there. The final estimating equation is

$$\sum_{i=1}^{D} \mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{2.3}$$

The summation in (2.3) indicates that only the animals captured at least once contribute to the estimation. Numerically, (2.3) is solved through the Newton-Raphson

algorithm. In practice, especially when D is small, there can be multiple roots to (2.3). In such case, empirical rule and practical knowledge is typically used to select a suitable root. The estimator from solving (2.3) has the asymptotic property described in Theorem 1.

Theorem 1. *Let $\widehat{\boldsymbol{\theta}}$ solve the estimating equation (2.3). Assume the covariance Σ to be known for the measurement error. Then,*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \mathbf{V})$$

in distribution when $N \to \infty$. Here $\mathbf{V} = \mathbf{A}^{-1}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\left\{\mathbf{A}^{-1}(\boldsymbol{\theta})\right\}^{\mathrm{T}}$,

$$\begin{aligned}
\mathbf{A}(\boldsymbol{\theta}) &= E\left\{\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}^{*}_{\mathrm{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})\right\}, \\
\mathbf{B}(\boldsymbol{\theta}) &= E\left\{\mathbf{S}^{*}_{\mathrm{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})\mathbf{S}^{*\mathrm{T}}_{\mathrm{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})\right\}.
\end{aligned}$$

In addition, when $f^{*}_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1}) = f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$, the estimator achieves the optimal estimation variance $\mathbf{V}_{\mathrm{opt}} = \mathbf{B}^{-1}(\boldsymbol{\theta})$.

When Σ is unknown and needs to be estimated, the asymptotic normality results in Theorem 1 still hold, while the variance $\mathbf{V}$ will have a different form due to the additional variability caused by estimating Σ. However, the optimality result will no longer hold when an estimated variance is used to replace Σ. The proof of Theorem 1 is in appendix. When performing inference in practice, we can approximate $\mathbf{A}(\boldsymbol{\theta}), \mathbf{B}(\boldsymbol{\theta})$ with their respective sample versions evaluated at $\widehat{\boldsymbol{\theta}}$.

Once we have $\widehat{\boldsymbol{\theta}}$, we can use the procedure proposed in Hwang, Huang & Wang (2007) to estimate the population size

$$\widehat{N}_C = \sum_{i=1}^{N} \frac{I(\mathcal{C}_{i1})}{\widehat{\mathrm{pr}}(\mathcal{C}_{i1} \mid \Delta_i)}, \tag{2.4}$$

with the associated asymptotic variance

$$\begin{aligned}\mathrm{var}(\widehat{N}_c) &= \sum_{i=1}^{N} I(\mathcal{C}_{i1}) \frac{1-\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i)}{\mathrm{pr}^2(\mathcal{C}_{i1} \mid \Delta_i)} \\ &+ \left\{ \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \sum_{i=1}^{N} \frac{I(\mathcal{C}_{i1})}{\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i)} \right\} \mathrm{var}(\widehat{\boldsymbol{\theta}}) \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{N} \frac{I(\mathcal{C}_{i1})}{\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i)} \right\}.\end{aligned}$$

Here $\Delta_i = \mathbf{W}_{i(1)} + Y_i \boldsymbol{\Sigma}\boldsymbol{\beta}$ and

$$\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i) = 1 - \left[\sum_{y_i=0}^{J} \binom{J}{y_i} \exp\left\{ y_i(\alpha + \Delta_i^{\mathrm{T}}\boldsymbol{\beta}) - \frac{1}{2} y_i^2 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\beta} \right\} \right]^{-1},$$

$\widehat{\mathrm{pr}}(\mathcal{C}_{i1} \mid \Delta_i)$ is $\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i)$ calculated under $\widehat{\boldsymbol{\theta}}$, and $\mathrm{var}(\widehat{\boldsymbol{\theta}})$ is given in Theorem 1. We can easily obtain the approximation $\widehat{\mathrm{var}}(\widehat{N}_c)$ through replacing $\boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}}$ in the expression of $\mathrm{var}(\widehat{N}_C)$.

The first term of $\mathrm{var}(\widehat{N}_c)$ captures the variability of estimating N_c caused by not observing all the animals. This is completely decided by the data and is not affected by how well we estimate the parameters. The second term describes the additional variability due to the estimation of the related parameters. Hence if we reduce the estimation variance of the parameters, we can reduce the second term and reduce the overall variance in estimating N. In the simulation section, we will illustrate that the semiparametric method achieves this goal.

2.4 Simulation

We conduct a series of simulation experiments to investigate the performance of the semiparametric methods, in comparison with GMM and the conditional score method Hwang, Huang & Wang (2007). In each simulation experiment, we generated 1000 data sets.

In the first simulation, the true population size is set to be $N = 500$. We gener-

ated the true covariates X_i from uniform distribution Unif$[-3, 3]$, and set the measurement error standard deviation $\sigma_u = 0.6$. We then generated the observations $(Y_{ij}, W_{ij}Y_{ij}), j = 1, \cdots, 5$ from the model with the true parameter values $\alpha = 0.2$, $\beta = 1.0$. This yields an average of 417 first time captures and 335 second time captures, corresponding to high capture probability. To implement different estimators, we replaced Σ with its estimate $\widehat{\Sigma}$, which has bias -0.009 and variance 0.0007. In the semiparametric estimation, we implemented the estimation when both the true uniform distribution of X is used and a false normal distribution is used. Two GMM methods, one uses a single $\mathbf{W}_{i(k)}$ and the other uses the average $k^{-1}\sum_{l=1}^{k}\mathbf{W}_{i(l)}$, were implemented. Throughout the simulation section, we call them GMM1 and GMM2 respectively. The results of the various estimators are given in Table 2.1, where we reported the mean and the standard error of the estimators as well as the average of estimated standard errors and the sample coverage of the 95% confidence intervals constructed using the asymptotic results. From the results in Table 2.1, we can see that all five estimators have small biases in estimating the parameters α, β. Under finite sample, each of the five methods has positive bias for the population size N estimation. However, the bias decreases towards zero when larger sample sizes are used.

The two GMM methods perform similarly. Although the GMM estimators are able to reduce the estimation standard error for the model parameters α and β, they do not seem to reduce the estimating variability of the population size N. This is likely because the reduction on parameter estimation variability is not large enough so that it is masked out by the first term of var$(\widehat{N}_c)$. However, the semiparametric method yields larger variability reduction in the parameter estimation, hence this reduction is able to propagate to a more visible variability reduction of the population size estimate as well in the finite sample.

		α	β	N
	true	0.2	1.0	500
CS	estimate	0.1986	1.0035	503.12
	emp se	0.0710	0.0607	23.79
	mse	0.0101	0.0074	1163.3
	est se	0.0712	0.0606	23.39
	95% cov	95.1%	95.5%	94.8%
GMM1	estimate	0.1970	1.0091	504.44
	emp se	0.0688	0.0591	23.96
	mse	0.0095	0.0070	1179.7
	est se	0.0690	0.0585	23.40
	95% cov	95.0%	95.8%	95.3%
GMM2	estimate	0.1977	1.0083	504.24
	emp se	0.0702	0.0598	24.06
	mse	0.0099	0.0071	1187.8
	est se	0.0701	0.0590	23.48
	95% cov	94.9%	95.0%	95.3%
Semi-Nor	estimate	0.2006	1.0033	502.14
	emp se	0.0640	0.0537	22.92
	mse	0.0084	0.0059	1056.5
	est se	0.0658	0.0546	22.28
	95% cov	95.4%	95.1%	93.3%
Semi-Uni	estimate	0.2005	1.0031	502.13
	emp se	0.0640	0.0536	22.91
	mse	0.0084	0.0059	1055.7
	est se	0.0658	0.0546	22.27
	95% cov	95.6%	95.0%	93.3%

Table 2.1: Simulation 1. Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean square error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

The second simulation uses $\alpha = -1.0$, $\beta = 1.0$. This yields an average of 298 first time captures and 125 second time captures. The situation is more likely to happen in practice. All the other settings are the same as above, except that X_i is generated from a standard normal distribution. An estimated $\widehat{\Sigma}$ is used in all the

estimation procedures, where $\widehat{\Sigma}$ has bias -0.0019 and variance 0.0021. The semiparametric estimation also proposes both true and false distribution for X_i. The results of the various estimators are given in Table 2.2. From the results in Table 2.2, we can see that all five estimators have non-substantial biases in estimating both the parameters α, β and the population size N. With the exception of GMM estimators, the sample standard error and the average of the estimated standard errors are close to each other, indicating the satisfactory performance of the asymptotic results for the relatively small N. This is further reflected in the close proximity of the observed 95% coverage to its nominal level. The poor performance of GMM is likely caused by the small number of animals captured more than once. Indeed, in simulations not reported here, when we increase the population size, the performance of the GMM estimators improves. The two GMM estimators perform similarly, with GMM2 yields slightly smaller MSE than GMM1. However, both GMM estimators are clearly inferior compared to the semiparametric estimators.

Summarizing the first two simulation results, we find that the GMM did not improve drastically over the conditional score method in terms of the estimation efficiency. Its finite sample performance also heavily relies on the capture probability and the sample size, in that smaller sample sizes tend to inhibit the gain. Intuitively, the gain of the GMM comes mainly from the appropriate usage of additional measurements. When N is relatively small, there are very small amount of additional measurements available. This not only limits the source of additional information, but also adversely affects how such information is used, because the GMM weighting matrix relies on asymptotic results and is not a suitable approximation to the true weights under small number of recaptures. In contrast, the semiparametric methods perform satisfactorily. The efficiency in estimating α, β is improved by 15% and 36% respectively, while that for the population estimation is improved by 58% in the

		α	β	N
	true	-1	1	500
CS	estimate	-1.0229	1.0239	519.87
	emp se	0.1604	0.1829	85.55
	mse	0.0528	0.0651	15928
	est se	0.1612	0.1736	73.29
	95% cov	96.2%	94.2%	94.1%
GMM1	estimate	-1.0686	1.0692	544.22
	emp se	0.1782	0.2082	142.21
	mse	0.0682	0.0868	64989
	est se	0.1676	0.1800	94.34
	95% cov	95.6%	93.4%	96.2%
GMM2	estimate	-1.0673	1.0685	544.37
	emp se	0.1843	0.2107	147.12
	mse	0.0672	0.0827	53500
	est se	0.1663	0.1786	90.21
	95% cov	94.7%	93.5%	95.9%
Semi-Nor	estimate	-1.0101	1.0102	512.27
	emp se	0.1497	0.1571	68.16
	mse	0.0442	0.0494	10149
	est se	0.1465	0.1556	63.85
	95% cov	95.7%	95.0%	93.3%
Semi-Uni	estimate	-1.0103	1.0106	512.40
	emp se	0.1500	0.1577	68.36
	mse	0.0444	0.0497	10220
	est se	0.1468	0.1561	64.00
	95% cov	96.0%	95.0%	93.5%

Table 2.2: Simulation 2. Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean square error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

worse capture scenario. The dramatic efficiency gain on the population estimation is caused by the multiplication of $\sum_{i=1}^{N} \partial\{I(\mathcal{C}_{i1})/\mathrm{pr}(\mathcal{C}_{i1} \mid \Delta_i)\}/\partial\boldsymbol{\theta}^{\mathrm{T}}$ in the second term of $\mathrm{var}(\widehat{N}_c)$, which amplifies the magnitude of the change in $\mathrm{var}(\widehat{\boldsymbol{\theta}})$ in this model.

In the third simulation study, we generated the data by mimicking the bird data

structure in Section 2.5. Specifically, we set the true population size $N = 913$ and used $J = 12$ capture occasions. We generated the covariates X_i from a normal distribution with mean $\mu_X = 45.2$ and standard deviation $\sigma_X = 1.0$, while used variance $\sigma_u = 0.8$ to generate the measurement errors. The observations $(Y_{ij}, W_{ij}Y_{ij}), j = 1, \ldots, J$ are generated from the logistic model with $\alpha = -36.34$ and $\beta = 0.72$. These values are all reasonably close to the estimation results of the bird data example. Under this data generation procedure, the observed average number of first captures and second captures are respectively 244 and 49. Similar as in the first simulation, we conducted the five estimations procedures and the results are in Table 2.3. In all five estimators, an estimated $\widehat{\Sigma}$ is used, with bias -0.0136 and variance 0.1261. We see small biases of $\widehat{\boldsymbol{\theta}}$ and the population estimate in all estimators. Both the conditional score and the semiparametric methods yield close results between the sample estimation standard error and the empirical one, and between the observed coverage of the 95% confidence intervals and the nominal level, indicating the validity and relevance of the asymptotic results. Once more, the semiparametric methods provided the smallest estimation variability for both parameter and population estimation, with a gain of 59%, 63% and 277% in terms of estimation efficiency in comparison with the conditional score method. Among the two semiparametric methods, the normal-based procedure is slightly better performed than the uniform-based procedure, indicating the optimality when the true f_X distribution is used. However, the performance difference is very small, which is quite encouraging considering that the true f_X is not easy to obtain in practice. Despite the fact that GMM2 improves upon GMM1 through reducing estimation bias, the GMM estimates of population are not competitive in comparison with the semiparametric methods due to the small number of recaptures.

		α	β	N
	true	-36.34	0.72	913
CS	mean	-37.41	0.74	1031.44
	emp se	10.43	0.2251	527.56
	mse	353.43	0.1651	$1.994e^6$
	est se	10.21	0.2202	381.67
	95% cov	94.7%	94.7%	93.3%
GMM1	mean	-37.94	0.75	1051.96
	emp se	9.7499	0.2118	264.97
	mse	376.34	0.1769	$4.539e^5$
	est se	11.25	0.2427	365.93
	95% cov	97.6%	97.6%	97.7%
GMM2	mean	-37.93	0.75	1048.81
	emp se	9.3641	0.2033	269.47
	mse	252.54	0.1183	$2.833e^5$
	est se	10.53	0.2271	332.50
	95% cov	98.3%	98.3%	98.3%
Semi-Nor	mean	-36.74	0.73	976.73
	emp se	8.2814	0.1787	271.72
	mse	137.01	0.0638	$1.705e^5$
	est se	8.0443	0.1735	245.43
	95% cov	94.2%	94.3%	92.4%
Semi-Uni	mean	-36.91	0.73	982.07
	emp se	8.4262	0.1818	279.99
	mse	141.89	0.0661	$1.798e^5$
	est se	8.2185	0.1773	250.39
	95% cov	94.2%	94.3%	92.6%

Table 2.3: Simulation 3. Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean square error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

2.5 Data Example

We implement the conditional score method, the generalized method of moments and the semiparametric method on a data set regarding the bird species *Prinia flaviventris* collected by the Hong Kong Bird Society from 1991 to 1995. In addition

to the capture record, the data set contains the measurements of a bird's wing length, which is believed to be directly linked with a bird's capture probability and is measured with error. We consider a subset of the data from 01/31/93 to 04/11/93. In this relatively short time period, the population size change is likely small and is negligible. Therefore, we treat it as a closed population. During this time period, 146 distinct birds were captured and measured in $J = 12$ occasions with 168 total captures. Among them, the average wing length is 45.20 and the variance is 1.64.

Under the normal additive measurement error assumption, taking advantage of the multiple measurements of the wing length of recaptured birds, we form the difference between the measurements and estimated the variance of the measurement error to be $\widehat{\sigma}_u^2 = 0.626$.

The results of parameter and population size estimation based on the five methods are summarized in Table 2.4, where we proposed both a normal and a uniform working model in the semiparametric estimation. In the GMM implementation, we only incorporated the first two captures despite that the maximum total recapture is five. This is because only four animals are captured three or more times and this sample size is certainly too small to justify any analysis based on asymptotic results. Table 2.4 indicates that both the GMM and the semiparametric methods result in estimation variance reduction in comparison to the conditional score method, while the improvement from the two semiparametric methods are especially important in terms of estimating the population size. The improvement in the semiparametric estimators is likely quite reliable as is reflected in the simulation studies. However, we would like to caution that the improvement from GMM is less trustworthy. This is because the inference result of GMM at this population size may not be sufficiently precise, and it tends to under-estimate the estimation variability, as is exhibited in the simulation studies.

	$\widehat{\alpha}$(se)	$\widehat{\beta}$(se)	$\widehat{N}$(se)
CS	-40.11 (31.742)	0.80 (0.694)	921.46 (713.767)
GMM(First 2 captures)	-35.46 (9.274)	0.69 (0.201)	1117.31 (431.082)
Semi-Nor	-37.03 (14.451)	0.73 (0.318)	839.58 (293.993)
Semi-Uni	-32.75 (9.66)	0.64 (0.213)	770.82 (225.332)

Table 2.4: Data analysis. Estimation and the associated standard error (se) of bird data analysis based on conditional score (CS), generalized method of moments (GMM) and Semiparametric methods using normal (Semi-Nor) and uniform (Semi-Uni) candidate distributions for the wing lengths.

2.6 Discussion

We have investigated the issue of using multiple error-prone covariate measurements in the capture-recapture models. Among the two methods that we propose, we have found that the GMM estimation tend not to perform satisfactorily while the semiparametric methods generally demonstrate good performance. We emphasize here that although multiple measurements are often easy to be taken into account in most measurement error models, it is not the case in the capture-recapture model context. This is a direct consequence of two difficulties, the violation of the commonly assumed surrogacy assumption and the loss of the complete sufficient statistic. Because of the loss of surrogacy, together with the need to estimate measurement error structure, to handle multiple measurements and missing observations, this work made a breakthrough in the field of capture-recapture models by modifying ideas from Tsiatis and Ma (2004). Although GMM is a possible way of taking into account the multiple measurements, its usage of the information is somewhat superficial. Our semiparametric approach, in contrast, forms an improved measurements and takes advantage of this information directly in the core construction of the estimator, hence is a more profound way of using the multiple measurements. Its effectiveness has been reflected in both the theoretical analysis and the numerical results.

