STATISTICAL METHODS FOR TRUNCATED SURVIVAL DATA

Lior Rennert

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

_____

Sharon X. Xie, Professor of Biostatistics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Warren B. Bilker, Professor of Biostatistics

Kevin G. Lynch, Associate Professor of Psychiatry

Murray Grossman, Professor of Neurology

STATISTICAL METHODS FOR TRUNCATED SURVIVAL DATA

ACKNOWLEDGEMENT

I would like to thank my dissertation advisor, Dr. Sharon X. Xie, for all of her support and encouragement during my time as a PhD student. I would also like to thank my committee member and former supervisor Dr. Kevin G. Lynch for his mentorship over the years. The professors, students, and staff in the Biostatistics Department here at Penn have provided an amazing environment to be in, and I am very grateful for my experience here. Finally, I would like to thank my family. Among many things, I would not be here if it were not for the love and motivation instilled in me from my mother. My father and brothers have also provided me with great love and encouragement and have believed in me throughout the years. Finally, I would like to thank all of my friends. While I probably would have graduated much sooner if it weren't for you, I would not trade the experiences we shared over the years for anything.

ABSTRACT

STATISTICAL METHODS FOR TRUNCATED SURVIVAL DATA

Lior Rennert

Sharon X. Xie

Truncation is a well-known phenomenon that may be present in observational studies of time-to-event data. For example, autopsy-confirmed survival studies of neurodegenerative diseases are subject to selection bias due to the simultaneous presence of left and right truncation, also known as double truncation. While many methods exist to adjust for either left or right truncation, there are very few methods that adjust for double truncation. When time-to-event data is doubly truncated, the regression coefficient estimators from the standard Cox regression model will be biased. In this dissertation, we develop two novel methods to adjust for double truncation when fitting the Cox regression model. The first method uses a weighted estimating equation approach. This method assumes the survival and truncation times are independent. The second method relaxes this independence assumption to an assumption of conditional independence between the survival and truncation times. As opposed to methods that ignore truncation, we show that both proposed methods result in consistent and asymptotically normal regression coefficient estimators and have little bias in small samples. We use these proposed methods to assess the effect of cognitive reserve on survival in individuals with autopsy-confirmed Alzheimers disease. We also conduct an extensive simulation study to compare survival distribution function estimators in the presence of double truncation and conduct a case study to compare the survival times of individuals with autopsy-confirmed Alzheimers disease and frontotemporal lobar degeneration. Furthermore, we introduce an R-package for the above methods to adjust for double truncation when fitting the Cox model and estimating the survival distribution function.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Truncation is a statistical phenomenon that has been shown to occur in a wide range of applications, including survival analysis, epidemiology, economics, and astronomy. Individuals who are subject to truncation provide no information to the investigator. *Left truncation* occurs when data is only recorded for individuals whose survival time exceeds a random time (i.e. left truncation time). *Right truncation* occurs when data is only recorded for individuals whose survival time proceeds a random time (i.e. right truncation time). When both left and right truncation are present, this is known as *double truncation*.

Double truncation is inherent in retrospective autopsy-confirmed studies of neurodegenerative diseases. Due to the inaccuracy of clinical diagnosis (Beach et al., 2012), autopsy confirmation is needed for a definitive diagnosis (Grossman and Irwin, 2016) of a particular neurodegenerative disease. The right truncation occurs because information is only obtained from a subject when they receive an autopsy. Subjects who survive past the end of the study are not diagnosed and therefore not included in the study sample, resulting in a sample that is biased towards subjects with smaller survival times. Furthermore, the retrospective sample is also left truncated because subjects who succumb to the disease before they enter the study are unobserved, resulting in a sample that is biased towards subjects with larger survival times. We note that right censoring is not possible in this setting, since any subject who has an autopsy performed will also have a known survival time. A diagram showing how double truncation occurs is provided in Figure 2.1.

The aim of our data analysis is to get accurate estimates of the effect of risk factors on survival from disease symptom onset in subjects with autopsy-confirmed neurodegenerative diseases. The default application for analysis in this setting is the Cox regression model (Cox, 1972). However, regression techniques which do account for truncation will result in biased regression coefficient estimators. This is because under left truncation, individuals with smaller event times are less likely to be observed, resulting in a study sample that is biased towards larger event times and risk factors associated with larger event times. Similarly, under right truncation, individuals with larger event times are less likely to be observed, resulting in a study sample that is biased towards

smaller event times and risk factors associated with smaller event times. If double truncation is not accounted for, then the regression coefficient estimators from the Cox regression model will be biased.

In Chapter 2, we introduce a weighted estimating equation approach to adjust the Cox regression model in the presence of double truncation, under the assumption that the survival times and truncation times are independent. In chapter 3, we use a conditional likelihood approach to relax this independence assumption to an assumption of conditional independence between the survival and truncation times. Here we estimate the regression coefficient estimators for the Cox regression model using an expectation-maximization (E-M) algorithm. In Chapter 4, we conduct a case study to compare estimators of the survival distribution function under double truncation. In Chapter 5, we introduce an R-package to adjust both the Cox regression model and the survival time distribution function in the presence of double truncation. The R-package is intended for the situation where the survival and truncation times are independent. Concluding remarks are given in Chapter 6. The code for the Cox regression coefficient estimators using the EM algorithm introduced in Chapter 3 is provided in Appendix D. The code for the functions contained in the R package described in Chapter 5, which use a nonparametric weighting approach to adjust the Cox regression model (Chapter 2) and the survival distribution function (Chapter 4), is provided in Appendix E and F, respectively.

CHAPTER 2

COX REGRESSION MODEL WITH DOUBLY TRUNCATED DATA

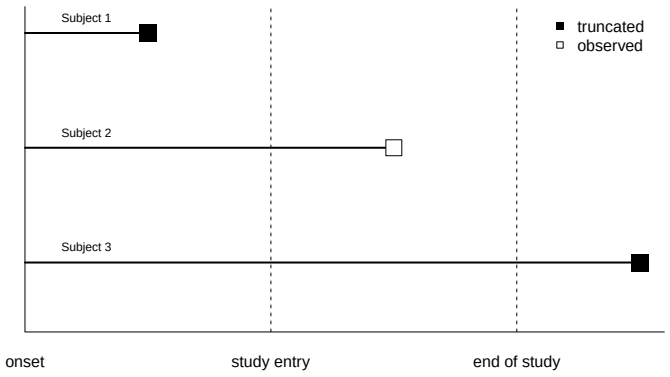
2.1. Introduction

Accurate regression coefficient estimation in survival analysis is crucial for studying factors that affect disease progression. However in some survival studies the outcome of interest may be subject to either left or right truncation. When both left and right truncation are present, this is known as double truncation. For example, double truncation is inherent in retrospective autopsy-confirmed studies of Alzheimer's disease (AD), where autopsy confirmation is the gold standard for diagnosing AD due to the inaccuracy of clinical diagnosis (Beach et al., 2012). The right truncation occurs because information is only obtained from a subject when they receive an autopsy. Subjects who survive past the end of the study are not diagnosed and therefore not included in the study sample, resulting in a sample that is biased towards subjects with smaller survival times. Furthermore, the retrospective sample is also left truncated because subjects who succumb to the disease before they enter the study are unobserved, resulting in a sample that is biased towards subjects with larger survival times. We note that right censoring is not possible in this setting, since any subject who has an autopsy performed will also have a known survival time.

A diagram showing how double truncation occurs is provided in Figure 2.1. In this hypothetical example, we assume subjects 1, 2, and 3 all have similar times of disease symptom onset. For illustrative purposes, we also assume that subjects 1, 2, and 3 have the same study entry time, however this need not be the case. Here the x-axis represents time, and the squares represent the terminating events. Subject 1 is left truncated because they die before they enter the study. Subject 2 enters the study and dies before the end of the study, and is therefore observed. Subject 3 is right truncated because they live past the end of the study, and therefore do not have an autopsy performed.

If the left and right truncation are not accounted for then the observed sample will be biased, which may lead to biased estimators of regression coefficients and hazard ratios. In this paper, we examine the relationship between education and survival from AD symptom onset in a retrospective autopsy-confirmed AD population. The default application for analysis in this setting is the Cox

Figure 2.1: Hypothetical example of double truncation



In this hypothetical example, we assume subjects 1, 2, and 3 all have similar times of disease symptom onset. For illustrative purposes, we also assume that subjects 1, 2, and 3 have the same study entry time, however this need not be the case. Here the x-axis represents time, and the squares represent the terminating events. Subject 1 is left truncated because they die before they enter the study. Subject 2 enters the study and dies before the end of the study, and is therefore observed. Subject 3 is right truncated because they live past the end of the study, and therefore do not have an autopsy performed.

regression model (Cox, 1972). However, to obtain consistent regression coefficient estimators, we must adjust for truncation. Regression techniques already exist under left truncation (Lai and Ying, 1991), right truncation (Kalbfleisch and Lawless, 1991), and length-biased data (Wang, 1996). In this paper, we propose a Cox regression model to adjust for double truncation using a weighted estimating equation approach, where the hazard rate for the failure times follows that of the standard Cox regression model.

Although double truncation may appear in many studies in which data is only recorded for subjects whose event times fall in an observable time interval, the amount of literature on methods to handle double truncation is small. Most of the literature deals with the estimation of the survival distribution rather than regression. Efron and Petrosian (1999) introduced the nonparametric maximum likelihood estimator (NPMLE) for the survival distribution function under double truncation. Shen (2010) investigated the asymptotic properties of the NPMLE and introduced a nonparametric estimator of the truncation distribution function. Shen (2010) and Moreira and de Ũna-Álvarez (2010) introduced a semiparametric maximum likelihood estimator (SPMLE) for the survival distribution function under double truncation. Shen (2013) introduced a method for regression analysis of interval censored

and doubly truncated data using linear transformation models, but these models only allow discrete covariates and the asymptotic properties of the resulting estimators are not established. Moreira, de Ũna-Álvarez, and Meira-Machado (2016) introduced nonparametric kernel regression for doubly truncated data, where a mean function conditional on a single covariate is estimated, rather than a hazard ratio. Furthermore, the resulting estimator is asymptotically biased. Since right censoring is rare under double truncation, the current literature assumes no censoring or interval censoring.

The concept of adjusting the Cox regression model for biased samples using a weighted estimating equation approach was first introduced by Binder (1992) for survey data. In this setting, the weights were known *a priori* and a biased study sample was selected directly from the target population (i.e. the population we wish to study). Lin (2000) proved the asymptotic normality of the regression coefficient estimator introduced by Binder, and extended the model to settings where the biased study sample is selected from a representative sample of the underlying target population. Pan and Schaubel (2008) introduced a Cox regression model with estimated weights, using logistic regression to estimate each subject's probability of selection into the study. In their setting, they assumed that baseline information was available from both subjects with observed and missing survival times. Due to truncation, we do not have any information on subjects with missing survival times. Therefore previous methods are unable to address the unique challenges present in our AD study.

There are several new contributions of this paper to the literature. We propose a Cox regression model using a weighted estimating equation approach to obtain a hazard ratio estimator under double truncation, where the weights are inversely proportional to the probability that a subject is *not* truncated. These selection probabilities are estimated both parametrically and nonparametrically using methods introduced by Shen (2010a, 2010b) and Moreira and de Ũna-Álvarez (2010). As opposed to using data from missing subjects, the selection probabilities here are estimated using survival and truncation times from observed subjects only. The parametric selection probabilities make distributional assumptions about the truncation times, while the nonparametric selection probabilities do not. We show that the proposed regression coefficient estimators are consistent, and greatly reduce the bias in finite samples compared to the standard Cox regression estimator which ignores double truncation. We prove the asymptotic normality of the regression coefficient estimator under parametric weights, and provide a consistent estimator of its asymptotic variance. We

use the bootstrap technique (Efron and Tibshirani, 1993) to estimate the variance and confidence intervals of the regression coefficient estimator under nonparametric weights.

The remainder of this paper is organized as follows. In Section 2.2 we introduce the weighted estimating equation and the proposed estimators, as well as the estimation procedure for the weights. The asymptotic properties of the proposed estimators are provided in Section 2.3. In Section 2.4 we conduct a simulation study to assess the finite sample performance of the proposed estimators. The proposed method is then applied to the AD data in Section 2.5. Discussion and concluding remarks are given in Section 2.6.

2.2. Proposed Parametric and Nonparametric Weighted Estimators

Throughout this paper, we refer to *population random variables* as random variables from the target population and denote them without subscripts. We refer to *sampling random variables* as random variables from the observed sample and denote them with subscripts. These two sets of variables may have different distributions due to double truncation, which is why standard methodology may be inappropriate.