We would like to point out three major differences of our problem setting and approach in comparison with Xi, Watson, Wang & Yip (2009). First, Xi, Watson, Wang & Yip (2009) assumes a parametric distribution model for the error prone covariates, hence they work in the structural measurement error model framework and their final model is a parametric one. In contrast, we leave the distribution of the error prone covariates completely unspecified, hence we work in a functional measurement error model and have to deal with a semiparametric problem. Second, the parametric model in Xi, Watson, Wang & Yip (2009) permits the construction of a likelihood, and a maximum likelihood estimator is therefore used for estimation. In contrast, we do not have a well-defined likelihood function, hence we seek various ways to construct estimating equations to develop estimators. Finally, in terms of computation, an EM algorithm is implemented to obtain the maximum likelihood estimator in Xi, Watson, Wang & Yip (2009), while we resort to a Newton-Raphson procedure in combination with integral equation solving to obtain the semiparametric estimator.

We are aware that when the sample size is moderate or small, the asymptotic properties of the semiparametric estimator may not exhibit well. In addition, the superior performance of the semiparametric estimator comes with the price of relatively more intensive computation. These are limitations of the semiparametric estimator.

Finally, throughout the article, we have worked under a closed population assumption in the capture-recapture framework. Because we only considered a short time period, it is a reasonable assumption for our data example (Hwang & Huang (2003), Hwang, Huang & Wang (2007) and Xi, Watson, Wang & Yip (2009)). On the other hand, it is certainly of interest to also study the population sizes over the whole five year period that the data were collected. In this case, open population models, such as the Jolly-Seber open population model (Jolly (1965), Seber (1982) and Seber

(1986)), will be more appropriate. See also recent developments in open population studies in Schwarz & Arnason (1996) and Pledger, Pollock & Norris (2003).

3. INSTRUMENT ASSISTED REGRESSION FOR ERRORS IN VARIABLES MODELS WITH BINARY RESPONSE

3.1 Introduction

Logistic and probit models are widely used in regression analysis with binary response. They belong to the family of generalized linear models. In real data analysis, particularly in the analysis of medical and clinical data, a ubiquitous problem is that some or all covariates cannot be directly or precisely measured and indirect or proxy measurements are used instead. For example, in studies of human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS), important variables such as CD4 lymphocyte count cannot be accurately measured due to instrument's limitation or individual biological variation. Other well-known examples include blood pressure and cholesterol level in cardiovascular disease research. It is well-known that ignoring the measurement error and simply replacing the true covariates with their mismeasured proxies will lead to biased estimates and thus invalid conclusions (Stefanski & Buzas (1995)).

Although the problem of measurement error in general has been extensively studied in the literature, research focusing specifically on binary regression with instrumental variables is limited. Stefanski & Carroll (1985) and Stefanski & Buzas (1995) proposed approximate estimators for functional logistic models, while Stefanski & Carroll (1987) and Ma & Tsiatis (2006b) studied consistent estimators for generalized linear models based on conditional score functions under the assumption of normal measurement errors or unknown measurement error distribution. Huang and Wang (2001) proposed alternative estimating function correction schemes to obtain consistent estimators for the cases where the measurement error distribution is known

or the replicate data are available. These works did not use instrumental variable approach, although Huang and Wang (2001) discussed the possibility in their setup. Buzas & Stefanski (1996) considered instrumental variable approach to functional generalized linear models. However, their approach requires the normality assumption for both the measurement error and instrumental variables.

Therefore, an interesting question is whether it is possible to use instrumental variables to obtain consistent estimators without normality assumption for both the unobserved covariates and the measurement errors. In this paper, we demonstrate that this is possible in a wide range of models. In particular, we show that this can be achieved by employing a prediction relationship for the unobserved covariates using the instruments. Similar use of the instruments in some special models also appeared in Buzas (1997). This way of incorporating instrumental variables is different than most other methods mentioned above, and its applicability in the generality of the model has also not been achieved before. Thus, our work is the first in using instruments in the general regression models with measurement error and binary response, where the link between the response and the covariates does not need to belong to any special regression family.

Instrumental variable approach has been used by other authors to deal with errors-in-variables problem in general nonlinear models, e.g., Amemiya (1985), Amemiya (1990), Schennach (2007), Wang & Hsiao (2011), and Abarin & Wang (2012). In particular, Schennach (2007) and Wang & Hsiao (2011) show that the nonlinear measurement error models are generally identified when instrumental variables are available. In recent years, instrumental variable approach has drawn more and more attention in the literature, partly due to its methodological flexibility and practical applicability. In practice, any observable variables that are correlated with unobserved covariates but independent of measurement error can be used as instruments.

In particular, the replicate measurements can be regarded as special instruments.

The rest of the section is organized as the following. We present the model we study and our main methodology in section 3.2. In this section, we also establish the asymptotic properties of our estimator. Numerical work including both simulations and real data analysis is given in section 3.3. We conclude the section with some discussions on the generalization and possible extension of the method in section 3.4. All the technical details are in appendix.

3.2 Main Results

3.2.1 The Model

The model we study can be explicitly written as

$$\mathrm{pr}(Y = 1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}) \tag{3.1}$$

where H is a known inverse link function, for example, the inverse logit link function $H(\cdot) = 1 - 1/\{\exp(\cdot) + 1\}$ or the inverse probit link function $H(\cdot) = \Phi(\cdot)$. While the response variable Y and the covariate $\mathbf{Z}$ are observed, the covariate $\mathbf{X}$ is a latent variable. Instead of observing $\mathbf{X}$, we observe an erroneous version of $\mathbf{X}$, written as $\mathbf{W}$ and an instrumental variable $\mathbf{S}$. The variables $\mathbf{W}$ and $\mathbf{S}$ are linked to $\mathbf{X}$ through

$$\mathbf{W} = \mathbf{X} + \mathbf{U} \qquad \text{and} \qquad \mathbf{X} = \mathbf{m}(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}) + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\mathbf{m}$ is a known function up to an unknown parameter $\boldsymbol{\alpha}$. Here we assume the conditional mean of $\boldsymbol{\epsilon}$ and the marginal mean of $\mathbf{U}$ to be zero, i.e. $E(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = \mathbf{0}$, $E(\mathbf{U}) = \mathbf{0}$. We further assume that $(\mathbf{S}, \mathbf{Z}, \mathbf{X})$ is independent of $\mathbf{U}$, $\mathbf{U}$ is independent of $\boldsymbol{\epsilon}$, $\mathbf{W}$ is independent of $(\mathbf{S}, \mathbf{Z})$ given $\mathbf{X}$, and Y is independent of $(\mathbf{S}, \mathbf{W})$ given $(\mathbf{X}, \mathbf{Z})$. The observed data are $(\mathbf{Z}_i, \mathbf{S}_i, \mathbf{W}_i, Y_i)$, $i = 1, \ldots, n$. They are independent

and identically distributed (iid) according to the model described in (3.1) and (3.2). Our main interest is in estimating $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. The problem considered here can be viewed as a generalization of the one considered in Buzas & Stefanski (1996), in that we have much less stringent conditions. For example, we do not impose the normality assumption on $\mathbf{X}, \mathbf{S}, \boldsymbol{\epsilon}, \mathbf{U}$, while this is required there. Note also that parametric assumption of the regression function m in (3.2) is not restrictive, because it can be easily checked using data on $(\mathbf{W}, \mathbf{S}, \mathbf{Z})$ (see (3.3) below).

3.2.2 A Simplification

To proceed with estimation, we first recognize that from the relations described in (3.2), we have

$$\mathbf{W} = \mathbf{m}(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}) + \mathbf{U} + \boldsymbol{\epsilon}, \tag{3.3}$$

where $E(\mathbf{U} + \boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = \mathbf{0}$. It is easy to see that this is a familiar mean regression model, so we can use least squares method to get a consistent estimator of $\boldsymbol{\alpha}$. Specifically, we can solve the estimating equation

$$\sum_{i=1}^{n} \mathcal{S}_{\alpha}(\mathbf{S}_i, \mathbf{Z}_i) = \sum_{i=1}^{n} \frac{\partial \mathbf{m}^{\mathrm{T}}(\mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \boldsymbol{\Omega}(\mathbf{S}_i, \mathbf{Z}_i)\{\mathbf{W}_i - \mathbf{m}(\mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\alpha})\} = \mathbf{0}, \tag{3.4}$$

where $\boldsymbol{\Omega}(\mathbf{S}, \mathbf{Z})$ is any weight matrix, to obtain a consistent estimator $\widehat{\boldsymbol{\alpha}}$. Obviously, if we set $\boldsymbol{\Omega}(\mathbf{S}, \mathbf{Z})$ to be the identity matrix, we obtain the ordinary least squares (OLS) estimator of $\boldsymbol{\alpha}$, while if we set $\boldsymbol{\Omega}(\mathbf{S}, \mathbf{Z})$ to be the inverse of the error variance-covariance matrix conditional on $(\mathbf{S}, \mathbf{Z})$, we obtain the optimal weighted least squares estimator (WLS) of $\boldsymbol{\alpha}$. Once we have an estimate $\widehat{\boldsymbol{\alpha}}$, we can plug the relation between

$\mathbf{X}$ and $(\mathbf{S}, \mathbf{Z})$ into model (3.1) to obtain the joint distribution of $(Y, \mathbf{S}, \mathbf{Z})$ as

$$\begin{aligned} &\mathrm{pr}(Y = y, \mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z}) = \\ &f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z}) \int \left[1 - y + (2y - 1)H\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\beta}\}\right] f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\epsilon}), \end{aligned} \tag{3.5}$$

where $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z})$ is a conditional probability density function (pdf) that satisfies $\int \boldsymbol{\epsilon} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\epsilon}) = \mathbf{0}$, and $f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z})$ is the joint pdf of $(\mathbf{S}, \mathbf{Z})$.

3.2.3 Semiparametric Derivation

We now derive the estimation procedure for $\boldsymbol{\beta}, \boldsymbol{\gamma}$ from the above form. For simplicity, we write $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and assume $\boldsymbol{\theta} \in \mathbb{R}^p$. Then the pdf in (3.5) involves the unknown parameter $\boldsymbol{\theta}$ and unknown functions $f_{\boldsymbol{\epsilon}}(\cdot), f_{\mathbf{S},\mathbf{Z}}(\cdot)$, while we are only interested in $\boldsymbol{\theta}$. Thus, $f_{\boldsymbol{\epsilon}}(\cdot), f_{\mathbf{S},\mathbf{Z}}(\cdot)$ can be viewed as two infinite dimensional nuisance parameters. This allows us to view the model as a semiparametric model and use the existing semiparametric approaches (Bickel, Klaassen, Ritov & Wellner (1993), Tsiatis (2006)). In the measurement error framework, semiparametric methods were first introduced in Tsiatis & Ma (2004) in the context of a known error distribution. Following the semiparametric approach, our estimator will be based on the efficient score function. In general, the efficient score function can be obtained through projecting the score function $\mathbf{S}_\theta(Y, \mathbf{S}, \mathbf{Z}) \equiv \partial \log f_{\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}}\{\boldsymbol{\epsilon}, \mathbf{s}, \mathbf{z}; \boldsymbol{\theta}, f_{\boldsymbol{\epsilon}}(\cdot), f_{\mathbf{S},\mathbf{Z}}(\cdot)\}/\partial\boldsymbol{\theta}$ onto the orthogonal complement of the nuisance tangent space. The nuisance tangent space is defined as the mean square closure of the nuisance tangent spaces associated with all possible parametric submodels of a semiparametric model (See Tsiatis (2006), Section 4), and is often hard to obtain. In appendix, we derive the nuisance

tangent space associated with model (3.5) as

$$\begin{aligned}\Lambda &= \Lambda_1 \oplus \Lambda_2 \\ &= \{\mathbf{f}(\mathbf{S},\mathbf{Z}) : \mathbf{f} \in \mathbb{R}^p, E(\mathbf{f}) = \mathbf{0}, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty, \forall \mathbf{f}\} \\ &\oplus \ [E\{\mathbf{f}(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid Y,\mathbf{S},\mathbf{Z}\} : \mathbf{f} \in \mathbb{R}^p, E(\mathbf{f} \mid \mathbf{S},\mathbf{Z}) = \mathbf{0}, \\ &\quad E(\boldsymbol{\epsilon}\mathbf{f}^{\mathrm{T}} \mid \mathbf{S},\mathbf{Z}) = \mathbf{0}, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty, \forall \mathbf{f}].\end{aligned}$$

Here, we use the notation $\oplus$ to emphasize that an arbitrary function $\mathbf{f}_1(\mathbf{S},\mathbf{Z})$ in Λ_1 and an arbitrary function $\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})$ in Λ_2 satisfy $E\{\mathbf{f}_1(\mathbf{S},\mathbf{Z})\mathbf{f}_2^{\mathrm{T}}(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})\} = \mathbf{0}$. The orthogonal complement of Λ can then be derived as

$$\Lambda^{\perp} = \{\mathbf{f}(Y,\mathbf{S},\mathbf{Z}) : \mathbf{f} \in \mathbb{R}^p, E(\mathbf{f} \mid \boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) = \mathbf{a}(\mathbf{S},\mathbf{Z})\boldsymbol{\epsilon}, E(\mathbf{a}^{\mathrm{T}}\mathbf{a}) < \infty\},$$

where $\mathbf{a}(\mathbf{S},\mathbf{Z})$ contains p rows and conforms with the dimension of $\boldsymbol{\epsilon}$. We also need to calculate the score function with respect to $\boldsymbol{\theta}$, which has the form

$$\mathcal{S}_\theta(Y,\mathbf{S},\mathbf{Z}) = (2Y-1)\cdot \frac{\int \left\{ \begin{array}{c} \mathbf{m}(\mathbf{S},\mathbf{Z},\widehat{\boldsymbol{\alpha}}) + \boldsymbol{\epsilon} \\ \mathbf{Z} \end{array} \right\} H'\{\mathbf{m}(\mathbf{S},\mathbf{Z},\widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\beta}\} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S},\mathbf{Z}) d\mu(\boldsymbol{\epsilon})}{\int [1-Y+(2Y-1)H\{\mathbf{m}(\mathbf{S},\mathbf{Z},\widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\beta}\}] f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S},\mathbf{Z}) d\mu(\boldsymbol{\epsilon})}.$$

The efficient score can now be obtained by projecting $\mathcal{S}_\theta$ to $\Lambda^{\perp}$, and can be verified as

$$\mathcal{S}_{\text{eff}}(Y,\mathbf{S},\mathbf{Z}) = \mathcal{S}_\theta(Y,\mathbf{S},\mathbf{Z}) - E\{\mathbf{b}(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid Y,\mathbf{S},\mathbf{Z}\},$$

where $\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})$ satisfies

$$E\left[\, \mathcal{S}_\theta(Y, \mathbf{S}, \mathbf{Z}) - E\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\right] = \mathbf{a}(\mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon} \tag{3.6}$$

for some function $\mathbf{a}(\mathbf{S}, \mathbf{Z})$. Unfortunately, $\mathbf{a}(\mathbf{S}, \mathbf{Z})$ is unspecified in (3.6), hence we cannot directly solve for $\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})$ from (3.6). In order to determine the function $\mathbf{a}(\mathbf{S}, \mathbf{Z})$, we multiply $\boldsymbol{\epsilon}$ on both sides of (3.6), take expectation conditional on $(\mathbf{S}, \mathbf{Z})$, and obtain

$$E\left[\, \mathcal{S}_\theta(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} - E\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right] = \mathbf{a}(\mathbf{S}, \mathbf{Z})E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}).$$

This implies

$$\mathbf{a}(\mathbf{S}, \mathbf{Z}) = E\left[\, \mathcal{S}_\theta(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} - E\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right] \left\{E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1}.$$

We can now plug the form of $\mathbf{a}(\mathbf{S}, \mathbf{Z})$ into (3.6) to obtain an explicit integral equation

$$\begin{aligned} & E\left[\, \mathcal{S}_\theta(Y, \mathbf{S}, \mathbf{Z}) - E\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\right] \\ = \;& E\left[\, \mathcal{S}_\theta(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} - E\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right] \left\{E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1} \boldsymbol{\epsilon}. \end{aligned}$$

This integral equation no longer contains unspecified component, and $\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})$ can be obtained as a solution to the equation.

3.2.4 Estimation Under Working Model

The above derivation is performed under a true density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ which is usually unknown. In order to be able to compute $\mathcal{S}_\theta$ or $\mathcal{S}_{\text{eff}}$, we propose to use a working model $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, which may or may not be equal to $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, and

perform all the calculations under this working model. The name "working model" means that $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ is not a part of the model assumption. It is merely used for constructing our estimator. This is in contrast to $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, which is the true model that defines the data generation process. Using * to denote all the affected quantities by the substitution of $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ with $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, our estimation procedure is the following.

1. Propose a working model $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ that has mean zero. For example, we can propose $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ to be a normal pdf with mean $\mathbf{0}$ and variance $\mathbf{I}$.

2. Calculate the score function $\mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z})$ under the working model.

3. Obtain $\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})$ through solving the integral equation

$$
\begin{aligned}
&E\left[\, \mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z}) - E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\right] = \\
&E^*\left[\, \mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} - E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right] \cdot \\
&\left\{E^*(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1} \boldsymbol{\epsilon}. \qquad (3.7)
\end{aligned}
$$

4. Form

$$
\mathcal{S}^*_{\text{eff}}(Y, \mathbf{S}, \mathbf{Z}) = \mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z}) - E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}
$$

and solve the estimating equation

$$
\sum_{i=1}^{n} \mathcal{S}^*_{\text{eff}}(Y_i, \mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \mathbf{0}
$$

to obtain the estimator $\widehat{\boldsymbol{\theta}}$.