Let T_i denote the observed survival times for subject $i = 1, ..., n \leq N$, where n is the size of the observed sample and N is the size of the *target sample*. Here we use the term target sample to denote a representative sample from the underlying target population. In our setting, this consists of all subjects that would have been included in the observed sample had truncation not occurred. For a given time t, define $Y_i(t) = 1_{\{T_i \geq t\}}$ and $N_i(t) = 1_{\{T_i \leq t\}}$. Let τ be a constant set to the end of study time. The Cox regression model assumes that for a given subject with $p \times 1$ covariate vector $\boldsymbol{Z}_i(t)$, the hazard function at time t is given by $\lambda_i(t) = \lambda_0(t)e^{\boldsymbol{\beta}_0' \boldsymbol{Z}_i(t)}$, where $\lambda_0(t)$ is the true baseline hazard function and is unspecified. The true $p \times 1$ regression coefficient vector, $\boldsymbol{\beta}_0$, is estimated by $\widehat{\boldsymbol{\beta}}$, the solution to

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{Z}_i(t) - \frac{\sum_{j=1}^{n} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)} \boldsymbol{Z}_j(t)}{\sum_{j=1}^{n} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)}} \right\} dN_i(t) = \boldsymbol{0}, \tag{2.1}$$

where $dN_i(t) = N_i(t) - N_i(t^-)$. Since right censoring is not possible under our sampling scheme, we do not include it in the estimation procedures. Therefore $dN_i(T_i)$ = 1 in this setting, since all subjects in our study sample experience an event.

When subjects have unequal probabilities of selection, then the study sample will not be a representative sample of the underlying target population. To adjust for biased samples, Binder (1992) proposed weighting each subject in the score equation 2.1 by the inverse probability of their inclusion in the sample. The true regression coefficient β_0 is then estimated by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}$, the solution to the weighted score equation

$$\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \int_0^{\tau} w_i \Big\{ \boldsymbol{Z}_i(t) - \frac{\sum_{j=1}^{n} w_j Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)} \boldsymbol{Z}_j(t)}{\sum_{j=1}^{n} w_j Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)}} \Big\} dN_i(t) = \mathbf{0}. \tag{2.2}$$

Here $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)$ and $w_i = \pi_i^{-1}$, where π_i is the selection probability for subject i, and is conditional on subject specific characteristics. The method described above assumes that the selection probabilities π_i are known *a priori*. When these probabilities are not known, they must be estimated.

In our setting, we can estimate the probability that a subject was selected in our sample (i.e. not truncated), conditional on their observed survival time. Thus a natural solution to adjust for double truncation is to use these estimated selection probabilities in (2). These selection probabilities are estimated using the survival and truncation times from observed subjects only. The estimation procedure is given in Section 2.2.1.

In our data example, the left truncation time is taken to be the time from AD symptom onset to entry into the study. The right truncation time is set to the time from AD symptom onset to the end of the study. Let U and V denote the left and right truncation times, respectively. Due to double truncation, we observe $\{T, U, V, \boldsymbol{Z}(t)\}$ if and only if $U \leq T \leq V$.

Conditional on T_i, subject i is observed with probability $\pi_i = P(U \leq T \leq V | T = T_i)$. Here π_i is the probability that a subject from the target sample with survival time $T = T_i$ is observed, and is called the selection bias function (Bilker and Wang, 1996). For an intuition as to why this weighting scheme works, we consider the following. If x individuals with survival time T_i are observed in the sample, then by the definition of π_i, there must be x/π_i individuals in the target sample with survival time T_i. Without loss of generality, suppose x = 1, so that there are $1/\pi_i$ individuals with survival time T_i in the target sample. Of these, $(1/\pi_i) \times \pi_i = 1$ will be observed and the other $1/\pi_i - 1$ individuals are referred to as ghosts (Turnbull, 1976) and are unobserved. In this case, each T_i represents $1/\pi_i$ individuals from the target sample with survival time $T = T_i$. We can therefore

adjust for the biased sample by weighting each observation in the estimating equation 2.1 by $1/\pi_i$.

To give another intuitive view as to how the weighting works, it can be shown that π_i is proportional to the probability of observing a survival time T_i in the observed sample relative to the probability of observing a survival time T_i in the target sample. That is, $\pi_i \propto P(T = T_i | U \leq T \leq V)/P(T = T_i)$. Using these selection probabilities in (2) works because observations with survival times which are oversampled in the observed sample relative to the target sample are downweighted and those which are undersampled are upweighted, yielding a score function consisting of survival times (and corresponding covariates) that are distributed according to those of the target population. We show in Section 2.3 that if these selection probabilities are estimated consistently and plugged into the score equation 2.2, then this score function is asymptotically equivalent to the unweighted score function using all observations from the target sample, and is therefore asymptotically unbiased. This results in the consistency of the proposed regression coefficient estimators presented below.

2.2.1. Estimation of selection probabilities

The methods used to estimate the selection probabilities assume that the survival and truncation times are independent in the observable region $U \leq T \leq V$. This independence assumption is needed to estimate $\boldsymbol{\pi}$ using the estimation procedures below. We note that under independence, π_i is simply $P(U \leq T_i \leq V)$. Situations where the independence assumption can be relaxed by covariate adjustment are discussed in Section 2.6.

Before we describe the parametric and nonparametric procedures for estimating the selection probabilities, we introduce additional notation and assumptions. Let $f(t)$ and $F(t)$ denote the density and cumulative distribution functions of T. Let $k(u, v)$ and $K(u, v)$ denote the joint density and cumulative distribution functions of (U, V). For any cumulative distribution function H, define the left endpoint of its support by $a_H = inf\{x : H(x) > 0\}$ and the right endpoint of its support by $b_H = inf\{x : H(x) = 1\}$. Let $H_U(u) = K(u, \infty)$ and $H_V(v) = K(\infty, v)$ denote the marginal cumulative distribution functions of U and V, respectively. For the following methods, we assume that $a_{H_U} < a_F \leq a_{H_V}$ and $b_{H_U} \leq b_F < b_{H_V}$. These conditions are needed for identifiability of the selection probability estimators presented below (Shen, 2010a,b; Woodroofe, 1985).

Letting $\pi(t) = P(U \leq t \leq V)$, our methods rest on the assumption that $\pi(t) > 0$ for all $t \in [a_F, b_F]$. That is, we assume all survival times have a positive probability of being observed. A near violation

of this positivity assumption can lead to a π_i that is very small and thus gives undue influence to the i^{th} observation in the score equation 2.2. We discuss a remedy to this situation at the end of Section 2.6. We note that this positivity assumption is generally implied through the identifiability constraints $a_{H_U} < a_F \leq a_{H_V}$ and $b_{H_U} \leq b_F < b_{H_V}$. Justification of these constraints and positivity assumption for our data example, and a discussion on when these may be violated, are given in Web Appendix D.

Nonparametric estimation

We now present the nonparametric estimation of the selection probabilities π_i. As shown in Shen (2010a, p. 837), the distribution of the observed survival times, $\tilde{F}(t)$, can be written as $\tilde{F}(t) = P(T_i \leq t) = P(T \leq t | U \leq T \leq V) = p^{-1} P(T \leq t, U \leq T \leq V) = p^{-1} \int_0^t [K(s, b_{H_V}) - K(s,s)] F(ds)$, where $p = P(U \leq T \leq V)$ is the probability of observing a random subject from the target sample. The last equality follows from the independence of T and (U, V) in the observable region $U \leq T \leq V$. In this case, the density of the observed survival times is given by $\tilde{f}(t) = p^{-1} \times \pi(t) f(t)$, where $\pi(t) = K(t, b_{H_V}) - K(t,t) = P(U \leq t \leq V)$. It can also be shown that under this independence assumption, the joint density of the observed truncation times can be written as $\tilde{k}(u,v) = p^{-1} \times \varphi(u,v) k(u,v)$, where $\varphi(u,v) = F(v) - F(u-) = P(u \leq T \leq v)$.

Let $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_n)$, where $\varphi_i = \varphi(U_i, V_i)$. Since $k(u,v) = p \times \tilde{k}(u,v)/\varphi(u,v)$, we have that when $\boldsymbol{\varphi}$ and p are known, $K(u,v)$ can be estimated by $n^{-1} p \sum_{j=1}^n \frac{1_{\{U_j \leq u, V_j \leq v\}}}{\varphi_j}$. Setting u and v to ∞, we can estimate p by $n \big[\sum_{j=1}^n 1/\varphi_j\big]^{-1}$. Therefore when $\boldsymbol{\varphi}$ is known, we can estimate $K(u,v)$ by $\big[\sum_{j=1}^n 1/\varphi_j\big]^{-1} \sum_{j=1}^n \frac{1_{\{U_j \leq u, V_j \leq v\}}}{\varphi_j}$ and thus $\pi_i = K(T_i, b_{H_V}) - K(T_i, T_i)$ can be estimated by $\big[\sum_{j=1}^n 1/\varphi_j\big]^{-1} \sum_{j=1}^n \frac{1_{\{U_j \leq T_i \leq V_j\}}}{\varphi_j}$. Similarly, since $f(t) = p \times \tilde{f}(t)/\pi(t)$, we have that when $\boldsymbol{\pi}$ is known, $F(t)$ can be estimated by $\big[\sum_{j=1}^n 1/\pi_j\big]^{-1} \sum_{j=1}^n \frac{1_{\{T_j \leq t\}}}{\pi_j}$ and thus $\varphi_i = F(V_i) - F(U_i-)$ can be estimated by $\big[\sum_{j=1}^n 1/\pi_j\big]^{-1} \sum_{j=1}^n \frac{1_{\{U_i \leq T_j \leq V_i\}}}{\pi_j}$.

Shen (2010) proved that the NPMLE's of φ_i and π_i, denoted by $\widehat{\varphi}_i$ and $\widehat{\pi}_i$, respectively, can be found using the following iterative algorithm:

Step 0) Set $\widehat{\varphi}_i^{(0)} = n^{-1} \sum_{j=1}^n 1_{\{U_i \leq T_j \leq V_i\}}$, for $i = 1, ..., n$.

Step 1) Set $\widehat{\pi}_i^{(1)} = \Big(\sum_{j=1}^n \frac{1}{\widehat{\varphi}_j^{(0)}} \Big)^{-1} \sum_{j=1}^n \frac{1_{\{U_j \leq T_i \leq V_j\}}}{\widehat{\varphi}_j^{(0)}}$, for $i = 1, ..., n$.

Step 2) Set $\widehat{\varphi}_i^{(1)} = \left(\sum_{j=1}^n \frac{1}{\widehat{\pi}_j^{(1)}} \right)^{-1} \sum_{j=1}^n \frac{1_{\{U_i \leq T_j \leq V_i\}}}{\widehat{\pi}_j^{(1)}}$, for $i = 1, ..., n$.

Step 3) For a prespecified error e, repeat steps 1 and 2 until $\sum_{i=1}^n |\widehat{\pi}_i^{(s)} - \widehat{\pi}_i^{(s-1)}| < e$.

The NPMLE of $\boldsymbol{\pi}$ is given by $\widehat{\boldsymbol{\pi}}^{np} = (\widehat{\pi}_1^{(s)}, ..., \widehat{\pi}_n^{(s)})$, with estimated weights $\boldsymbol{w}_{np} = 1/\widehat{\boldsymbol{\pi}}^{np}$. The corresponding estimator of $\boldsymbol{\beta}_0$ is denoted by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$, the solution to $\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}^{np}) = \mathbf{0}$.

Because we do not need estimates of the survival and truncation time distributions, the algorithm to estimate $\boldsymbol{\pi}$ presented here is a simplified version of the algorithm given in Shen (2010). We note that both algorithms result in the same estimator $\widehat{\boldsymbol{\pi}}^{np}$.

Parametric estimation

We can also estimate the selection probabilities parametrically using the methods introduced by Shen (2010) and Moreira and de Ũna-Álvarez (2010). In this setting, we assume that the truncation times U and V have a parametric joint density function $k_{\boldsymbol{\theta}}(u, v)$. Here $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a $q \times 1$ vector of parameters and $\boldsymbol{\Theta}$ is the parametric space.