In the above step 3, we solved the integration equation (3.7) via converting it to a linear algebra problem. Specifically, based on the working model, we first discretize the distribution of $\boldsymbol{\epsilon}$ on m points $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m$. A typical practice is to choose m equally spaced points on the support of the distribution. We then calculate the probability mass $\pi_i(\mathbf{S}, \mathbf{Z})$ at each of the m points and normalize the $\pi_i(\mathbf{S}, \mathbf{Z})$'s so that $\sum_{i=1}^m \pi_i(\mathbf{S}, \mathbf{Z}) = 1$. This allows us to approximate the calculation of E^* with $\widehat{E}^*$. For example, denoting

$$\widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_i, Y \mid \mathbf{S}, \mathbf{Z}) = \left[1 - y + (2y-1)H\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{\beta}\}\right] \pi_i(\mathbf{S}, \mathbf{Z}),$$

we replace $E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}$ with

$$\widehat{E}^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} = \frac{\sum_{i=1}^m \mathbf{b}(\boldsymbol{\epsilon}_i, \mathbf{S}, \mathbf{Z}) \widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_i, Y \mid \mathbf{S}, \mathbf{Z})}{\sum_{i=1}^m \widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_i, Y \mid \mathbf{S}, \mathbf{Z})}.$$

Let

$$\mathbf{B}(\mathbf{S}, \mathbf{Z}) = \{\mathbf{b}(\boldsymbol{\epsilon}_1, \mathbf{S}, \mathbf{Z}), \ldots, \mathbf{b}(\boldsymbol{\epsilon}_m, \mathbf{S}, \mathbf{Z})\}^{\mathrm{T}},$$

$$\mathbf{C}(\mathbf{S}, \mathbf{Z}) = \{\mathbf{c}(\boldsymbol{\epsilon}_1, \mathbf{S}, \mathbf{Z}), \ldots, \mathbf{c}(\boldsymbol{\epsilon}_m, \mathbf{S}, \mathbf{Z})\}^{\mathrm{T}},$$

where

$$\begin{aligned} & \mathbf{c}(\boldsymbol{\epsilon}_i, \mathbf{S}, \mathbf{Z}) \\ = \; & E\left\{ \mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z}) \mid \boldsymbol{\epsilon}_i, \mathbf{S}, \mathbf{Z}\right\} - E\left\{ \mathcal{S}^*_\theta(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right\} \left\{E^*(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1} \boldsymbol{\epsilon}_i. \end{aligned}$$

Further, let $\mathbf{A}(\mathbf{S}, \mathbf{Z})$ be an $m \times m$ matrix whose (i, j) block is

$$E\left\{\frac{\widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_j, Y, \mathbf{S}, \mathbf{Z})}{\sum_{i=1}^m \widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_i, Y, \mathbf{S}, \mathbf{Z})} \mid \boldsymbol{\epsilon}_i, \mathbf{S}, \mathbf{Z}\right\}$$
$$- \quad \widehat{E}^*\left\{\frac{\widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_j, Y, \mathbf{S}, \mathbf{Z})}{\sum_{i=1}^m \widehat{f}^*_{\boldsymbol{\epsilon},Y}(\boldsymbol{\epsilon}_i, Y, \mathbf{S}, \mathbf{Z})}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right\}\left\{\widehat{E}^*(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1}\boldsymbol{\epsilon}_i.$$

The integral equation (3.7) can then be converted into a linear algebra problem

$$\mathbf{A}(S, Z)\mathbf{B}(S, Z) = \mathbf{C}(S, Z),$$

and we can readily solve it for $\mathbf{b}(\boldsymbol{\epsilon}_i, S, Z)$'s.

3.2.5 Asymptotic Properties

Although the working model $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ does not necessarily equal to the true model $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, the above procedure still yields a consistent estimator $\widehat{\boldsymbol{\theta}}$. Let $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for all matrix or vector $\mathbf{a}$ throughout the text. In appendix, we prove the following theorem.

Theorem 2. *Under suitable regularity conditions, if $\boldsymbol{\alpha}$ is known, then $\widehat{\boldsymbol{\theta}}$ obtained from the procedure described above satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

when $n \to \infty$. Here

$$\mathbf{A} = E\left\{\frac{\partial\ \mathcal{S}^*_{\mathrm{eff}}(Y, \mathbf{S}, \mathbf{Z})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}, \quad \mathbf{B} = E\{\ \mathcal{S}^*_{\mathrm{eff}}(Y, \mathbf{S}, \mathbf{Z})^{\otimes 2}\}.$$

In addition, when $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, the variance is $[E\{\ \mathcal{S}_{\mathrm{eff}}(Y, \mathbf{S}, \mathbf{Z})^{\otimes 2}\}]^{-1}$, which is the minimum semiparametric variance bound for estimating $\boldsymbol{\theta}$.

In practice, $\boldsymbol{\alpha}$ is unknown and $\widehat{\boldsymbol{\theta}}$ is obtained from using $\widehat{\boldsymbol{\alpha}}$, an estimator obtained from solving (3.4). Hence additional variability associated with estimating $\boldsymbol{\alpha}$ occurs and needs to be taken into account. In this case, we have the following result.

Theorem 3. *When $\boldsymbol{\alpha}$ is estimated from (3.4) and $\widehat{\boldsymbol{\alpha}}$ is used in the estimation procedure, then the resulting plug-in estimator $\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}})$ satisfies*

$$\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}\} \to N(\mathbf{0}, \mathbf{V})$$

when $n \to \infty$. Here $\mathbf{V} = \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}} + \mathbf{V}_\alpha$ and

$$\mathbf{V}_\alpha = \mathbf{A}^{-1}\left\{\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_2(\mathbf{A}_1\mathbf{A}_2^{-1})^{\mathrm{T}} - \mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_1 - (\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_1)^{\mathrm{T}}\right\}(\mathbf{A}^{-1})^{\mathrm{T}},$$

where $\mathbf{A}, \mathbf{B}$ are given in Theorem 2, $\mathbf{A}_1 = E\{\partial\, \mathcal{S}^*_{\mathrm{eff}}/\partial\boldsymbol{\alpha}^{\mathrm{T}}\}$, $\mathbf{A}_2 = E\{\partial\, \mathcal{S}_\alpha/\partial\boldsymbol{\alpha}^{\mathrm{T}}\}$, $\mathbf{B}_1 = E(\, \mathcal{S}_\alpha\, \mathcal{S}^{*\mathrm{T}}_{\mathrm{eff}})$, $\mathbf{B}_2 = E(\, \mathcal{S}_\alpha^{\otimes 2})$. In addition, when $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, the resulting estimation variance is minimized among all the plug-in estimators.

3.3 Numerical Examples

We now demonstrate our method numerically through both simulated and real data examples. In all simulated examples, 1000 data sets were generated with sample size $n = 1000$.

3.3.1 Simulated Example One

In our first simulation, we generated the observations (Z_i, S_i, W_i, Y_i) from the model

$$\begin{aligned} \Pr(Y_i = 1 \mid X_i = x_i, Z_i = z_i) &= H(\beta x_i + \gamma z_i), \\ W_i &= X_i + U_i, \\ X_i &= \alpha_1 + \alpha_2 S_i + \epsilon_i. \end{aligned}$$

Here, $H(\cdot)$ is respectively set to be the inverse logit and the inverse probit link function, and $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta = 0.3$, $\gamma = 0.5$. The observable covariate Z_i and the instrument variable S_i are generated from the standard normal distribution. We generated U_i from a normal distribution with mean zero and variance 0.6. We further generated ϵ_i respectively from a normal distribution with mean 0 and variance $S_i^2/2$, and a t_5 distribution multiplied by $(|S_i|/3)^{1/2}$. Those two cases correspond to a normal and a non-normal regression model $W_i = \alpha_1 S_i + \alpha_2 Z_i + U_i + \epsilon_i$ with heteroscedastic error. finally, we proposed a normal working model on ϵ_i. Thus, the estimation in the two cases corresponds to a correct and a misspecified working model.

The combination of the logit and probit link functions with the normal and non-normal regression errors yields four different cases, and the performances of our method in all four scenarios are summarized in Table 3.1. Because the OLS and WLS are the most popular methods of estimating $(\alpha_1, \alpha_2)^{\mathrm{T}}$, we calculated both of them in our simulation and compared the performance with the estimation under the known $\boldsymbol{\alpha}$.

Based on Table 3.1, it is obvious that the estimators for (β, γ) have very small

	truth	α_1 1.0	α_2 1.0	β(logit) 0.3	γ(logit) 0.5	β(probit) 0.3	γ(probit) 0.5	β(as) 0.3	γ(as) 0.5
ϵ: Normal distribution									
α_0	mean			0.2994	0.4984	0.3006	0.4999	0.2992	0.4981
	median			0.3005	0.4948	0.3001	0.5002	0.2997	0.4945
	emp se	NA	NA	0.0526	0.0712	0.0366	0.0498	0.0521	0.0706
	est se			0.0509	0.0708	0.0355	0.0478	0.0501	0.0709
	95% cov			94.7%	95.3%	95.3%	93.0%	93.9%	95.6%
OLS	mean	0.9999	1.0013	0.2992	0.4981	0.3006	0.4997	0.2992	0.4980
	median	1.0015	1.0025	0.2990	0.4941	0.2994	0.4998	0.2998	0.4947
	emp se	0.0334	0.0443	0.0530	0.0707	0.0372	0.0496	0.0521	0.0707
	est se	0.0331	0.0456	0.0509	0.0708	0.0355	0.0478	0.0500	0.0709
	95% cov	94.3%	95.3%	94.0%	95.4%	93.9%	93.3%	93.9%	95.6%
WLS	mean	0.9999	0.9999	0.2994	0.4981	0.3008	0.4997	0.2992	0.4980
	median	1.0001	1.0007	0.2997	0.4943	0.2997	0.5000	0.2998	0.4946
	emp se	0.0299	0.0393	0.0531	0.0707	0.0371	0.0496	0.0521	0.0707
	est se	0.0297	0.0398	0.0510	0.0708	0.0356	0.0478	0.0500	0.0709
	95% cov	95.0%	96.1%	94.2%	95.4%	94.2%	93.3%	93.9%	95.6%
ϵ: Student t distribution t_5									
α_0	mean			0.2994	0.4984	0.3004	0.4992	0.2986	0.4983
	median			0.2993	0.4960	0.2986	0.4974	0.2996	0.4972
	emp se	NA	NA	0.0528	0.0718	0.0370	0.0487	0.0515	0.0713
	est se			0.0507	0.0707	0.0349	0.0476	0.0498	0.0709
	95% cov			93.7%	95.9%	93.8%	94.4%	94.3%	95.7%
OLS	mean	0.9984	0.9993	0.2998	0.4984	0.3007	0.4989	0.2985	0.4983
	median	0.9969	0.9998	0.2996	0.4959	0.2988	0.4975	0.2994	0.4972
	emp se	0.0316	0.0378	0.0528	0.0718	0.0371	0.0487	0.0516	0.0713
	est se	0.0303	0.0384	0.0508	0.0707	0.0350	0.0476	0.0498	0.0709
	95% cov	95.3%	95.8%	94.0%	95.8%	94.0%	94.3%	94.2%	95.6%
WLS	mean	0.9989	0.9989	0.2997	0.4984	0.3007	0.4989	0.2985	0.4983
	median	0.9977	0.9996	0.2995	0.4961	0.2985	0.4976	0.2991	0.4972
	emp se	0.0303	0.0370	0.0529	0.0718	0.0372	0.0487	0.0516	0.0713
	est se	0.0308	0.0373	0.0508	0.0707	0.0350	0.0476	0.0498	0.0709
	95% cov	95.4%	95.8%	94.1%	95.8%	94.0%	94.3%	94.2%	95.6%

Table 3.1: Simulation 1: Estimation and inference results on $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}$, $\widehat{\gamma}$. The estimation mean, median, empirical standard error, estimated standard error and coverage rate of the 95% confidence intervals are reported. α_0 means the true α's are used. "as" stands the adjusted score method, implemented in the logit model only.

bias in all cases. In addition, the empirical and average estimated standard errors match closely, and the empirical coverage of the 95% confidence intervals are very close to the nominal level. All these indicate satisfactory accuracy of our inference results in the finite sample situations.

In the logistic model context, Buzas (1997) developed an adjusted score method. For comparison, we included the adjusted score results in our simulation, see Table 3.1. Its performance in terms of means, estimation variability and coverage probabilities are similar to our method. The drawback of the adjusted score method is its limited applicability. For example, it can only be used for the logistic link function.

One can observe an interesting phenomenon regarding the relative efficiency of the estimators for β and γ under different $\boldsymbol{\alpha}$ estimators in comparison with the known $\boldsymbol{\alpha}$ case. On the one hand, it is clear that for estimating $\boldsymbol{\alpha}$, the WLS is much more efficient than the OLS estimator. On the other hand, the difference in the estimation variability for $\widehat{\boldsymbol{\alpha}}$ does not seem to influence much the estimation variability for $\widehat{\beta}$ and $\widehat{\gamma}$. In fact, even when the estimation is conducted under the known $\boldsymbol{\alpha}$, the variability of $\widehat{\beta}$ and $\widehat{\gamma}$ does not seem to improve much in this simulation example. However, we point out that this is not always the case. For example, when we generate U_i from a centered normal distribution with variance 8, the estimation variability of $\widehat{\beta}$ and $\widehat{\gamma}$ decreased visibly when $\boldsymbol{\alpha}$ is known, see Table 3.2 for details. In fact, how does the variability of $\widehat{\boldsymbol{\alpha}}$ affect that of $\widehat{\beta}$ and $\widehat{\gamma}$ is difficult to quantify, despite the analytic result in Theorem 3.

	Initial values	α_1 1.0	α_2 1.0	β (logit) 0.3	γ (logit) 0.5
α known	mean			0.3002	0.4993
	median			0.2990	0.4995
	emp se	NA	NA	0.0766	0.0938
	est se			0.0754	0.0950
	95% cov			94.8%	96.3%
OLS	mean	0.9983	1.0018	0.3023	0.4978
	median	1.0000	1.0006	0.2980	0.5004
	emp se	0.0930	0.0950	0.0813	0.0999
	est se	0.0923	0.0973	0.0811	0.1028
	95% cov	94.9%	95.6%	94.8%	96.8%
WLS	mean	0.9984	1.0009	0.3026	0.4975
	median	1.0000	1.0005	0.2979	0.5005
	emp se	0.0929	0.0950	0.0815	0.1001
	est se	0.0920	0.0968	0.0812	0.1030
	95% cov	95.0%	95.8%	94.7%	96.8%

Table 3.2: Simulation 1: Estimation and inference results on $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}$, $\widehat{\gamma}$ based on logit function and normal regression error. Measurement error variance is 8. The mean, median, empirical standard error, estimated standard error and coverage rate of the 95% confidence intervals are reported.

3.3.2 Simulated Example Two

Our second simulation is designed to reflect the structure of the AIDS data which will be analyzed next. We generated the observations (Z_i, S_i, W_i, Y_i) from the model

$$\begin{aligned} & \mathrm{pr}(Y_i = 1 \mid X_i = x_i, Z_i = z_i) && (3.8) \\ = \; & H\{x_i(\beta_4 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i}) + \beta_{c4} + \beta_{c1} z_{1i} + \beta_{c2} z_{2i} + \beta_{c3} z_{3i}\}, \\ W_i = \; & X_i + U_i, && (3.9) \\ X_i = \; & \alpha_1 + \alpha_2 S_i + \epsilon_i. && (3.10) \end{aligned}$$

Here, $H(\cdot)$ is chosen to be the inverse logit link function. We set $(\alpha_1, \alpha_2) = (1.0, 1.0)$ and $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_{c1}, \beta_{c2}, \beta_{c3}, \beta_{c4}) = (-0.5, 0.6, -0.4, 0.3, 1.0, -1.0, 0.5, -0.5)$. The observable covariates z_{1i}, z_{2i} and z_{3i} are all dichotomous variables, where $z_{1i} = z_{2i} = z_{3i} = 0$ indicates that the ith individual receives the reference treatment (treatment 1) and $z_{ki} = 1$ $(k = 1, 2, 3)$ means that the ith individual receives treatment $k + 1$. For the ith observation, at most one of the three $Z_{ki} (k = 1, 2, 3)$ is 1, and the chances of receiving each of the four treatments are equal. The instrumental variable S_i is generated from the standard normal distribution, and we generated ϵ_i from the normal distribution with mean 0 and variance $S_i^2/8$, and U_i from a normal distribution with mean 0 and variance 0.4.

The simulation results are summarized in Table 3.3. It is evident that all the estimators show little bias. Although there are 10 unknown parameters in the problem, which is a relatively large number, the inference performance of our method is still satisfactory. In particular, the empirical and average estimated standard errors are close to each other, and the coverage rate of the 95% confidence intervals are all around the nominal level. We further conducted the simulation by replacing the

	α_1	α_2	β_1	β_2	β_3
Initial value	1	1	-0.5	0.6	-0.4
median	1.0011	1.0006	-0.5028	0.6029	-0.4064
emp se	0.0227	0.0270	0.1976	0.2319	0.1918
est se	0.0222	0.0264	0.1923	0.2339	0.1889
95% cov	94.1%	94.2%	94.1%	95.1%	95.6%
	β_4	β_{c1}	β_{c2}	β_{c3}	β_{c4}
Initial value	0.3	1.0	-1.0	0.5	-0.5
median	0.3006	1.0247	-0.9934	0.5068	-0.4973
emp se	0.1366	0.2752	0.3325	0.2559	0.1829
est se	0.1368	0.2705	0.3263	0.2645	0.1919
95% cov	95.9%	94.7%	95.7%	96.2%	96.7%

Table 3.3: Simulation 2: Model structure similar to the AIDS data; Estimation and inference results on $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, $\widehat{\beta}_4$, $\widehat{\beta}_{c1}$, $\widehat{\beta}_{c2}$, $\widehat{\beta}_{c3}$, $\widehat{\beta}_{c4}$. The median, empirical standard error, estimated standard error and coverage rate of the 95% confidence intervals are reported.

logit link with a probit link, and observed very similar results, which are omitted here. Since this simulation is designed to have similar structure as the AIDS data, it provides certain confidence in our real data analysis result in the next subsection.

3.3.3 Real Data Analysis

We applied our method on the data set from an AIDS Clinical Trials Group (ACTG) study. This study evaluated four different treatments on HIV infected adults whose CD4 cell counts were from 200 to 500 per cubic millimeter. These four treatments are "ZDV", "ZDV+ddI", "ZDV+ddC" and "ddC", labeled as treatment 1 to 4 in this order. Treatment 1 is a standard treatment hence is considered as the reference treatment; see Hammer, Katzenstein, Hughes, Gundacker, Schooley, Haubrich, Henry, Lederman, Phair, Niu, Hirsch & Merigan (1996), Huang & Wang (2000) and Huang & Wang (2001) for more detailed descriptions of the data set. We included 1036 patients who had no antiretroviral therapy at enrollment in our analysis. We are interested in studying the treatment difference in terms of whether

a patient has his CD4 count drop below 50%, a clinically important indicator for the HIV infected patients, develops AIDS or dies from HIV related disease ($Y = 1$). Thus, our main model is given in (3.8), where Z_{ik} has the same meaning as in the second simulation study. Here, X is the baseline log(CD4 count) prior to the start of treatment and within 3 weeks of randomization. Of course X is not measured precisely, and we use the average of two available measurements as W. From the two repeated measurements, the measurement error variance is estimated as 0.3. In addition, a screening log(CD4 count) is available and is used as the instrumental variable S. The relationship between W and S is depicted in Figure 3.2. Apparently, a linear model will fit the data well. Therefore we assume the relation between W, X and S, Z can be described using (3.9) and (3.10).

We conducted the analysis under both the logit and probit models, but report only the results in the logit model because the probit model yields very similar results. The estimate for (α_1, α_2) is $(0.0001, 0.67)$ with the standard error $(0.02, 0.02)$ using the OLS method. The result from the WLS is very similar. The subsequent estimate of $\boldsymbol{\beta}$ is given in Table 3.4. We further plotted the corresponding relations between the baseline log CD4 counts (X) and the estimated linear function of X under the four treatments in Figure 3.1. Different methods of estimating the $\boldsymbol{\alpha}$ parameter make little difference in the $\boldsymbol{\beta}$ estimation since the estimations from OLS and WLS are themselves very similar. This is reflected in the information in both Table 3.4 and Figure 3.1. As manifested in the plots in Figure 3.1, treatment 1 shows a negative slope, indicating that the standard treatment seems to be more effective for patients with larger baseline CD4, or patients whose situation is less severe. On the contrary, the treatments 2 and 4 show positive slopes, indicating that these treatments are more effective for patients with smaller baseline CD4 counts, or patients with more grave situation.

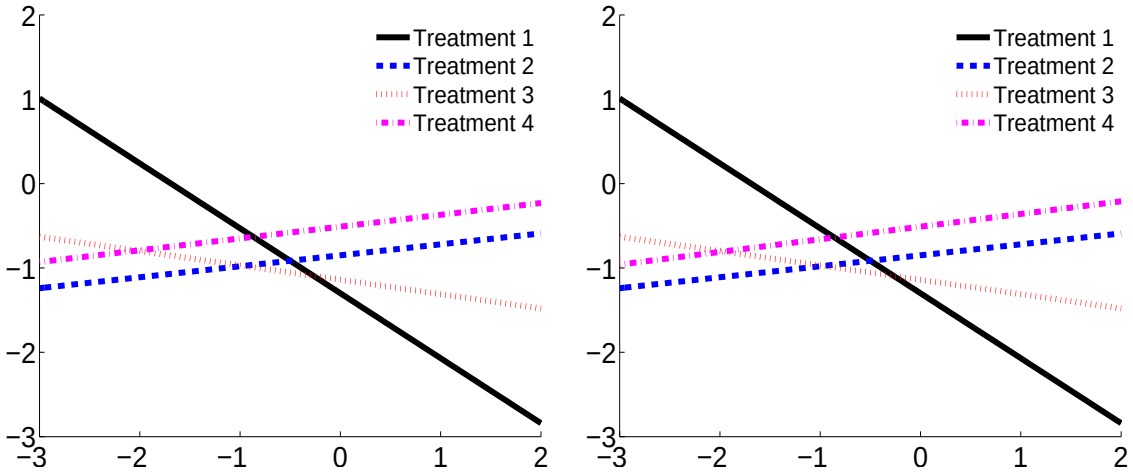

Figure 3.1: Plots of the linear function of x inside the link H in four treatments, where x is the baseline CD4 count in the logarithm scale. The OLS (left) and the WLS (right) methods are used to estimate $\boldsymbol{\alpha}$.

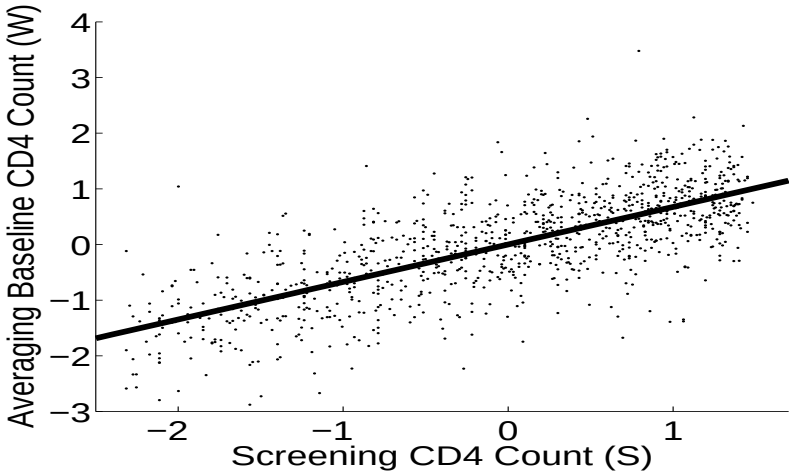

Figure 3.2: Plot of the covariate averaged baseline CD4 count versus the instrument variable screening CD4 count. Unit is "Cells per cubic millimeter". The measurements are on logarithm scales. A straight line is fitted to the scattered points.

		β_1	β_2	β_3	β_4
OLS	Estimate	0.13	-0.17	0.14	-0.77
	two-sided	$(-0.60, 0.86)$	$(-0.95, 0.61)$	$(-0.52, 0.81)$	$(-1.22, -0.31)$
	one-sided	$(-0.48, \infty)$	$(-\infty, 0.49)$	$(-0.41, \infty)$	$(-\infty, -0.39)$
IWLS	Estimate	0.13	-0.17	0.15	-0.77
	two-sided	$(-0.60, 0.86)$	$(-0.96, 0.62)$	$(-0.52, 0.81)$	$(-1.23, -0.31)$
	one-sided	$(-0.49, \infty)$	$(-\infty, 0.49)$	$(-0.41, \infty)$	$(-\infty, -0.39)$
		β_{c1}	β_{c2}	β_{c3}	β_{c4}
OLS	Estimate	-0.85	-1.14	-0.51	-1.30
	two-sided	$(-1.37, -0.32)$	$(-1.72, -0.56)$	$(-1.00, -0.03)$	$(-1.61, -0.98)$
	one-sided	$(-\infty, -0.41)$	$(-\infty, -0.65)$	$(-\infty, -0.10)$	$(-\infty, -1.03)$
IWLS	Estimate	-0.85	-1.14	-0.51	-1.30
	two-sided	$(-1.37, -0.32)$	$(-1.72, -0.56)$	$(-1.00, -0.03)$	$(-1.61, -0.98)$
	one-sided	$(-\infty, -0.41)$	$(-\infty, -0.65)$	$(-\infty, -0.10)$	$(-\infty, -1.03)$

Table 3.4: Analysis of the ACTG 175 data: Estimates, two-sided and one-sided 95% confidence intervals for the model are reported. Results are based on logit model in combination with the OLS and the WLS method respectively for $\boldsymbol{\alpha}$ estimation.

In both the OLS (left plot in Figure 3.1) and the WLS (right plot in Figure 3.1) estimation, the lines from treatment 1 and the other three treatments intercept around $x = -0.5$, corresponding to the baseline CD4 level of 288. Thus, for patients with a baseline CD4 count larger than 288, treatment 1 is probably a good treatment since the corresponding probability of having a $\geq 50\%$ drop of CD4 count is quite small compared to other treatments. On the other hand, if a patient's baseline CD4 count is smaller than 288, there is probably good reason to use the new treatments.

To further confirm our intuitive conclusion from observing the plots, we perform statistical inference regarding the four treatments. Our first attempt is to test the treatment differences between treatment $k, (k = 2, 3, 4)$ and treatment 1. From the second row of Table 3.4, it is clear that at 95% confidence level, all of the three new treatments $(k = 2, 3, 4)$, are significantly different from the standard treatment.

Considering that our original goal of the study is to discover better new treatments $(k = 2, 3, 4)$ than the standard one, we further constructed one-sided con-

fidence intervals. The third row in Table 3.4 summarizes the one-sided confidence intervals. The fact that under both OLS and WLS, β_{c1}, β_{c2} and β_{c3} are significantly smaller than zero suggests that at 95% confidence level, treatments 2, 3 and 4 are better than treatment 1 for severe patients, in that these three treatments decrease the probability of severe CD4 count declination for patients with low baseline CD4 counts. On the other hand, with high baseline CD4 counts, no certain variation in the treatment effect can be declared since the intervals regarding β_1, β_2 and β_3 include zero. In other words, the improvement of the new treatments only applies to patients with low CD4 counts and is more significant if the patients' situation are more grave in terms of their baseline CD4 counts. For patients whose baseline CD4 counts are sufficiently high, the standard treatment could be a preferred choice.

3.4 Discussion

The problem of measurement error arises in real data analysis in many scientific disciplines. Generally speaking, there are two approaches to dealing with this problem. The first approach assumes the distribution of the unobserved covariates or of the measurement error to be known, or can be estimated using replicate data. Therefore this approach has limited applicability in practice. Another approach uses the instrumental variables which are easier to obtain than replicate data. Hence this approach has wider applicability in practice.

Although the instrumental variable approach has been widely used in nonlinear models, its applicability in binary response models is unclear. In this section we demonstrate that this is possible without making any parametric assumption for the distribution of the unobserved variables in the model. In particular, the proposed estimator is fairly efficient under semiparametric setup. The simulation studies show satisfactory performance of the proposed estimator in finite sample situation.

Through combining the relations of the unobservable variable $\mathbf{X}$ with the observed $\mathbf{W}$ and with the instruments $\mathbf{S}$, we establish a direct relation between $\mathbf{W}$ and $\mathbf{S}$, and estimate the parameter $\boldsymbol{\alpha}$ before performing the estimation for the parameter of interest $\boldsymbol{\beta}$. Although Theorem 3 clearly indicates that this estimated $\boldsymbol{\alpha}$ alters the final estimation variability of $\widehat{\boldsymbol{\beta}}$, it is still unclear if such alteration is detrimental or beneficial. The only clear message is that if a true error distribution $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$ is implemented, then the estimation of $\boldsymbol{\alpha}$ causes estimation variance inflation for $\boldsymbol{\beta}$. Overall, how to best handle the estimation of $\boldsymbol{\alpha}$ so that under a same working model $f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z})$, the estimation variability of $\boldsymbol{\beta}$ is minimized is still unknown. Further study is certainly needed.

Although we present our main estimator in the context of logistic or probit models, the method is certainly not restricted only to these contexts. In fact, any regression model of Y conditional on $\mathbf{X}, \mathbf{Z}$ can be handled by our method via a suitable H function. This indicates that Y is also not restricted to binary variables. Thus, for example, the method can readily be extended to generalized linear models.

4. NONPARAMETRIC ESTIMATION OF AGE DISTRIBUTION OF HUNTINGTON'S DISEASE WHEN FAMILIAL CORRELATION EXIST

4.1 Introduction

The Cooperative Huntington's Observational Research Trial (COHORT, Dorsey & the Huntington Study Group COHORT Investigators (2012)) is a kin-cohort study. During the study, patients of the Huntington's disease are genotyped while their relatives are not. Instead, only the survival information of their relatives are collected. This brings challenge in analyzing the relative data, mainly because it is impossible to identify if a relative is a Huntingtin gene mutation carrier or not. This further causes difficulty in characterizing the difference between the Huntingtin gene carrier and non-carrier populations, a key step in understanding the disease hence effectively intervening the disease progress or controlling the damage caused by the disease. Nevertheless, a relative's relation to the proband together with the Mendalian law allows the calculation of the probability of a relative to carry the Huntingtin gene mutation.

Using the probabilities associated with each relative, assuming that the observations are independent, it is possible to study the distribution of any trait of interest for both the mutation carrier population and non-carrier population. This has been extensively studied in Ma & Wang (2012) and efficient estimation procedures have been developed. When data are also subject to censoring, Ma & Wang (2013) further developed effective methods to perform the estimation and inference. However, both works have assumed that the observations concerning the relatives are independent, hence it is unclear how to properly handle the possible correlation, likely to exist among relatives from a same family.

It is the goal of this work to address the within family correlation between observations when the data are subject to censoring and their population identifiers are only know up to the probabilities. Because the correlation with a family can be a result of similar life style or similar biological elements other than the gene under study, it is very difficult to quantify or even model such correlation. In addition, different family has different sizes. Due to these considerations, we leave the within family correlation completely unspecified. The attractiveness of our method lies in its novelty, its simplicity and its flexibility. We first eliminate the effect of the within family correlation by using only one member per family to form our base estimator. We then devise an optimal way to synthesize a new estimator that takes advantage of the multiple members in the families. To the best of our knowledge, no such treatment has been considered in the literature and the idea is completely new. When forming the base estimator, we use the method by Ma & Wang (2013), which is simple and practically as effective as the efficient estimator. This results a final procedure that is also straightforward to implement. Finally, our method is able to handle arbitrary distribution function forms and arbitrary correlation structures with the need to tune or adapt any part of the method. This makes the method extremely flexible and user friendly.

The rest of the section is organized as the following. We illustrate the method, describe it implementation and demonstrate its optimality property in section 4.2. Simulations are carried out in section 4.3 to illustrate the performance of the estimators in both simple and complex settings. Finally, we analyse the COHORT data which motivated this work in section 4.4 and conculde the section with some discussions in section 4.5. All the technical derivations are in appendix.

4.2 Methodology

We first define some notations. Suppose there are N families in the study, and the ith family has n_i members, $i = 1, \ldots, N$. The survival time for the jth member of the ith family is S_{ij}, where S_{ij} is a random event time. The event is subject to censoring at time C_{ij}. Let $Y_{ij} = \min(S_{ij}, C_{ij})$ and the censoring indicator $\delta_{ij} = I(S_{ij} \leq C_{ij})$. Furthermore, we assume there are p different populations, whose event time has cumulative distribution functions $F_1(t), F_2(t), \ldots, F_p(t)$ respectively. Write $\mathbf{F}(t) = \{F_1(t), F_2(t), \ldots, F_p(t)\}^{\mathrm{T}}$. Assume for all $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, S_{ij} is a random sample from one of the p populations, although we do not know which population it is. We use q_{ijk} to denote the probability of the event S_{ij} belonging to the kth population, for $k = 1, \ldots, p$. Let $\mathbf{q}_{ij} = (q_{ij1}, q_{ij2}, \ldots, q_{ijp})^{\mathrm{T}}$. Obviously, $\sum_{k=1}^{p} q_{ijk} = 1$. We assume the p distribution processes are independent of the censoring process. Using these notations, the observed data can be written as $\mathbf{O} = \{(\mathbf{q}_{ij}, Y_{ij}, \delta_{ij}),\ i = 1, \ldots, N,\ j = 1, \ldots, n_i\}$.

In kin-cohort and QTL studies, there are only finitely many, say m, $m < \infty$, possible values $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ for $\mathbf{q}_{ij}$. We count the frequencies of the occurrences of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ and record them as $d_1, d_2, \ldots, d_m$ with $\sum_{i=1}^{m} d_i = \sum_{i=1}^{N} n_i$. The mixture distributions are defined as $\mathbf{H}(t) = \{H_1(t), H_2(t), \ldots, H_m(t)\}^{\mathrm{T}}$, where $H_l(t) = \mathbf{u}_l^{\mathrm{T}}\mathbf{F}(t)$ for $l = 1, \ldots, m$.

4.2.1 Special Configuration

We consider a special case where the observations $\mathbf{O} = \{(\mathbf{q}_{ij}, Y_{ij}, \delta_{ij}),\ i = 1, \ldots, N,\ j = 1, \ldots, n_i\}$ are independent of each other. Obviously, this happens when each family has only one observation, i.e. $n_i = 1$ for $i = 1, \ldots, N$. This also happens when there is no within family correlation. In this case, following Ma &

Wang (2013), we use the relation

$$\mathbf{F}(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l H_l(t) \right\}, \tag{4.1}$$

and estimate $\mathbf{F}(t)$ through

$$\widehat{\mathbf{F}}(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l \widehat{H}_l(t) \right\}. \tag{4.2}$$

Here $\widehat{H}_l(t)$ is a Kaplan-Meier (KM) estimate (Kaplan & Meier (1958)) for $H_l(t)$. Kaplan & Meier (1958) has established the consistency for the KM estimator, while Breslow & Crowley (1974) has shown that it converges weakly to a Gaussian process. Since $\widehat{\mathbf{F}}(t)$ is a linear transformation of $\widehat{\mathbf{H}}(t)$, it is also consistent and converges weakly to a Gaussian process when $N \to \infty$. These observations will be used in our following derivation when within family correlation exists.