Under the assumption of independence in the region $U \leq T \leq V$, the conditional likelihood of the (U_i, V_i) given T_i is given by $L_c(\boldsymbol{\theta}) = \prod_{i=1}^n k_{\boldsymbol{\theta}}(U_i, V_i)/\pi_i^{\boldsymbol{\theta}}$, where $\pi_i^{\boldsymbol{\theta}} = \int_{u \leq T_i \leq v} k_{\boldsymbol{\theta}}(u, v) du dv = P_{\boldsymbol{\theta}}(U \leq T_i \leq V)$. Here the subscript $\boldsymbol{\theta}$ denotes that the probability depends on $\boldsymbol{\theta}$. In this setting, we estimate π_i by $\pi_i^{\widehat{\boldsymbol{\theta}}} = \int_{u \leq T_i \leq v} k_{\widehat{\boldsymbol{\theta}}}(u, v) du dv$. The conditional likelihood estimator, $\widehat{\boldsymbol{\theta}}$, is the solution to $\boldsymbol{U}_c(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} log L_c(\boldsymbol{\theta}) = \mathbf{0}$.

The MLE of $\boldsymbol{\pi}$ is given by $\boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}} = (\pi_1^{\widehat{\boldsymbol{\theta}}}, ..., \pi_n^{\widehat{\boldsymbol{\theta}}})$. The weights w_i are then estimated by $w_i(\widehat{\boldsymbol{\theta}}) = p(\widehat{\boldsymbol{\theta}})/\pi_i^{\widehat{\boldsymbol{\theta}}}$, where $p(\widehat{\boldsymbol{\theta}}) = P_{\widehat{\boldsymbol{\theta}}}(U \leq T \leq V) = \left(n^{-1} \sum_{j=1}^n 1/\pi_j^{\widehat{\boldsymbol{\theta}}}\right)^{-1}$. The corresponding estimator of $\boldsymbol{\beta}_0$ is denoted by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$, the solution to $\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}}) = \mathbf{0}$. Here the estimated parametric weights $w_i(\widehat{\boldsymbol{\theta}})$ scale $1/\pi_i^{\widehat{\boldsymbol{\theta}}}$ by $p(\widehat{\boldsymbol{\theta}})$ so that they sum up to the original sample size n, which is needed for the derivation of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$.

2.2.2. Estimating the regression coefficients

The estimated parametric and nonparametric selection probabilities, $\boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\boldsymbol{\pi}}^{np}$, can be computed using the code provided in the online supplementary materials. The regression coefficient estimators $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ can be obtained by specifying the weight option in SAS (phreg, surveyphreg) or

R (coxph) with weights $p(\widehat{\boldsymbol{\theta}})/\boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}}$ and $1/\widehat{\boldsymbol{\pi}}^{np}$. More details, including standard error estimates and confidence intervals of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$, as well as sample data, are provided in our code.

2.3. Asymptotic Properties of Proposed Estimators

In this section, we describe the asymptotic properties of our proposed estimators $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$. The asymptotic properties of the proposed estimators refer to the situation when the total number of observed (non-truncated) subjects $n \to \infty$. The following theorems assume that the regularity conditions listed below hold.

The regularity conditions listed here are adapted from Andersen and Gill (1982), Pan and Schaubel (2008), and Shen (2010ab). For a $p \times 1$ vector $\boldsymbol{a}$, we denote $\boldsymbol{a}^{\otimes 0}$ = 1, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2}$ as the $p \times p$ matrix $\boldsymbol{a}\boldsymbol{a}'$. Conditions (a)-(f) below are needed for the consistency of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$:
(a) $\{T_i, U_i, V_i, \boldsymbol{Z}_i\}$ are independent and identically distributed for $i = 1, ..., N$,
(b) $\int_0^\tau d\Lambda_0(t) < \infty$, where $\Lambda_0(t)$ is the baseline cumulative hazard function,
(c) For $\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) = n^{-1} \sum_{i=1}^n \pi_i^{-1} Y_i(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_i(t)} \boldsymbol{Z}_i(t)^{\otimes j}$, $j = 0, 1, 2$, we assume the existence of a neighborhood $\boldsymbol{B}_0$ of $\boldsymbol{\beta}_0$ and $\boldsymbol{\Pi}_0$ of $\boldsymbol{\pi}_0$ such that $sup_{t\in[0,\tau], \boldsymbol{\beta}\in\boldsymbol{B}_0, \boldsymbol{\pi}\in\boldsymbol{\Pi}_0} \| \boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) - \boldsymbol{s_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) \| \xrightarrow{p} \boldsymbol{0}$ as $n \to \infty$, for $j = 0, 1, 2$, where $\boldsymbol{s_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) = E\{\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t)\}$ and $s_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) > 0$,
(d) There exists a $\delta > 0$ such that $\pi(t) = P(U \le t \le V) > \delta$ almost surely for every $t \in [a_F, b_F]$,
(e) $\int_0^\tau |Z_{ik}(t)| dt < \infty$ almost surely, where $Z_{ik}(t)$ is the k^{th} covariate value for subject i at time t,
(f) The Cox model assumption $\lambda(t) = \lambda_0(t) e^{\boldsymbol{\beta}_0' \boldsymbol{Z}(t)}$ holds for both observed and unobserved subjects.

Condition (a) is used when applying the central limit theorem, and this assumption is reasonable in practice assuming the subjects are independent. Condition (b) is used to ensure that several terms in the proofs of Theorems 2.1 and 2.2 are bounded. Condition (c) ensures that $\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t)$ converges in probability, and that $\boldsymbol{e}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) = \frac{\boldsymbol{s_w}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t)}{s_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\pi}; t)}$ is bounded. This assumption is applied several times throughout the proofs below. Condition (d) states that the probability of observing any survival time t in $[0, \tau]$ is non-zero, which leads to the boundedness of several quantities in the proofs below and ensures that N and n go to ∞ at the same rate. Condition (e) is a boundedness condition of the covariate $Z_{ik}(t)$. While it is not required, it is applicable in most situations and is used to simplify the proofs of Theorems 2.1 and 2.2. For a fixed covariate vector $\boldsymbol{Z}(t)$ and fixed time t, condition (f) ensures that the relationship between survival and $\boldsymbol{Z}(t)$ is the same (i.e. assumes a

Cox model) regardless of whether a subject was observed or truncated. This assumption is used implicitly in the proof of Theorem 2.1 when concluding $N^{-1}\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}})$ and $N^{-1}\boldsymbol{U}^{*}(\boldsymbol{\beta})$ (defined in the proof of Theorem 2.1) converge to the same limit.