4.2.2 Resampled and Bootstrapped Linear Combination Estimator (RBLCE)

In the general situation when members from a same family may be correlated, we propose a two stage procedure that utilizes the results described in Section 4.2.1. In the first stage, we randomly sample one member from each family, regardless of the family size, and then use (4.2) to obtain a crude estimation of $\mathbf{F}(t)$. Repeat this process multiple, say R, times to collect multiple estimators for $\mathbf{F}(t)$, denoting these estimators $\widehat{\mathbf{F}}^1(t), \ldots, \widehat{\mathbf{F}}^R(t)$. In the second stage, we aim to combine the multiple estimators from the first stage in an optimal way.

Since each $\widehat{\mathbf{F}}^r(t), r = 1, \ldots, R$ is a consistent estimator of $\mathbf{F}(t)$, it is natural to use a weighted average of these estimators to form an estimator that is also consistent

and hopefully more efficient. In general, we write the combined estimator

$$\widehat{\mathbf{F}}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t), \tag{4.3}$$

where

$$\widehat{\mathbf{F}}_L(t) = \left[\{\widehat{\mathbf{F}}^1(t)\}^{\mathrm{T}}, \{\widehat{\mathbf{F}}^2(t)\}^{\mathrm{T}}, \ldots, \{\widehat{\mathbf{F}}^R(t)\}^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{4.4}$$

and $\mathbf{A}$ is a $p \times pR$ weight matrix. The consistent requirement mandates $\mathbf{AJ} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the size p identity matrix, and $\mathbf{J}$ is a $pR \times p$ matrix formed by $\mathbf{I}_p$'s, i.e. $\mathbf{J} = (\mathbf{I}_p, \ldots, \mathbf{I}_p)^{\mathrm{T}}$. In appendix, we further show that the optimal choice of $\mathbf{A}$ in terms of minimizing the variance of $\widehat{\mathbf{F}}(t)$ is $(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}$, where $\mathbf{U}$ is the asymptotic variance-covariance matrix of $\sqrt{N}\widehat{\mathbf{F}}_L(t)$. We summarize the above results in Theorem 1.

Theorem 4. *Let $\widehat{\mathbf{F}}(t)$ be as in (4.3). Then as long as $\mathbf{AJ} = \mathbf{I}_p$, $\widehat{\mathbf{F}}(t)$ is consistent. In addition, var$\{\widehat{\mathbf{F}}(t)\}$ is minimized when*

$$\mathbf{A}_{\mathrm{opt}} = (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}. \tag{4.5}$$

The resulting optimal variance of $\sqrt{N}\widehat{\mathbf{F}}(t)$ is

$$\mathbf{V}_1^{\mathrm{opt}} = (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}.$$

To take advantage of the result in Theorem 4, we still need to obtain $\mathbf{U}$. Because of our construction of $\widehat{\mathbf{F}}(t)$, $\mathbf{U}$ is naturally an $R \times R$ block matrix with each block size $p \times p$. Although the diagonal blocks of $\mathbf{U}$ can be approximated using results in Section 4.2.1, the analysis of the off-diagonal blocks is impossible because of the

unspecified correlation structure among family members and the potentially complex pattern resulting from the sampling procedure. Thus, we resort to a bootstrap procedure (Efron (1981) and Akritas (1986)) to assess $\mathbf{U}$. However, caution needs to be taken in performing the bootstrap procedure, which is somewhat different from the obvious. In particular, although our interest is to repeatedly draw family members to form estimators, we need to bootstrap families, not members of the families. Specifically, for $b = 1, \ldots, B$, we randomly draw N families with replacement and with equal probability, and denote the bootstrap sample $\mathbf{O}_b^*$. We then repeat the estimation procedure described above on $\mathbf{O}_b^*$ to obtain $\widehat{\mathbf{F}}_L^{*b}(t)$. The sample variance of $\widehat{\mathbf{F}}_L^{*1}(t), \widehat{\mathbf{F}}_L^{*2}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$ is then used to estimate $\mathbf{U}$.

The complete procedure of our RBLCE is the following.

Algorithm 1.

Step 1. Randomly draw one member from each family, assume the resulting sample contains m different $\mathbf{q}$ values, written as $\mathbf{u}_1, \ldots, \mathbf{u}_m$, with frequency $d_1, \ldots, d_m$. Form

$$\widehat{\mathbf{F}}^r(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l \widehat{H}_l(t) \right\}.$$

Step 2. Repeat Step 1 R times ($r = 1, \ldots, R$), and form $\widehat{\mathbf{F}}_L(t)$ using (4.4).

Step 3. Randomly sample N families with replacement from the original families.

Step 4. Perform Steps 1 and 2 on the sampled data, obtain the corresponding $\widehat{\mathbf{F}}_L^{*b}(t)$.

Step 5. Repeated Steps 3 and 4 B times ($b = 1, \ldots, B$) to obtain $\widehat{\mathbf{F}}_L^{*1}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$.

Step 6. Calculate the sample variance $\widehat{\mathbf{U}}$ of $\widehat{\mathbf{F}}_L^{*1}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$.

Step 7. Form the estimator $\widehat{\mathbf{F}}(t) \equiv (\mathbf{J}^{\mathrm{T}}\widehat{\mathbf{U}}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\widehat{\mathbf{U}}^{-1}\widehat{\mathbf{F}}_L(t)$.

4.2.3 Resampled and Bootstrapped Quadratic Inference Function Estimator (RBQIF)

Considering that using each one-member-per-family data sampled from the original data, we can collect, in the general case, a set of estimating equations, it is also quite natural to combine different estimating equations to obtain a final estimator. Specifically, when there is only one member per family, we rewrite the relation in (4.1) as

$$\sum_{i=1}^{N} \left\{ \mathbf{q}_{i1} H_i(t) - \mathbf{q}_{i1}\mathbf{q}_{i1}^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}, \tag{4.6}$$

and view $\widehat{\mathbf{F}}(t)$ as the root that solves

$$\sum_{i=1}^{N} \left\{ \mathbf{q}_{i1} \widehat{H}_i(t) - \mathbf{q}_{i1}(\mathbf{q}_{i1})^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}, \tag{4.7}$$

where $\widehat{H}_i(t)$ is the same KM estimator as before.

We use the sampling scheme in the first stage of RBLCE in section 4.2.2 to sample R data sets, and write the estimating equation (4.7) based on the rth sampled data $\sum_{i=1}^{N} \mathbf{g}_i^r(t) = \sum_{i=1}^{N} \left\{ \mathbf{q}_i^r \widehat{H}_i^r(t) - \mathbf{q}_i^r (\mathbf{q}_i^r)^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}$, $r = 1, \ldots, R$. Here, $\mathbf{q}_i^r$ denotes the $\mathbf{q}$ value of the member from the ith family in the rth sample. Because the number of equations, pR, can be much larger than the number of the parameters p, we resort

to the Quadratic Inference Function (QIF) method (Lindsay & Qu (2003)). Write

$$\sum_{i=1}^{N} \mathbf{g}_i(t) = \sum_{i=1}^{N} \begin{Bmatrix} \mathbf{g}_i^1(t) \\ \mathbf{g}_i^2(t) \\ \vdots \\ \mathbf{g}_i^R(t) \end{Bmatrix} = \sum_{i=1}^{N} \begin{Bmatrix} \mathbf{q}_i^1 \widehat{H}_i^1(t) - \mathbf{q}_i^1 (\mathbf{q}_i^1)^{\mathrm{T}} \mathbf{F}(t) \\ \mathbf{q}_i^2 \widehat{H}_i^2(t) - \mathbf{q}_i^2 (\mathbf{q}_i^2)^{\mathrm{T}} \mathbf{F}(t) \\ \vdots \\ \mathbf{q}_i^R \widehat{H}_i^R(t) - \mathbf{q}_i^R (\mathbf{q}_i^R)^{\mathrm{T}} \mathbf{F}(t) \end{Bmatrix}, \tag{4.8}$$

We minimize the quadratic form,

$$\left\{ \sum_{i=1}^{N} \mathbf{g}_i(t) \right\}^{\mathrm{T}} \mathbf{W} \left\{ \sum_{i=1}^{N} \mathbf{g}_i(t) \right\} \tag{4.9}$$

for a weight matrix $\mathbf{W}$. In typical QIF construction, $\mathbf{g}_i(t)$'s are functions of the ith observation respectively and hence are independent. Therefore root-N consistency, asymptotic normality, etc. has been established in Lindsay & Qu (2003). However here, it is important to recognize that $\mathbf{g}_i(t)$'s are not independent since they contain $\widehat{H}_i^r(t)$'s, which are estimated based on all the observations from the rth sample for $r = 1, \ldots, R$. Nevertheless, in Theorem 5, we show that the resulting estimator still enjoys the usual asymptotic normality property. The proof is in appendix.

Theorem 5. *Let $\widehat{\mathbf{F}}(t)$ be the minimizer of the quadratic form in (4.9). Then $\sqrt{N}\{\widehat{\mathbf{F}}(t) - \mathbf{F}(t)\} \to$ Normal$(\mathbf{0}, \mathbf{V}_2)$ in distribution when $N \to \infty$, where $\mathbf{V}_2$ is a $p \times p$ positive-definite matrix. Let $\mathbf{M}$ be the asymptotic variance-covariance matrix of $N^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbf{g}_i(t)$. When $\mathbf{W}_{\mathrm{opt}} = \mathbf{M}^{-1}$, $\sqrt{N}\{\widehat{\mathbf{F}}(t) - \mathbf{F}(t)\}$ achieves the efficiency bound*

$$\mathbf{V}_2^{\mathrm{opt}} = \left[E \left\{ \frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)} \right\}^{\mathrm{T}} \mathbf{M}^{-1} E \left\{ \frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)} \right\} \right]^{-1}. \tag{4.10}$$

Theorem 5 prescribes the choice of the optimal weight matrix. To achieve effi-

ciency, it is essential to estimate $\mathbf{M}$. Because no correlation structure is modeled for members from the same family, we resort to the bootstrap procedure mentioned in Section 4.2.2 to approximate $\mathbf{M}$. Using the bth bootstrap sample $\mathbf{O}_b^*$, we follow the procedure described above to construct estimation equation $\sum_{i=1}^N \mathbf{g}_i^{*b}(t)$. The sample variance of $\sum_{i=1}^N \mathbf{g}_i^{*1}(t), \ldots, \sum_{i=1}^N \mathbf{g}_i^{*B}(t)$ is then used to estimate $\mathbf{M}$.

The detailed algorithm based on RBQIF is the following.

Algorithm 2.

Step 1. Randomly draw one member from each family. Form

$$\sum_{i=1}^N \mathbf{g}_i^r(t) = \sum_{i=1}^N \left\{ \mathbf{q}_i^r \widehat{H}_i^r(t) - \mathbf{q}_i^r (\mathbf{q}_i^r)^{\mathrm{T}} \mathbf{F}(t) \right\}.$$

Step 2. Repeat Step 1 R times ($r = 1, \ldots, R$), and form $\sum_{i=1}^N \mathbf{g}_i(t)$ using (4.8).

Step 3. Randomly sample N families with replacement from the original families.

Step 4. Perform Steps 1 and 2 on the sampled data, obtain the corresponding $\sum_{i=1}^N \mathbf{g}_i^{*b}(t)$.

Step 5. Repeated Steps 3 and 4 B times ($b = 1, \ldots, B$) to obtain $\sum_{i=1}^N \mathbf{g}_i^{*1}(t), \ldots, \sum_{i=1}^N \mathbf{g}_i^{*B}(t)$.

Step 6. Calculate the sample variance $\widehat{\mathbf{M}}$ of $\sum_{i=1}^N \mathbf{g}_i^{*1}(t), \ldots, \sum_{i=1}^N \mathbf{g}_i^{*B}(t)$. Let $\mathbf{W} = \widehat{\mathbf{M}}^{-1}$.

Step 7. Obtain the estimator $\widehat{\mathbf{F}}(t)$ from minimizing (4.9).

4.2.4 Equivalence of the Two Methods

To understand the advantages and disadvantages of RBLCE and RBQIF introduced respectively in Section 4.2.2 and 4.2.3, we perform further analysis to compare their relative performance. Given that RBLCE is a combination of the estimators

from R samples, while RBQIF results from solving a combination of estimating equations from the same R samples, it is not surprising that these two procedures are in fact equivalent. In the following, we formally establish that there is a one-to-one mapping between the estimators in the two classes, and in particular, the optimal estimation variances from the two estimators are identical asymptotically.

Because the RBLCE is uniquely decided by the weight matrix choice $\mathbf{A}$ while the RBQIF is uniquely decided by the weight matrix $\mathbf{W}$, we only need to establish the one-to-one mapping between $\mathbf{A}$ and $\mathbf{W}$ in order to show our results. Define a $pR \times pR$ block diagonal matrix

$$\mathbf{D} = \text{diag}\left[\left\{E\left(\mathbf{q}_{ij}\mathbf{q}_{ij}^{\mathrm{T}}\right)\right\}^{-1}, \ldots, \left\{E\left(\mathbf{q}_{ij}\mathbf{q}_{ij}^{\mathrm{T}}\right)\right\}^{-1}\right].$$

For any weight matrix $\mathbf{W}$ defined in the RBQIF estimator, consider

$$\mathbf{A} = (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1} \tag{4.11}$$

as the weight matrix in RBLCE. Obviously, $\mathbf{A}\mathbf{J} = \mathbf{I}_p$. We now investigate the resulting RBLCE and RBQIF from the corresponding $\mathbf{A}$ and $\mathbf{W}$. Let the RBLCE estimator $\widehat{\mathbf{F}}^{(1)}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t)$, where $\widehat{\mathbf{F}}_L(t)$ is defined in (4.4). Define $\mathbf{F}_L(t)$ analogously as $\widehat{\mathbf{F}}_L(t)$ and recall the definition of $\mathbf{g}_i(t)$ in (4.8). We can write

$$\sqrt{N}\{\widehat{\mathbf{F}}_L(t) - \mathbf{F}_L(t)\} = N^{-1/2}\mathbf{D}\sum_{i=1}^{N}\mathbf{g}_i(t) + o_p(1), \tag{4.12}$$

which leads to

$$\widehat{\mathbf{F}}^{(1)}(t) = \mathbf{A}\mathbf{D}N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(t) + \mathbf{F}(t) + o_p(N^{-1/2}). \tag{4.13}$$

On the other hand, the RBQIF, denoted $\widehat{\mathbf{F}}^{(2)}(t)$, is obtained from minimizing (4.9), thus standard Taylor expansion leads to

$$\begin{aligned}
& \widehat{\mathbf{F}}^{(2)}(t) \\
= & -\left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W} N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(t) \\
& + \mathbf{F}(t) + o_p(N^{-1/2}) \\
= & (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W} N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(t) + \mathbf{F}(t) + o_p(N^{-1/2}) \qquad (4.14)
\end{aligned}$$

where the last equality follows from the relation

$$E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\} = -1_R \otimes E\left\{\mathbf{q}_{ij}(\mathbf{q}_{ij})^{\mathrm{T}}\right\} = -\mathbf{D}^{-1}\mathbf{J}.$$

Further using the connection between $\mathbf{A}$ and $\mathbf{W}$ in (4.11), we immediately have $\widehat{\mathbf{F}}^{(1)}(t) = \widehat{\mathbf{F}}^{(2)}(t) + o_p(N^{-1/2})$. Conversely, if $\widehat{\mathbf{F}}^{(1)}(t) = \widehat{\mathbf{F}}^{(2)}(t) + o_p(N^{-1/2})$, subtraction of (4.14) from (4.13) yields (4.11).

Having established the one-to-one mapping between RBLCE and RBQIF via (4.11), it is not surprising to expect that the optimal weight matrix choices in the two estimator classes, $\mathbf{A}_{\mathrm{opt}}$ and $\mathbf{W}_{\mathrm{opt}}$, also satisfies (4.11). This can be easily verified through using the equality $\mathbf{U} = \mathbf{DMD}$, which follows from (4.12). Furthermore, we can explicitly verify that the two optimal asymptotic estimation variances are identical, i.e.

$$\begin{aligned}
\mathbf{V}_1^{\mathrm{opt}} &= (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} = (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{D}^{-1}\mathbf{J})^{-1} \\
&= \left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{M}^{-1} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1} = \mathbf{V}_2^{\mathrm{opt}}.
\end{aligned}$$

4.3 Simulation

We now demonstrate the finite sample performance of the RBLCE and RBQIF methods via two simulation studies. The first simulation is a relatively simple one. We use it to illustrate the effectiveness of the theoretical properties derived in Section 4.2. In the second simulation, we generated data following the similar pattern as the Huntington's Disease data studied in Section 4.4, hence it represents quite realistic scenario. Throughout both simulations, we repeated 1000 simulations.

In the first simulation, we set the sample size $N = 1000$, $p = 2$, $m = 3$ and $n_i = 2$ for $i = 1, \ldots, N$. The two ($p = 2$) true functions $F_1(t)$ and $F_2(t)$ are respectively the distribution functions of two normal density, $N(4.0, 1.0^2)$ and $N(6.0, (5/3)^2)$. To generate correlated survival times for members from a same family, we implement the following procedure. For the ith family, we construct a multivariate normal distribution with mean $(4.0, 4.0, 6.0, 6.0)$ and a randomly-generated positive-definite matrix with diagonal $(1.0^2, 1.0^2, (5/3)^2, (5/3)^2)$. We then generate a random vector $(S^1_{i1}, S^1_{i2}, S^2_{i1}, S^2_{i2})$ from this multivariate normal distribution. It provides the jth member of the ith family with possible survival time S^1_{ij} or S^2_{ij}, corresponding to the two functions $F_1(t)$ and $F_2(t)$. We select $S_{ij} = S^1_{ij}$ or S^2_{ij} with probabilities in the $\mathbf{q}_{ij}$ vector, where $\mathbf{q}_{ij}$ is randomly assigned to three ($m = 3$) different vector values $(0.25, 0.75)^{\mathrm{T}}$, $(0.75, 0.25)^{\mathrm{T}}$ and $(0.5, 0.5)^{\mathrm{T}}$, with probabilities 0.4, 0.55 and 0.05 respectively. Lastly, we generate the censoring time from a uniform distribution on $(0, 9.9)$, resulting in a censoring rate of 50% approximately. We then create $Y_{ij} = \min(S_{ij}, C_{ij})$ and $\delta_{ij} = I(S_{ij} \leq C_{ij})$.

We use Algorithms 1 and 2 to carry out the RBLCE and RBQIF methods, and considered $R = 1, 2, 3$ and 4. We implemented $B = 1000$ bootstrap repetitions to estimate the variance-covariance matrices $\mathbf{U}$ in RBLCE and $\mathbf{M}$ in RBQIF respec-

		RBLCE		RBQIF	
		$F_1^{(1)}(t)$	$F_2^{(1)}(t)$	$F_1^{(2)}(t)$	$F_2^{(2)}(t)$
	true	0.1842	0.6915	0.1842	0.6915
$R=1$	mean	0.1855	0.6907	0.1855	0.6907
	emp se	0.0393	0.0479	0.0393	0.0479
	mse	0.0032	0.0047	0.0032	0.0047
	est se	0.0408	0.0488	0.0408	0.0488
	95% cov	95.6%	94.8%	95.5%	94.8%
$R=2$	mean	0.1837	0.6925	0.1837	0.6925
	emp se	0.0354	0.0412	0.0354	0.0412
	mse	0.0025	0.0035	0.0025	0.0035
	est se	0.0352	0.0422	0.0352	0.0422
	95% cov	95.4%	95.1%	95.5%	95.3%
$R=3$	mean	0.1846	0.6899	0.1845	0.6900
	emp se	0.0334	0.0408	0.0334	0.0407
	mse	0.0022	0.0032	0.0022	0.0032
	est se	0.0408	0.0488	0.0408	0.0488
	95% cov	94.7%	93.5%	94.3%	93.5%
$R=4$	mean	0.1815	0.6935	0.1814	0.6936
	emp se	0.0321	0.0385	0.0322	0.0385
	mse	0.0021	0.0030	0.0021	0.0030
	est se	0.0321	0.0385	0.0321	0.0385
	95% cov	94.8%	95.9%	95.0%	95.7%

Table 4.1: Simulation study 1. The mean of the estimates (mean), empirical standard error (emp se), mean square error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

tively. We summarize the results of our analysis at $t = 4.5$ in Table 4.1. From Table 4.1, it is clear that Algorithms 1 and 2 produce very similar results across all choices of R. This fact concurs with our theoretical results on the asymptotic equivalence of the two methods in Section 4.2.4. In addition, the mean of the 1000 estimates are very close to the true function values, the sample standard deviations of the 1000 estimates are very close to the average of the estimated standard deviations, and coverage rate of the 95% confidence interval is indeed close to the nominal value. When R is increased from 1 to 2, the mean squared error (MSE) is reduced by 22%

for $F_1^{(1)}(t)$ and 26% for $F_1^{(1)}(t)$, indicating large improvement in estimation. When R is increased to 3, MSE is further reduced by 12% and 9% respectively. When we increase R to 4, the improvement on MSE is 5% and 6%. Since the last improvement is rather small, we did not further increase R. The estimation result of the entire functions $F_1(t)$ and $F_2(t)$ is given in the left panel of Figure 4.1, where the mean estimated curves almost overlap with the true curves.

In the second simulation, we set sample size $N = 750$, $m = 3$, $p = 2$ and generate the family sizes $n_i, i = 1, \ldots, N$ from the same distribution as in the COHORT data in Section 4.4. The distribution of the n_i's is described via a barplot in Figure 4.2. We set $F_1(t)$ as the distribution function of a normal distribution $N(64, (43/3)^2)$ and $F_2(t)$ as that of a skew-normal distribution (Genton (2004)) with location 93.5, scale 20 and shape parameter -50. Our selection of $F_1(t)$ and $F_2(t)$ results similar curves as those obtained from the data analysis in Figure 4.3. As in simulation 1, we generate correlated survival time for members in a same family from a multivariate distribution. Specifically, for the ith family, we first generate a vector $(S_{i1}^1, \ldots, S_{in_i}^1, S_{i1}^2, \ldots, S_{in_i}^2)$ from a Clayton copula with parameter 2. The survival time S_{ij} of the jth member in the ith family is then assigned to S_{ij}^1 or S_{ij}^2, corresponding to $F_1(t)$ and $F_2(t)$, with probability q_{ij1} or q_{ij2} respectively. Here $\mathbf{q}_{ij} = (q_{ij1}, q_{ij2})^{\mathrm{T}}$ is set to be $(1.0, 0.0)^{\mathrm{T}}$, $(0.5, 0.5)^{\mathrm{T}}$ or $(0.0, 1.0)^{\mathrm{T}}$, with probability 0.15, 0.5 or 0.35 respectively. We generate censoring time from a uniform distribution on $(0, 100)$, resulting in a censoring rate around 29%. Finally we let $Y_{ij} = \min(S_{ij}, C_{ij})$ and $\delta_{ij} = I(S_{ij} \leq C_{ij})$.

We implement Algorithm 1 (RBLCE) and 2 (RBQIF) with $R = 1, 2, 3$ and 4. We use $B = 1000$ bootstraps to estimate $\mathbf{U}$ and $\mathbf{M}$. The simulation results at $t = 55$ are summarized in Table 4.2 and the right panel of Figure 4.1 depicts the estimated distribution curves and their confidence bands. From Table 4.2, we find that MSE

		RBLCE		RBQIF	
		$F_1^{(1)}(t)$	$F_2^{(1)}(t)$	$F_1^{(2)}(t)$	$F_2^{(2)}(t)$
	true	0.2650	0.0542	0.2650	0.0542
$R=1$	mean	0.2651	0.0536	0.2651	0.0536
	emp se	0.0420	0.0204	0.0420	0.0204
	mse	0.0035	0.0016	0.0035	0.0016
	est se	0.0418	0.0206	0.0420	0.0207
	95% cov	96.10%	95.20%	95.70%	95.60%
$R=2$	mean	0.2639	0.0536	0.2640	0.0535
	emp se	0.0343	0.0170	0.0344	0.0169
	mse	0.0024	0.0013	0.0024	0.0013
	est se	0.0345	0.0171	0.0345	0.0171
	95% cov	95.10%	94.60%	94.90%	94.60%
$R=3$	mean	0.2643	0.0541	0.2643	0.0541
	emp se	0.0317	0.0153	0.0316	0.0153
	mse	0.0020	0.0013	0.0020	0.0013
	est se	0.0316	0.0157	0.0315	0.0157
	95% cov	94.70%	95.30%	94.30%	95.10%

Table 4.2: Simulation study 2. Performance of two algorithms with $t = 55$. The mean of the estimates (mean), empirical standard error (emp se), mean square error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

decreases 31% and 19% from $R = 1$ to 2, and further decreases 21% for the first parameter when R increases to 3. We omit the results from $R = 4$ since it is very similar to the one from $R = 4$.

4.4 Data Example

We apply our methods to analyze the COHORT data which motivated this work. The data set includes 771 families with different numbers of members within each family. There are a total of 3661 observations. The barplot in Figure 4.2 characterizes the distribution of the family sizes. Using the available relationship (parents, children, siblings etc.) between each family member and his/her proband, we calculated the probability of the member to carry the Huntingtin gene mutation and

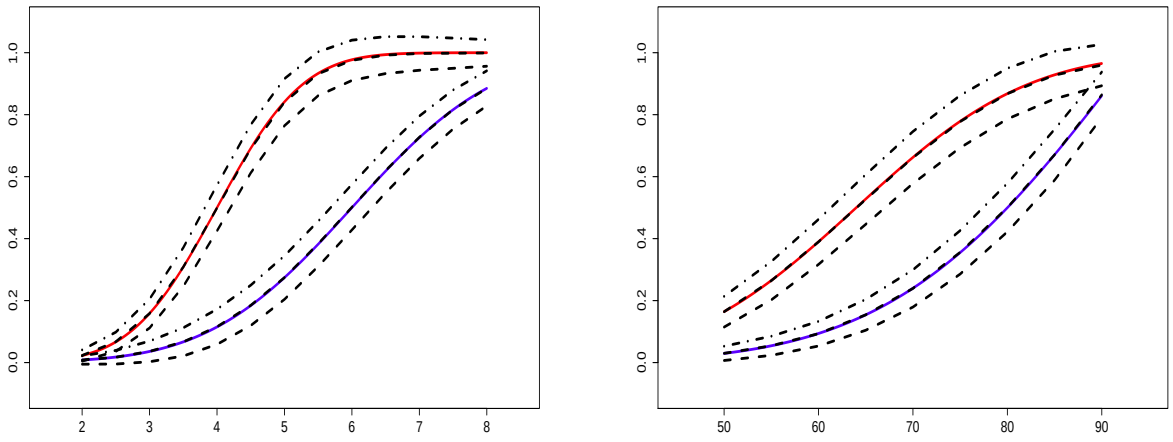

Figure 4.1: Simulation study on $F_1(t)$ and $F_2(t)$. True CDFs (solid) and mean (dashed), 95% confidence band (upper band dot-dashed, lower band dashed) of the estimated CDFs. Left: simulation 1. Right: simulation 2.

not carry the mutation. We obtained three ($m = 3$) different $\mathbf{q}_{ij}$ values in total, $(1.0, 0.0)^{\mathrm{T}}$, $(0.5, 0.5)^{\mathrm{T}}$ and $(0.0, 1.0)^{\mathrm{T}}$, with frequency 558, 1805 and 1298 respectively. Write the survival time of mutation carrier population have distribution function $F_1(t)$ and the non-carrier group $F_2(t)$. Our goal is to estimate $\mathbf{F}(t) = \{F_1(t), F_2(t)\}^{\mathrm{T}}$. The COHORT data has approximately 29% censoring, mainly due to administration in the data collection procedure or early data collection. Thus, we assume the censoring time is independent of the event time.

We implemented both Algorithm 1 (RBLCE) and Algorithm 2 (RBQIF) developed in section 4.2. We performed $B = 500$ bootstraps to estimate the variance-covariance $\mathbf{U}$ in RBLCE and $\mathbf{M}$ in RBQIF. The results corresponding to $R = 120$ are given in left panel of Figure 4.3, where the estimated $F_1(t)$ and $F_2(t)$, and their 95% confidence bands are provided. It is clear that the huntingtin gene mutation carriers have much smaller survival rates than non-carriers, especially in the age range 50 to 90. This indicates that the detrimental effect of the Huntington's disease on survival

is most severe in the middle to old age range. This is possibly because the disease takes time to progress so it does not cause early death very often, while for those who live beyond 90, various other causes of death are sufficiently grave to compete or even out perform the Huntington's disease. For comparison, we also perform the analysis of Ma & Wang (2013), where the within family correlation is ignored. We present the estimated curves of $F_1(t)$ and $F_2(t)$, and their 95% pointwise confidence bands in the right panel of Figure 4.3. It is clear that the 95% confidence bands are wider when the observations are treated as independent. These observations indicate that within family correlation indeed exists in the COHORT data and estimation efficiency is indeed improved if we take into account of the correlation and properly handle it.

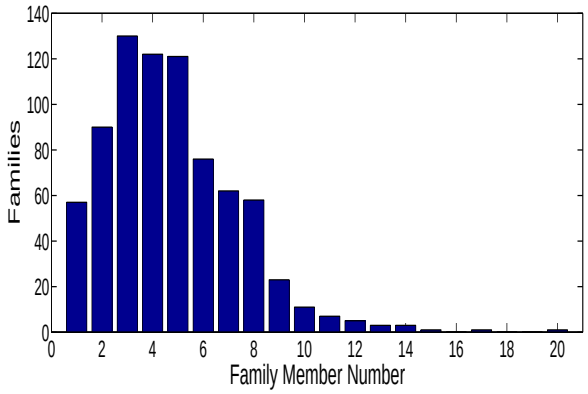

Figure 4.2: Huntington's Disease family members' distribution in Barplot. The highest percentage 16.86% happens when $n_i = 3$. The largest family has $n_i = 20$ members with the smallest percentage 0.13%.

4.5 Discussion

In this section, we devised resample and bootstrap based methods to explore within family correlation for mixed data from multiple populations, while the population label is only known to a probability. Such data frequently arise from Kin-cohort studies, such as COHORT study of Huntington's disease. The estimators in

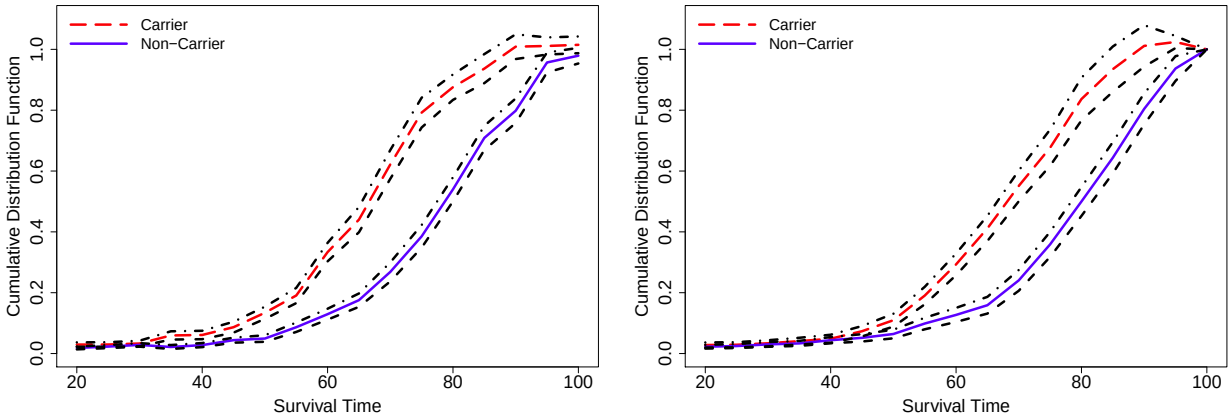

Figure 4.3: Distribution of the survival time for gene mutation carriers and non-carriers in Huntington's Disease study: estimated CDFs (solid) and the 95% confidence band (upper band dot-dashed, lower band dashed). Left: Treat within-family correlation; Right: Ignore within-family correlation.

the family is easy to implement, while the optimal estimators relies on proper choices of the weight matrices, which we propose to estimate via bootstrap. The estimators are applicable for longitudinal studies, where only marginal model is required. The finite sample efficiency of the estimators relies on both the number of resamples and the bootstrap size. Although in theory, large values of both are preferable, in practice, one can always gradually increase these values and stop when the improvement becomes sufficiently small.

5. CONCLUSIONS

The prevalence of semiparametric regression models is clearly seen in the previous sections where we discuss capture-recapture models, instrumental variable regression with binary response and mixed data in kin-cohort longitudinal studies. A class of semiparametric estimators are derived for the first two problems by using a semiparametric treatment (Tsiatis (2006)). Two resampled and bootstrapped methods are developed for the last study. All estimators are asymptotically consistent and optimal. During the process of deriving those estimators, we have some unique findings regarding each model.

Hwang, Huang & Wang (2007) studies the measurement error problem in the literature of capture-recapture models. It corrects the estimation bias by implementing a conditional score method. However, it completely ignores the multiple measurements of the covariates of recaptures, thus leading to a waste of information. We propose firstly a GMM method to combine estimating equations for observations with different times of captures. This way of utilizing information is not as efficient as constructing estimating equations directly for $\overline{\mathbf{W}}$. Therefore, the improvement of estimation efficiency is significant only when there are large samples and capture probabilities are high. Both of the scenarios are unlikely to happen in real life. In order to solve the problem, we implement a semiparametric treatment to use the averaged information directly because there are no complete sufficient statistics existing. An exciting finding is that the semiparametric treatment actually provides us with an approach to deal with measurement error models without surrogacy assumption. Simulations show the superior performance of the semiparametric method. We have successfully utilized multiple information in capture-recapture models. In the

future we can improve our work by proposing a better population estimator than that in equation (2.4). In addition, there is room for research of taking advantage of multiple measurements in open population problems.

The work of semiparametric method in instrumental variable regression fills the gap of generalized linear models with measurement errors in variables. Because there are no distribution assumptions on both the true covariates and measurement errors, our work can be implemented to a wider range of statistical problems. Not surprisingly, the measurement error problem is bypassed due to a configuration of prediction relationship of instrumental variable and the covariate. This way of handling measurement error models is rarely seen in the literature. So it offers another method to deal with measurement error models. Another contribution of the work is the realization of different treatment effects of different patients by the analysis of our semiparametric method. The threshold values of baseline CD4 count are valuable because they offers guidance for answering questions like what kind of patients need to take new treatments. The estimation procedure is achieved by two steps. In the second step, we need to plug in an estimator which is estimated from the first step. Because the parameter is estimated, there must be a variance inflation, as illustrated in Theorem 3. But the inflation is hardly seen in simulation studies. Therefore, future investigations are needed to examine the problem. We also want to see what kind of effects would be brought about by proposed working models.