For the consistency of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$, we need the following three conditions in addition to (a)-(f):
(g) $G_{\boldsymbol{\theta}}(t) = P_{\boldsymbol{\theta}}(U \leq t \leq V) = \int_{u \leq t \leq v} k_{\boldsymbol{\theta}}(u, v) du dv$ is continuous in t for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,
(h) $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$ implies $G_{\widehat{\boldsymbol{\theta}}_n}(t) \xrightarrow{p} G_{\boldsymbol{\theta}}(t)$ for every $t \in [0, \tau]$,
(i) Existence of a neighborhood $\boldsymbol{B}_0$ of $\boldsymbol{\beta}_0$ and $\boldsymbol{\Theta}_0$ of $\boldsymbol{\theta}_0$ such that $sup_{t \in [0,\tau], \boldsymbol{\beta} \in \boldsymbol{B}_0, \boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \| \boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) - \boldsymbol{s_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) \| \xrightarrow{p} \mathbf{0}$ as $n \to \infty$, for $j = 0, 1, 2$, where $\boldsymbol{s_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) = E\{\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)\}$ and $s_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) > 0$. The quantity $\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)$ is defined in the proof of Theorem 2.2.

Conditions (g) and (h) are used for the uniform consistency of $\pi_i^{\widehat{\boldsymbol{\theta}}}$ to $\pi_i^{\boldsymbol{\theta}_0}$ across all possible values of T_i in $[0, \tau]$. Note that $\pi_i^{\boldsymbol{\theta}} = G_{\boldsymbol{\theta}}(T_i)$. Condition (i) ensures that $\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)$ converges in probability, and that $\boldsymbol{e}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) = \frac{\boldsymbol{s_w}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)}{s_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)}$ is bounded.

The regularity conditions (j) and (k) below are needed in addition to (a)-(i) for the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$:
(j) For every $t \in [0, \tau]$ and $\boldsymbol{\theta}$ in a neighborhood $\boldsymbol{\Theta}_0$ of $\boldsymbol{\theta}_0$, $G_{\boldsymbol{\theta}}(t)$ is continuously differentiable in $\boldsymbol{\theta}$,
(k) Positive-definiteness of the matrices $\boldsymbol{A}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{I}(\boldsymbol{\theta})$ (defined in proof of Theorem 2.2).

Condition (j) is used to ensure the existence of $\boldsymbol{Q}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}})$ defined in the proof of Theorem 2.2, along with its convergence to $\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. Condition (k) ensures the existence of the inverses of the matrices $\boldsymbol{A}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{I}(\boldsymbol{\theta})$.

Theorem 2.1: $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}$ are consistent estimators of $\boldsymbol{\beta}_0$ as $n \to \infty$.

Proof of Theorem 2.1: The following proof holds for both $\widehat{\boldsymbol{w}} = \boldsymbol{w}_{np}$ and $\widehat{\boldsymbol{w}} = \boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}$. We therefore denote $\widehat{\boldsymbol{\pi}}^{np}$ and $\boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}}$ by $\widehat{\boldsymbol{\pi}}$ to simplify notation in this setting. The score equation 2.2 can be written as

$$\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \int_0^{\tau} \frac{\xi_i}{\pi_i} \{\boldsymbol{Z}_i(t) - \boldsymbol{E}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}; t)\} dN_i(t), \tag{2.3}$$

where $\boldsymbol{E}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}; t) = \sum_{j=1}^{N} \left\{ \frac{\xi_j}{\pi_j} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)} \boldsymbol{Z}_j(t) \right\} / \sum_{j=1}^{N} \left\{ \frac{\xi_j}{\pi_j} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)} \right\}$, and ξ_i is an indicator function set to 1 if subject i is observed, and 0 otherwise. Note that 2.3 consists of observations

from both truncated and observed subjects, with ξ_i = 0 for truncated subjects.

Let $\widehat{\boldsymbol{\beta}}_{\widehat{w}}$ be the solution to $\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}) = \mathbf{0}$. We will show that $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}})$ and $N^{-1}\boldsymbol{U}^*(\boldsymbol{\beta})$ converge to the same limit, where $\boldsymbol{U}^*(\boldsymbol{\beta}) = \sum_{i=1}^{N}\int_0^\tau \{\boldsymbol{Z}_i(t) - \boldsymbol{E}(\boldsymbol{\beta}; t)\}dN_i(t)$ is the complete case score function which includes all observations from both truncated and observed subjects, and $\boldsymbol{E}(\boldsymbol{\beta}; t) = \sum_{j=1}^{N}\{Y_j(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}_j(t)}\boldsymbol{Z}_j(t)\}/\sum_{j=1}^{N}\{Y_j(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}_j(t)}\}$. We then apply results from Lin (2000) and convex function theory to conclude that $\widehat{\boldsymbol{\beta}}_{\widehat{w}} \xrightarrow{p} \boldsymbol{\beta}_0$.

For $\pi(t) = P(U \leq t \leq V)$, Shen (2010a,2010b) proved that $\widehat{\pi}(t)$ converges uniformly in probability (with respect to t) to $\pi_0(t)$. Here $\widehat{\pi}(t)$ denotes the estimator of $\pi(t)$ under both parametric and nonparametric assumptions, and $\pi_0(t)$ is the true probability of observing a subject with survival time t. We will denote $\widehat{\pi}_i$ and $\pi_{0,i}$ as the estimated and true probability of observing a subject with survival time T_i, respectively. Note that $\widehat{\pi}_i = \widehat{\pi}(T_i)$ and $\pi_{0,i} = \pi_0(T_i)$.

We can re-express $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}})$ as

$$N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}) = N^{-1}\sum_{i=1}^{N}\int_0^\tau \pi_{0,i}^{-1}\xi_i\{\boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}; t)\}dN_i(t) \tag{2.4}$$

$$+ N^{-1}\sum_{i=1}^{N}\int_0^\tau \{\widehat{\pi}_i^{-1} - \pi_{0,i}^{-1}\}\xi_i\{\boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}; t)\}dN_i(t). \tag{2.5}$$

We will now state and prove a lemma used throughout the proof of Theorem 2.1.

Lemma: $N^{-1}\sum_{i=1}^{N}(\widehat{\pi}_i^{-1} - \pi_{0,i}^{-1})\boldsymbol{g}(\cdot) \xrightarrow{p} \mathbf{0}$ for any stochastically bounded function $\boldsymbol{g}$.