Finally, we estimate the survival time distribution functions for Huntington's disease gene mutation carriers and non-carriers. It is of great interest before to know the age distribution of people whose family history contains Huntington's disease. The successful estimation will provide both doctors and patients with the knowledge to treat the gene mutation disease more efficiently. Though the estimation is achieved in Ma & Wang (2013), the family correlation, which definitely determines

the development of Huntington's disease, is not considered at all. Thus we design two methods, namely RBLCE and RBQIF, to take into account the family correlation. We also show that RBLCE and RBQIF are equivalent asymptotically. Our methods are flexible and easy to implement. Besides, both the simulation studies and data analysis show good performance of our proposed methods. Although the choice of R is not determined, we can increase the R value until the efficiency improvement becomes not visible in practice. This work can be extended to many kin-cohort applications.

REFERENCES

Abarin, T. & Wang, L. (2012). Instrumental variable approach to covariate measurement error in generalized linear models. *Annals of the Institute of Statistical Mathematics* **64**, 475–493.

Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *Journal of the American Statistical Association* **81**, 1032–1038.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623–635.

Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors-in-variables model. *Journal of Econometrics* **28**, 273–289.

Amemiya, Y. (1990). Two-stage instrumental variable estimators for the nonlinear errors-in-variables model. *Journal of Econometrics* **44**, 311–332.

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

Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estiamtes under random censorship. *The Annals of Statistics* **2**, 437–453.

Buzas, J. S. (1997). Instrumental variable estimation in nonlinear measurement error models. *Communications in Statistics, Part A* **26**, 2861–2877.

Buzas, J. S. & Stefanski, L. A. (1996). Instrumental variable estimation in generalized linear measurement error models. *Journal of American Statistical Association* **91**, 999–1006.

Dorsey, E. & the Huntington Study Group COHORT Investigators (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the cohort study. *PLoS ONE* **7**, 429–522.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.

Genton, M. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton: Chapman & Hall / CRC.

Hall, P. & Ma, Y. (2007). Measurement error models with unknown error structure. *Journal of Royal Statistical Society, Series B* **69**, 429–446.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S. & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine* **335**, 1081–1090.

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

Huang, Y. J. & Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction. *Journal of the American Statistical Association* **95**, 1209–1219.

Huang, Y. J. & Wang, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association* **96**, 1469–1482.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.

Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47**, 725–732.

Huggins, R. M. (2006). Semiparametric estimation of animal abundance using capture-recapture data from open populations. *Biometrics* **62**, 684–690.

Huggins, R. M. & Hwang, W. H. (2009). A measurement error model for heterogeneous capture probabilities in mark-recapture experiments: An estimating equation approach. *Journal of Agricultural, Biological, and Environmental Statistics* **15**, 198–208.

Hwang, W. H. & Huang, S. Y. H. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrika* **59**, 1113–1122.

Hwang, W. H., Huang, S. Y. H. & Wang, C. Y. (2007). Effects of measurement error and conditional score estimation in capture-recapture models. *Statistica Sinica* **17**, 301–316.

Hwang, W. H. & Huggins, R. (2007). Application of semiparametric regression

models the analysis of capture-recapture experiments. *Australian and New Zealand Journal of Statistics* **49**, 191–202.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration – stochastic model. *Biometrika* **52**, 225–247.

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

Lindsay, B. G. & Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science* **18**, 394–410.

Ma, Y. & Tsiatis, A. A. (2006a). Closed form semiparametic estimators for measurement error models. *Statistica Sinica* **16**, 183–193.

Ma, Y. & Tsiatis, A. A. (2006b). Closed form semiparametric estimators for measurement error models. *Statistica Sinica* **16**, 183–193.

Ma, Y. & Wang, Y. (2012). Efficient semiparametric estimation for mixture data. *Electronic Journal of Statistics* **6**, 710–737.

Ma, Y. & Wang, Y. (2013). Estimating disease onset distribution functions in mutation carriers with censored mixture data. *Journal of the Royal Statistical Society, Serie C* , in press.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56**, 434–442.

Pledger, S., Pollock, K. H. & Norris, J. L. (2003). Open capture-recapture models with heterogeneity: I. Cormack-Jolly-Seber model. *Biometrics* **59**, 786–794.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: An overview. *Journal of Applied Statistics* **29**, 85–102.

Pollock, K. H., Hines, J. E. & Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* **40**, 329–340.

Schennach, M. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* **75**, 201–239.

Schwarz, C. J. & Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52**, 860–873.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. London: The Blackburn Press, 2nd ed.

Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42**, 267–292.

Stefanski, A. L. & Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703–716.

Stefanski, L. A. & Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of American Statistical Association* **90**, 541–550.

Stefanski, L. A. & Carroll, R. (1985). Covariate measurement error in logistic regression. *Annals of Statistics* **13**, 1335–1351.

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

Tsiatis, A. A. & Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error modals. *Biometrika* **91**, 835–848.

Wang, C. Y. (2000). Flexible regression calibration for covariate measurement error with longitudinal surrogate variables. *Statistica Sinica* **10**, 905–921.

Wang, L. & Hsiao, C. (2011). Method of moments estimation ad identifiability of semiparametric nonlinear errors-in-variables models. *Journal of Econometrics* **165**, 30–44.

Xi, L. Q., Watson, R., Wang, J. P. & Yip, P. S. F. (2009). Estimation in capture-recapture models when covariates are subject to measurement errors and missing data. *The Canadian Journal Statistics* **37**, 645–658.

APPENDIX A

SECTION 2

A.1 Derivation of Λ

For the model

$$\begin{aligned} & f_{Y_i,\overline{\mathbf{W}}_i,\mathbf{X}_i|\mathcal{C}_{i1}}(y_i,\overline{\mathbf{w}}_i,\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\theta}) \\ = \; & f_{\overline{\mathbf{W}}_i|Y_i,\mathbf{X}_i,\mathcal{C}_{i1}}(\overline{\mathbf{w}}_i \mid y_i,\mathbf{x}_i,\mathcal{C}_{i1}) f_{Y_i|\mathbf{X}_i,\mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i,\mathcal{C}_{i1};\boldsymbol{\theta}) f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1}), \end{aligned} \tag{1}$$

where $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$ is unknown, its nuisance tangent space is

$$[\mathbf{h}(\mathbf{X}_i) : \; E\{\mathbf{h}(\mathbf{X}_i) \mid \mathcal{C}_{i1}\} = \mathbf{0}]. \tag{2}$$

To see this, suppose the true model for $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1})$ is $f_0(\mathbf{x}_i \mid \mathcal{C}_{i1})$, and let

$$f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}) = f_0(\mathbf{x}_i \mid \mathcal{C}_{i1})\{1+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{h}(\mathbf{X}_i\},$$

where $\mathbf{h}(\mathbf{X}_i)$ satisfies (2) and is a bounded random function, and $\boldsymbol{\eta}_{r\times 1}$ is required to be sufficiently small such that $1+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{h}(\mathbf{X}_i) \geq 0$. A simple calculation yields that

$$\begin{aligned} \int f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}) d\mu(\mathbf{x}_i) &= \int f_0(\mathbf{x}_i \mid \mathcal{C}_{i1}) d\mu(\mathbf{x}_i) + \int f_0(\mathbf{x}_i \mid \mathcal{C}_{i1}) \boldsymbol{\eta}^{\mathrm{T}}\mathbf{h}(\mathbf{X}_i) d\mu(\mathbf{x}_i) \\ &= 1+\boldsymbol{\eta}^{\mathrm{T}} E\{\mathbf{h}(\mathbf{X}_i) \mid \mathcal{C}_{i1}\} \\ &= 1. \end{aligned}$$

Therefore, $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta})$ is a valid probability density function. When $\boldsymbol{\eta}=\mathbf{0}$, $f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta})$ equals to $f_0(\mathbf{x}_i \mid \mathcal{C}_{i1})$. So it contains the true model. Thus it is a

parametric submodel. Since $\partial f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta})/\partial\boldsymbol{\eta} = \mathbf{h}(\mathbf{X}_i)$, we have shown that any element in the set defined in (2) is indeed one element in the nuisance tangent space of model (1). On the other hand, for any parametric submodel of (1)

$$\begin{aligned} & f_{Y_i,\overline{\mathbf{W}}_i,\mathbf{X}_i|\mathcal{C}_{i1}}(y_i,\overline{\mathbf{w}}_i,\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\theta},\boldsymbol{\eta}) \\ = \; & f_{\overline{\mathbf{W}}_i|Y_i,\mathbf{X}_i,\mathcal{C}_{i1}}(\overline{\mathbf{w}}_i \mid y_i,\mathbf{x}_i,\mathcal{C}_{i1}) f_{Y_i|\mathbf{X}_i,\mathcal{C}_{i1}}(y_i \mid \mathbf{x}_i,\mathcal{C}_{i1};\boldsymbol{\theta}) f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}), \end{aligned}$$

let

$$\begin{aligned} \mathbf{S}_{\boldsymbol{\eta}}(\mathbf{x}_i;\boldsymbol{\eta}) &= \frac{\partial}{\partial\boldsymbol{\eta}} \log\Big\{ f_{Y_i,\overline{\mathbf{W}}_i,\mathbf{X}_i|\mathcal{C}_{i1}}(y_i,\overline{\mathbf{w}}_i,\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\theta},\boldsymbol{\eta})\Big\} \\ &= \frac{\partial}{\partial\boldsymbol{\eta}} \log f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}). \end{aligned}$$

Because $\int f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}_0) d\mu(\mathbf{x}_i) = 1$ when evaluating at the true value $\boldsymbol{\eta}_0$,

$$\begin{aligned} & \frac{\partial}{\partial\boldsymbol{\eta}} \int f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}_0) d\mu(\mathbf{x}_i) \\ = \; & \int \frac{\partial}{\partial\boldsymbol{\eta}} \log f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}_0) f_{\mathbf{X}_i|\mathcal{C}_{i1}}(\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\eta}_0) d\mu(x_i) \\ = \; & E\{\mathbf{S}_{\boldsymbol{\eta}}(\mathbf{x}_i;\boldsymbol{\eta}_0) \mid \mathcal{C}_{i1}\} = \mathbf{0}. \end{aligned}$$

Thus, any element belongs to the nuisance tangent space of model (1) must also belong to the set given in (2).

Finally, because the conditional joint distribution of $(Y_i, \overline{\mathbf{W}}_i)$ can be written as the conditional expectation

$$f_{Y_i,\overline{\mathbf{W}}_i|\mathcal{C}_{i1}}(y_i,\overline{\mathbf{w}}_i \mid \mathcal{C}_{i1}) = E\{f_{Y_i,\overline{\mathbf{W}}_i,\mathbf{X}_i|\mathcal{C}_{i1}}(y_i,\overline{\mathbf{w}}_i,\mathbf{x}_i \mid \mathcal{C}_{i1};\boldsymbol{\theta}) \mid Y_i,\overline{\mathbf{W}}_i,\mathcal{C}_{i1}\},$$

the semiparametric nuisance tangent space is

$$\Lambda = \left[E\left\{\mathbf{h}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\right\} : \; E\left\{\mathbf{h}(\mathbf{X}_i) \mid \mathcal{C}_{i1}\right\} = \mathbf{0}\right].$$

□

A.2 Derivation of $\Lambda^{\perp}$

Suppose $\mathbf{g}(Y_i, \overline{\mathbf{W}}_i)$ is an element in $\Lambda^{\perp}$ and $E\left\{\mathbf{h}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\right\} \in \Lambda$. Since the two spaces are orthogonal, by using the conditional expectations iteratively, we have

$$\begin{aligned} \mathbf{0} &= E\left[\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) E\left\{\mathbf{h}(\mathbf{X}_i) \mid Y_i, \overline{\mathbf{W}}_i, \mathcal{C}_{i1}\right\}\right] \\ &= E\left[\mathbf{h}(\mathbf{X}_i) E\left\{\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) \mid \mathbf{X}_i, \mathcal{C}_{i1}\right\}\right]. \end{aligned}$$

The above equation is true for any random function $\mathbf{h}(\mathbf{X}_i)$ in the Hilbert space. Thus, $E\left\{\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) \mid \mathbf{X}_i, \mathcal{C}_{i1}\right\} = \mathbf{0}$. Therefore, the orthogonal complement of the nuisance tangent space is

$$\Lambda^{\perp} = \left[\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) : E\left\{\mathbf{g}(Y_i, \overline{\mathbf{W}}_i) \mid \mathbf{X}_i, \mathcal{C}_{i1}\right\} = \mathbf{0}\right].$$

□

A.3 Proof of Theorem 1

A standard Taylor expansion on the estimating equation yields

$$\begin{aligned} \mathbf{0} &= N^{-1/2} \sum_{i=1}^{N} \mathbf{S}^{*}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \widehat{\boldsymbol{\theta}}) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{S}^{*}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta}) + \frac{1}{\sqrt{N}} \left\{ \sum_{i=1}^{N} \frac{\partial \mathbf{S}^{*}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right\} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1). \end{aligned}$$

This implies

$$\begin{aligned}
& \sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
= & \left\{ -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right\}^{-1} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta}) \right\} + o_p(1) \\
= & -\mathbf{A}(\boldsymbol{\theta}) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{S}^*_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta}) + o_p(1),
\end{aligned}$$

and the central limit theorem then immediately yields the asymptotic result in Theorem 1.

When the true distribution model $f_{\mathbf{X}_i|\mathcal{C}_{i1}}$ is used, all the * can be eliminated. Using integration by parts and observing that $\mathbf{S}_{\text{eff}}$ is the orthogonal projection of the score function $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$, we have

$$\begin{aligned}
& \mathbf{A}(\boldsymbol{\theta}) \\
= & E \left\{ \frac{\partial \mathbf{S}_{\text{eff}}(Y_i, \overline{\mathbf{W}}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right\} \\
= & \int \frac{\partial \mathbf{S}_{\text{eff}}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f_{Y_i, \overline{\mathbf{W}}_i}(y_i, \overline{\mathbf{w}}_i) d\mu(y_i, \overline{\mathbf{w}}_i) \\
= & \mathbf{0} - \int \mathbf{S}_{\text{eff}}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta}) \frac{\partial \log f_{Y_i, \overline{\mathbf{W}}_i}(y_i, \overline{\mathbf{w}}_i)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f_{Y_i, \overline{\mathbf{W}}_i}(y_i, \overline{\mathbf{w}}_i) d\mu(y_i, \overline{\mathbf{w}}_i) \\
= & -E \left\{ \mathbf{S}_{\text{eff}}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta}) \mathbf{S}^{\mathrm{T}}_{\theta}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta}) \right\} \\
= & -E \left\{ \mathbf{S}_{\text{eff}}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta}) \mathbf{S}^{\mathrm{T}}_{\text{eff}}(y_i, \overline{\mathbf{w}}_i; \boldsymbol{\theta}) \right\} \\
= & -\mathbf{B}(\boldsymbol{\theta}).
\end{aligned}$$

The general expression of $\mathbf{V}$ indicates that the variance is $\mathbf{B}^{-1}(\boldsymbol{\theta})$. Finally, it is the optimal variance because it is the variance of the efficient influence function. □

APPENDIX B

SECTION 3

B.1 Derivation of Λ

If we consider the parametric submodel,

$$\begin{aligned} & \mathrm{pr}(Y = y, \mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\ = & \int \mathrm{pr}(Y = y \mid \mathbf{S}, \mathbf{Z}, \boldsymbol{\epsilon}; \boldsymbol{\beta}, \boldsymbol{\gamma}) f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}, \boldsymbol{\eta}_2) f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z}; \boldsymbol{\eta}_1) d\mu(\boldsymbol{\epsilon}), \end{aligned}$$

the nuisance score vectors with respect to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are $\partial \log f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z}; \boldsymbol{\eta}_1)/\partial \boldsymbol{\eta}_1$ and $E\{\partial \log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2 \mid Y, \mathbf{S}, \mathbf{Z}\}$ respectively. The former has the property

$$E\{\partial \log f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z}; \boldsymbol{\eta}_1)/\partial \boldsymbol{\eta}_1\} = \mathbf{0},$$

and $\partial \log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2$ satisfies

$$\begin{aligned} E\{\partial \log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2 \mid \mathbf{S}, \mathbf{Z}\} &= \mathbf{0}, \\ E[\{\partial \log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s}, \mathbf{z}, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2\} \boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}] &= \mathbf{0}. \end{aligned}$$

The last equation comes from the condition that $E(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = 0$. This completes the nuisance tangent space derivation for a parametric submodel. Since the nuisance tangent space of our original model is the mean square closure of the nuisance tangent space of all parametric submodels, the conjecture for the desired nuisance tangent

space is the direct sum of two subspaces Λ_1 and Λ_2, where

$$\begin{aligned}
\Lambda_1 &= \{\mathbf{f}(\mathbf{S},\mathbf{Z}) : \mathbf{f} \in \mathbb{R}^p, E(\mathbf{f}) = \mathbf{0}, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty\} \\
\Lambda_2 &= [E\{\mathbf{f}(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid Y,\mathbf{S},\mathbf{Z}\} : \mathbf{f} \in \mathbb{R}^p, E(\mathbf{f} \mid \mathbf{S},\mathbf{Z}) = \mathbf{0}, \\
&\qquad E(\boldsymbol{\epsilon}\mathbf{f}^{\mathrm{T}} \mid \mathbf{S},\mathbf{Z}) = \mathbf{0}, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty].
\end{aligned}$$

In the second part of the proof, we must show that for any bounded random functions $\mathbf{f}_1(\mathbf{S},\mathbf{Z}) \in \Lambda_1$ and $E\{\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid Y,\mathbf{S},\mathbf{Z}\} \in \Lambda_2$, they are the nuisance score vectors of a particular parametric submodel. When the true models for $f_{\mathbf{S},\mathbf{Z}}(\mathbf{s},\mathbf{z})$ and $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})$ are $f_0(\mathbf{s},\mathbf{z})$ and $f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})$ respectively, we define new functions with the aid of $\mathbf{f}_1(\mathbf{S},\mathbf{Z})$ and $\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})$ such that

$$\begin{aligned}
f_{\mathbf{S},\mathbf{Z}}(\mathbf{s},\mathbf{z};\boldsymbol{\eta}_1) &= f_0(\mathbf{s},\mathbf{z})\{1 + \boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{f}_1(\mathbf{S},\mathbf{Z})\} \\
f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z},\boldsymbol{\eta}_2) &= f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})[1 + \boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})].
\end{aligned}$$

$\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ must be sufficiently small such that

$$1 + \boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{f}_1(\mathbf{S},\mathbf{Z}) \geq 0, \text{ and } 1 + \boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \geq 0.$$

Both $f_{\mathbf{S},\mathbf{Z}}(\mathbf{s},\mathbf{z};\boldsymbol{\eta}_1)$ and $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z},\boldsymbol{\eta}_2)$ are valid probability density function because they are positive and their integration from negative infinity to infinity are 1, as can

be seen below.

$$\begin{aligned}
& \int\int f_{\mathbf{S},\mathbf{Z}}(\mathbf{s},\mathbf{z};\boldsymbol{\eta}_1)d\mu(\mathbf{s},\mathbf{z}) \\
= & \int\int f_0(\mathbf{s},\mathbf{z})d\mu(\mathbf{s},\mathbf{z}) + \int\int f_0(\mathbf{s},\mathbf{z})\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{f}_1(\mathbf{S},\mathbf{Z})d\mu(\mathbf{s},\mathbf{z}) \\
= & 1 + \boldsymbol{\eta}_1^{\mathrm{T}} E\{\mathbf{f}_1(\mathbf{S},\mathbf{Z})\} = 1, \\
& \int f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z},\boldsymbol{\eta}_2)d\mu(\boldsymbol{\epsilon}) \\
= & \int f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})d\mu(\boldsymbol{\epsilon}) + \int f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})d\mu(\boldsymbol{\epsilon}) \\
= & 1 + \boldsymbol{\eta}_2^{\mathrm{T}} E\{\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid \mathbf{S},\mathbf{Z}\} = 1.
\end{aligned}$$

Moreover,

$$\begin{aligned}
& \int f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z},\boldsymbol{\eta}_2)\boldsymbol{\epsilon}^{\mathrm{T}}d\mu(\boldsymbol{\epsilon}) \\
= & \int f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})\boldsymbol{\epsilon}^{\mathrm{T}}d\mu(\boldsymbol{\epsilon}) + \int f_0(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z})\boldsymbol{\eta}_2^{\mathrm{T}}\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}}d\mu(\boldsymbol{\epsilon}) \\
= & 0 + \boldsymbol{\eta}_2^{\mathrm{T}} E\{\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S},\mathbf{Z}\} = 0.
\end{aligned}$$

So the density for $\boldsymbol{\epsilon}$ given $\mathbf{S}$ and $\mathbf{Z}$ also satisfies $E(\boldsymbol{\epsilon} \mid \mathbf{S},\mathbf{Z}) = \mathbf{0}$. One the other hand, the score vectors for the parametric submodel are

$$\begin{aligned}
\mathcal{S}_{\eta_1} &= \frac{\partial \log f_{\mathbf{S},\mathbf{Z}}(\mathbf{s},\mathbf{z};\boldsymbol{\eta}_1)}{\partial \boldsymbol{\eta}_1} = \mathbf{f}_1(\mathbf{S},\mathbf{Z}), \\
\mathcal{S}_{\eta_2} &= \int \frac{\partial \log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{s},\mathbf{z},\boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_2} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid Y = y,\mathbf{s},\mathbf{z})d\mu(\boldsymbol{\epsilon}) \\
&= E\{\mathbf{f}_2(\boldsymbol{\epsilon},\mathbf{S},\mathbf{Z}) \mid Y,\mathbf{S},\mathbf{Z}\}.
\end{aligned}$$

This leads to the result. □

B.2 Derivation of $\Lambda^{\perp}$

Using the form of the nuisance tangent space, it can be shown that

$$\Lambda = [E\{\mathbf{f}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} : \mathbf{f} \in \mathbb{R}^p, E(\boldsymbol{\epsilon}\mathbf{f}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}) = \mathbf{0}, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty].$$

Therefore, any element $g(Y, \mathbf{S}, \mathbf{Z}) \in \Lambda^{\perp}$ must satisfy

$$\begin{aligned} 0 &= E[g^{\mathrm{T}}(Y, \mathbf{S}, \mathbf{Z}) E\{\mathbf{f}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}] \\ &= E[E\{g^{\mathrm{T}}(Y, \mathbf{S}, \mathbf{Z})\mathbf{f}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\}] \\ &= E[E\{g^{\mathrm{T}}(Y, \mathbf{S}, \mathbf{Z}) \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\} f(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})] \end{aligned}$$

for any $f(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z})$ such that $E(\boldsymbol{\epsilon}\mathbf{f}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}) = \mathbf{0}$. Therefore, $E\{g(Y, \mathbf{S}, \mathbf{Z}) \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\}$ must have the form $\mathbf{a}(\mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}$ such that $E(\mathbf{a}^{\mathrm{T}}\mathbf{a}) < \infty$. This yields the desired result. □

B.3 Proof of Theorem 2

Note that even under a possibly incorrect working model, we have

$$\begin{aligned} & E\{\ \mathcal{S}^*_{\text{eff}}(Y_i, \mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\theta})\} \\ = \ & E[E\{\ \mathcal{S}^*_{\text{eff}}(Y_i, \mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\theta}) \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}\}] \\ = \ & E(E[\ \mathcal{S}^*_{\theta}(Y, \mathbf{S}, \mathbf{Z}) - E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\} \mid \boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}]) \\ = \ & E\left(E\left[\ \mathcal{S}^*_{\theta}(Y, \mathbf{S}, \mathbf{Z})\boldsymbol{\epsilon}^{\mathrm{T}} - E^*\{\mathbf{b}(\boldsymbol{\epsilon}, \mathbf{S}, \mathbf{Z}) \mid Y, \mathbf{S}, \mathbf{Z}\}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z}\right] \left\{E^*(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \mid \mathbf{S}, \mathbf{Z})\right\}^{-1} \boldsymbol{\epsilon}\right) \\ = \ & \mathbf{0}, \end{aligned}$$

which implies that the corresponding estimator $\boldsymbol{\theta}$ is consistent. In the above display, the second equality is due to the construction of $\mathcal{S}_{\text{eff}}$, the third equality is because $\mathbf{b}$ satisfies the integral equation (3.7), and the last equality is because $E(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = \mathbf{0}$.

Standard Taylor expansion then yields the desired result. Namely,

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{N}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\widehat{\boldsymbol{\theta}}) \\ = \ & n^{-1/2}\sum_{i=1}^{N}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\boldsymbol{\theta}) \\ + \ & n^{-1/2}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\left\{\sum_{i=1}^{N}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\boldsymbol{\theta})\right\}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+o_p(1). \end{aligned}$$

The left side of the equation is zero by observing since $\widehat{\boldsymbol{\theta}}$ is the solution of the estimating equation. It implies,

$$\begin{aligned} & \sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \\ = \ & \left\{-\frac{1}{n}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\boldsymbol{\theta})\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{N}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\boldsymbol{\theta})\right\}+o_p(1) \\ = \ & -\mathbf{A}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{N}\ \mathcal{S}^{*}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i,\boldsymbol{\theta})\right\}+o_p(1). \end{aligned}$$

The asymptotic result in Theorem 2 is deduced by implementing the central limit theorem. Furthermore, if the true model $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}\mid\mathbf{S},\mathbf{Z})$ is used, the variance will achieve the minimum semiparametric bound because

$$\begin{aligned} & E\left\{\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\ \mathcal{S}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i)\right\} \\ = \ & -E\left\{\ \mathcal{S}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i)\ \mathcal{S}^{\mathrm{T}}_{\boldsymbol{\theta}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i)\right\} \\ = \ & -E\left\{\ \mathcal{S}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i)\ \mathcal{S}^{\mathrm{T}}_{\text{eff}}(Y_i,\mathbf{S}_i,\mathbf{Z}_i)\right\}. \end{aligned}$$

The last equation is true since $\mathcal{S}_{\text{eff}}$ is the projection of $\mathcal{S}_{\boldsymbol{\theta}}$ onto the space $\Lambda^{\perp}$. It means that $\mathbf{A}=-\mathbf{B}$, and the variance becomes $\mathbf{B}=\left[E\{\ \mathcal{S}_{\text{eff}}(Y,\mathbf{S},\mathbf{Z})^{\otimes 2}\}\right]^{-1}$. □

B.4 Proof of Theorem 3

Consider the joint estimating equation

$$\begin{aligned}\sum_{i=1}^{N} \mathcal{S}^*_{\text{eff}}(Y_i, \mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\alpha}) &= \mathbf{0} \\ \sum_{i=1}^{N} \mathcal{S}_{\boldsymbol{\alpha}}(Y_i, \mathbf{S}_i, \mathbf{Z}_i, \boldsymbol{\alpha}) &= \mathbf{0}\end{aligned}$$

for estimating $\boldsymbol{\alpha}, \boldsymbol{\theta}$ simultaneous, the Taylor expansion yields

$$\begin{aligned}&\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \end{pmatrix} \\ = & -\left\{ \frac{1}{n}\sum_{i=1}^{N} \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathcal{S}^*_{\text{eff}} & \frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mathcal{S}^*_{\text{eff}} \\ \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathcal{S}_{\boldsymbol{\alpha}} & \frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mathcal{S}_{\boldsymbol{\alpha}} \end{pmatrix} \right\}^{-1} \left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^{N} \begin{pmatrix} \mathcal{S}^*_{\text{eff}} \\ \mathcal{S}_{\boldsymbol{\alpha}} \end{pmatrix} \right\} + o_p(1) \\ = & -\begin{pmatrix} \mathbf{A} & \mathbf{A}_1 \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}^{-1} \left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^{N} \begin{pmatrix} \mathcal{S}^*_{\text{eff}} \\ \mathcal{S}_{\boldsymbol{\alpha}} \end{pmatrix} \right\} + o_p(1).\end{aligned}$$

It indicates the normal limiting distribution with variance

$$\begin{pmatrix} \mathbf{A} & \mathbf{A}_1 \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{B} & \mathbf{B}_1^{\mathrm{T}} \\ \mathbf{B}_1 & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{A}_1 \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}^{-\mathrm{T}},$$

by the central limit theorem. The $(1,1)$th cell of the resulting matrix by expanding the above expression is $\mathbf{V} = \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}} + \mathbf{V}_\alpha$ where

$$\mathbf{V}_\alpha = \mathbf{A}^{-1}\left\{\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_2(\mathbf{A}_1\mathbf{A}_2^{-1})^{\mathrm{T}} - \mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_1 - (\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{B}_1)^{\mathrm{T}}\right\}(\mathbf{A}^{-1})^{\mathrm{T}}.$$

When

$$f^*_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}) = f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \mathbf{S}, \mathbf{Z}),$$

$$-\mathbf{A} = \mathbf{B}.$$

The resulting estimation variance is minimized. □

APPENDIX C

SECTION 4

C.1 Proof of Theorem 4

Write $\mathbf{A}$ as $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_R)$. Under the constraint $\mathbf{AJ} = \mathbf{I}_p$, we have

$$\begin{aligned} E\{\widehat{\mathbf{F}}(t)\} &= E\{\mathbf{A}\widehat{\mathbf{F}}_L(t)\} = \sum_{r=1}^{R} \mathbf{A}_r E\{\widehat{\mathbf{F}}^r(t)\} \\ &= \sum_{r=1}^{R} \mathbf{A}_r \mathbf{F}(t) + o_p(1) = \mathbf{AJF}(t) + o_p(1) = \mathbf{F}(t) + o_p(1). \end{aligned}$$

This shows that $\widehat{\mathbf{F}}(t)$ is a consistent estimator.

The variance of $\widehat{\mathbf{F}}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t)$ is $\mathbf{AUA}^{\mathrm{T}}$ for a general $\mathbf{A}$ matrix. For any $\mathbf{A}$ that satisfies $\mathbf{AJ} = \mathbf{I}_p$, we have

$$\begin{aligned} & \mathbf{AUA}^{\mathrm{T}} - \mathbf{A}_{\mathrm{opt}}\mathbf{U}\mathbf{A}_{\mathrm{opt}}^{\mathrm{T}} \\ = \; & \mathbf{AUA}^{\mathrm{T}} - (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} \\ = \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AUA}^{\mathrm{T}}\mathbf{AJ}(\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1} - (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} \\ = \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}\left\{\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AUA}^{\mathrm{T}}\mathbf{AJ} - \mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ}\right\} \\ & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1} \\ = \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AU}^{\frac{1}{2}}\left\{I - \mathbf{U}^{-\frac{1}{2}}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-\frac{1}{2}}\right\} \\ & \mathbf{U}^{\frac{1}{2}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ}(\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}. \end{aligned}$$

It is easy to verify that $I - \mathbf{U}^{-\frac{1}{2}}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-\frac{1}{2}}$ is an idempotent matrix, hence it is semi-positive definite. Therefore, $\mathbf{AUA}^{\mathrm{T}} - (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}$ is also semi-positive

definite. □

C.2 Proof of Theorem 5

Taking derivative of the quadratic form (4.9) with respect to $\mathbf{F}(t)$ and omit the higher order terms, we obtain

$$E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W} \sum_{i=1}^{N} \mathbf{g}_i(t) = \mathbf{0}. \tag{3}$$

In the following, we first investigate $N^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbf{g}_i(t)$.

For $r = 1, \ldots, R$, write the observations in the rth sample as $\{O_i^r : O_i^r = (\mathbf{q}_i^r, Y_i^r, \delta_i^r), i = 1, \ldots, N\}$. Because there are m possible values for $\mathbf{q}_i^r$'s, we can divide these N observations into m groups $\mathbf{O}_1^r, \ldots \mathbf{O}_m^r$, where

$$\mathbf{O}_l^r = \left\{O_{l,k}^r : \ O_{l,k}^r = (\mathbf{u}_l, Y_{l,k}^r, \delta_{l,k}^r), \ k = 1, \ldots, d_l^r\right\},$$

and the Kaplan-Meier estimator in the respective group is denoted $\widehat{H}_l^r(t)$ for $l = 1, \ldots, m$.

From Breslow & Crowley 1974, we have the asymptotic expansion

$$\sqrt{d_l^r}\{\widehat{H}_l^r(t) - H_l(t)\} = (d_l^r)^{-1/2} \sum_{k=1}^{d_l^r} a(O_{l,k}^r) + o_p(1),$$

where $a(O_{l,k}^r)$ is a function of the kth observation $O_{l,k}^r$ and $E\{a(O_{l,k}^r)\} = 0$. Inserting this relation into (4.8), we have

$$
\begin{aligned}
N^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbf{g}_i(t) &= N^{-\frac{1}{2}} \sum_{l=1}^{m} \begin{bmatrix} \sqrt{d_l^1}\mathbf{u}_l \sqrt{d_l^1}\{\widehat{H}_l^1(t) - H_l(t)\} \\ \vdots \\ \sqrt{d_l^R}\mathbf{u}_l \sqrt{d_l^R}\{\widehat{H}_l^R(t) - H_l(t)\} \end{bmatrix} \\
&= N^{-\frac{1}{2}} \sum_{l=1}^{m} \begin{bmatrix} \mathbf{u}_l \sum_{k=1}^{d_l^1} a(O_{l,k}^1) \\ \vdots \\ \mathbf{u}_l \sum_{k=1}^{d_l^R} a(O_{l,k}^R) \end{bmatrix} + o_p(1) \\
&= N^{-\frac{1}{2}} \sum_{i=1}^{N} \begin{bmatrix} \mathbf{q}_i^1 a(O_i^1) \\ \vdots \\ \mathbf{q}_i^R a(O_i^R) \end{bmatrix} + o_p(1), \qquad (4)
\end{aligned}
$$

where the first equality is obtained through rewriting the summation in (4.8), and the last equality is obtained similarly. Viewing $\mathbf{q}_i^r$'s as random quantities, we have that $N^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbf{g}_i(t)$ is the average of independently identically distributed mean zero random quantities hence it converges to a mean zero normal distribution with variance denoted $\mathcal{M}$.

Standard Taylor expansion of (3) then yields

$$
\sqrt{N}\{\widehat{\mathbf{F}}(t) - \mathbf{F}(t)\} \to N\{\mathbf{0}, \mathbf{B}^{-1}\mathbf{C}(\mathbf{B}^{-1})^{\mathrm{T}}\}
$$

in distribution when $N \to \infty$, where

$$
\begin{aligned}
\mathbf{B} &= E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}, \\
\mathbf{C} &= \left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W}\, \mathcal{M} \mathbf{W} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right].
\end{aligned}
$$

Similar derivation as in the proof of Theorem 4 can be used to show that the optimal choice of the weight matrix is $\mathbf{W} = \mathcal{M}^{-1}$, and the resulting variance is

$$\left[E\left\{ \frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)} \right\}^{\mathrm{T}} \mathcal{M}^{-1} E\left\{ \frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)} \right\} \right]^{-1}.$$

□