Proof: Let $H_N(\widehat{\boldsymbol{\pi}}; \cdot) = N^{-1}\sum_{i=1}^{N}(\widehat{\pi}_i^{-1} - \pi_{0,i}^{-1})\boldsymbol{g}(\cdot)$. We need to show that $\forall\epsilon > 0, \exists N \geq N_\epsilon$ such that $P(|H_N(\widehat{\boldsymbol{\pi}}; \cdot)| > \epsilon) < \epsilon$.

By the uniform consistency of $\widehat{\pi}(t)$ in t for $t \in [a_F, b_F]$ and the continuous mapping theorem, we have that $\widehat{\pi}^{-1}(t)$ is also uniformly consistent in t for $t \in [a_F, b_F]$. That is, $\forall\epsilon > 0, \exists N_{\epsilon_1}$ such that $N \geq N_{\epsilon_1} \implies P(\sup_{t\in[a_F,b_F]}|\widehat{\pi}^{-1}(t) - \pi_0^{-1}(t)| > \epsilon) < \epsilon$. Since $\boldsymbol{g}$ is stochastically bounded, $\exists M < \infty$ and N_{ϵ_2} such that $\forall\epsilon > 0, N > N_{\epsilon_2} \implies P(|\boldsymbol{g}(\cdot)| > M) < \epsilon$.

Let $N_\epsilon = \max(N_{\epsilon_1}, N_{\epsilon_2})$. Then for $N \geq N_\epsilon$,

$$P(|H_N(\widehat{\boldsymbol{\pi}};\cdot)|>\epsilon)=P(N^{-1}|\sum_{i=1}^{N}(\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1})\boldsymbol{g}(\cdot)|>\epsilon)$$

$$\leq P(N^{-1}\sum_{i=1}^{N}|(\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1})\boldsymbol{g}(\cdot)|>\epsilon)\leq P(N^{-1}\sum_{i=1}^{N}|\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1}|>\epsilon/M)$$

$$\leq P(\underset{i=1,..,N}{max}|\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1}|>\epsilon/M)\leq P(\sup_{t\in[a_F,b_F]}|\widehat{\pi}^{-1}(t)-\pi_0^{-1}(t)|>\epsilon/M)<\epsilon.$$

For $j=0,1$, we can write

$$\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)=N^{-1}\sum_{i=1}^{N}\pi_{0,i}^{-1}\xi_iY_i(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}_i(t)}\boldsymbol{Z}_i(t)^{\otimes j}+N^{-1}\sum_{i=1}^{N}(\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1})\xi_iY_i(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}_i(t)}\boldsymbol{Z}_i(t)^{\otimes j}$$

Since $Z_i(t)$ is stochastically bounded by regularity assumption (e), the term $\xi_iY_i(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}_i(t)}\boldsymbol{Z}_i(t)^{\otimes j}$ is also stochastically bounded. Application of the lemma therefore yields $\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)=\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta},\boldsymbol{\pi}_0;t)+o_p(1)$. Since $\boldsymbol{E_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)=\frac{\boldsymbol{S_w}^{(1)}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)}{\boldsymbol{S_w}^{(0)}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)}$, application of Slutsky's theorem yields $\boldsymbol{E_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}};t)=\boldsymbol{E_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0;t)+o_p(1)$.

We therefore have that 2.4 is equivalent to $N^{-1}\sum_{i=1}^{N}\int_0^{\tau}\pi_{0,i}^{-1}\xi_i\{\boldsymbol{Z}_i(t)-\boldsymbol{E_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0;t)+o_p(1)\}dN_i(t)$. Equation 2.5 is equivalent to $N^{-1}\sum_{i=1}^{N}\int_0^{\tau}\{\widehat{\pi}_i^{-1}-\pi_{0,i}^{-1}\}\xi_i\{\boldsymbol{Z}_i(t)-\boldsymbol{E_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0;t)+o_p(1)\}dN_i(t)$, which converges in probability to $\mathbf{0}$ by the lemma.

Finally, another application of Slutsky's theorem yields

$$N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}})=N^{-1}\sum_{i=1}^{N}\int_0^{\tau}\pi_{0,i}^{-1}\xi_i\{\boldsymbol{Z}_i(t)-\boldsymbol{E_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0;t)\}dN_i(t)+o_p(1)=N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0)+o_p(1).$$

Thus $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}})$ and $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\boldsymbol{\pi}_0)$ converge to the same limit. Since $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\pi_0)$ and $N^{-1}\boldsymbol{U}^*(\boldsymbol{\beta})$ converge to the same limit (Lin 2000), $N^{-1}\boldsymbol{U_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}})$ and $N^{-1}\boldsymbol{U}^*(\boldsymbol{\beta})$ must also converge to the same limit. Therefore our proposed estimating equation, $\boldsymbol{U_w}(\boldsymbol{\beta},\widehat{\boldsymbol{\pi}})$, is asymptotically equivalent to the standard Cox estimating equation containing all of the observations from the target sample, $\boldsymbol{U}^*(\boldsymbol{\beta})$. Since $\boldsymbol{U}^*(\boldsymbol{\beta})$ is maximized at $\boldsymbol{\beta}_0$ (Andersen and Gill 1982), it follows from convex function theory that $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{w}}}\xrightarrow{p}\boldsymbol{\beta}_0$ (Lin 2000). ■

Theorem 2.2 Under correct specification of the truncation distribution, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}-\boldsymbol{\beta}_0)$ is asymptoti-

cally normal as $n \to \infty$ with mean zero and covariance matrix
$\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1} \boldsymbol{V_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1}$.

To estimate the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$, we need some additional definitions. Let $w_i(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/\pi_i^{\boldsymbol{\theta}}$, where $p(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(U \leq T \leq V)$. Denote $\boldsymbol{\theta}_0$ as the true value of $\boldsymbol{\theta}$. For a $p \times 1$ vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 0}$ = 1, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2}$ denotes the $p \times p$ matrix $\boldsymbol{a}\boldsymbol{a}'$. Let

$$
\begin{aligned}
dM_i(\boldsymbol{\beta}; t) &= dN_i(t) - Y_i(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_i(t)} d\Lambda_0(t), \text{where } d\Lambda_0(t) \text{ is the hazard function,} \\
\boldsymbol{S_w}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) &= n^{-1} \sum_{i=1}^{n} w_i(\boldsymbol{\theta}) Y_i(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_i(t)} \boldsymbol{Z}_i(t)^{\otimes j}, j = 0, 1, 2, \\
\boldsymbol{E_w}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) &= \boldsymbol{S_w}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) / S_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t), \\
\boldsymbol{Q}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= E\Big[\int_0^{\tau} \frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} w_i(\tilde{\boldsymbol{\theta}}) \{ \boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) \} dM_i(\boldsymbol{\beta}; t) \Big]_{|_{\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}}}, \\
\boldsymbol{U}_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \boldsymbol{U}_{c_i}(\boldsymbol{\theta}), \text{where } \boldsymbol{U}_{c_i}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} log(k_{\boldsymbol{\theta}}(U_i, V_i)/\pi_i^{\boldsymbol{\theta}}), \\
\boldsymbol{I}(\boldsymbol{\theta}) &= -E\Big\{ n^{-1} \frac{\partial \boldsymbol{U_c}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \Big\}_{|_{\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}}}, \\
\boldsymbol{\phi_i}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int_0^{\tau} w_i(\boldsymbol{\theta}) \{ \boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}, \boldsymbol{\theta}; t) \} dM_i(\boldsymbol{\beta}; t) + \boldsymbol{Q}(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{U}_{c_i}(\boldsymbol{\theta}), \\
\boldsymbol{V_w}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= E\{ \boldsymbol{\phi}_i(\boldsymbol{\beta}, \boldsymbol{\theta})^{\otimes 2} \}, \\
\boldsymbol{A_w}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= E\Big[- \int_0^{\tau} w_i(\boldsymbol{\theta}) \Big\{ \frac{\boldsymbol{S_w}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)}{S_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)} - \frac{\boldsymbol{S_w}^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)^{\otimes 2}}{S_w^{(0)}(\boldsymbol{\beta}, \boldsymbol{\theta}; t)^2} \Big\} dN_i(t) \Big], \\
\boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \boldsymbol{A_w}(\boldsymbol{\beta}, \boldsymbol{\theta})^{-1} \boldsymbol{V_w}(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{A_w}(\boldsymbol{\beta}, \boldsymbol{\theta})^{-1}.
\end{aligned}
$$

The asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$ is given by $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1} \boldsymbol{V_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1}$. We can estimate $d\Lambda_0(t)$ by $d\widehat{\Lambda}_0(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}; t) = n^{-1} \sum_{j=1}^{n} w_j(\widehat{\boldsymbol{\theta}}) dN_j(t) / S_w^{(0)}\big(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}; t\big)$, and $dM_i(\boldsymbol{\beta}; t)$ by $d\widehat{M}_i(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}; t) = dN_i(t) - Y_i(t) e^{\widehat{\boldsymbol{\beta}}'_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}} \boldsymbol{Z}_i(t)} d\widehat{\Lambda}_0(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}; t)$. It can be shown that the remaining matrices defined above can be consistently estimated by their empirical counterparts, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ are replaced by their corresponding estimators $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$ and $\widehat{\boldsymbol{\theta}}$, respectively. It follows that the estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$, $\boldsymbol{\Sigma}\big(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}\big)$, is consistent for $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. The derivatives $\frac{\partial}{\partial \boldsymbol{\theta}} w_i(\boldsymbol{\theta})$, $\boldsymbol{U_c}(\boldsymbol{\theta})$, and $\frac{\partial \boldsymbol{U_c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ can be computed directly (when possible) or numerically.

Proof of Theorem 2.2: The proof proceeds by multiple applications of Taylor's theorem, results from empirical processes, the multivariate central limit theorem, and Slutsky's theorem. It can easily be shown that all of the matrices listed above can be consistently estimated by their empirical

counterparts using the strong law of large numbers and Slutsky's theorem.

Using simple algebra, we rewrite the score function given in equation 2.2 in Section 2.2, with parametric weights, as

$$\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{1}{\pi_i^{\widehat{\boldsymbol{\theta}}}} \{\boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}; t)\} d\widehat{M}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}; t).$$

Taylor expansion of $\boldsymbol{U_w}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ yields

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}} - \boldsymbol{\beta}_0) = \left\{ n^{-1} \frac{\partial \boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}} \right\}^{-1}_{|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}} n^{-\frac{1}{2}} \boldsymbol{U_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}),$$

where $\boldsymbol{\beta}^*$ lies between $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$ and $\boldsymbol{\beta}_0$ in $\mathbb{R}^p$. The uniform convergence in probability of $\boldsymbol{\pi}^{\widehat{\boldsymbol{\theta}}}$ to $\boldsymbol{\pi}^{\boldsymbol{\theta}_0}$, the consistency of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}$, and the continuous mapping theorem implies the uniform convergence of $\boldsymbol{S}_{\boldsymbol{w}}^{(j)}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}}; t)$ to $\boldsymbol{s}_{\boldsymbol{w}}^{(j)}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0; t)$ in t, for $j = 0, 1, 2$. Application of the strong law of large numbers yields

$$n^{-1} \frac{\partial \boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}}} \xrightarrow{p} \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0).$$

Applying the mean value theorem yields

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}} - \boldsymbol{\beta}_0) = \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1} n^{-\frac{1}{2}} \boldsymbol{U_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) + o_p(1).$$

Following similar arguments to Pan and Schuabal (2008), we set
$\boldsymbol{U_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = \boldsymbol{U_{w_1}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) + \boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}})$, where

$$\boldsymbol{U_{w_1}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{1}{\pi_i^{\boldsymbol{\theta}_0}} \{\boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}; t)\} d\widehat{M}_i(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}; t),$$

$$\boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \frac{1}{\pi_i^{\widehat{\boldsymbol{\theta}}}} - \frac{1}{\pi_i^{\boldsymbol{\theta}_0}} \right\} \{\boldsymbol{Z}_i(t) - \boldsymbol{E_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}; t)\} d\widehat{M}_i(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}; t).$$

Using results from empirical process theory, it can be shown that

$$n^{-\frac{1}{2}} \boldsymbol{U_{w_1}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} \frac{1}{\pi_i^{\boldsymbol{\theta}_0}} \{\boldsymbol{Z}_i(t) - \boldsymbol{e_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0; t)\} dM_i(\boldsymbol{\beta}_0; t) + o_p(1).$$

Applying Taylor expansion of $\boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ yields

$$n^{-\frac{1}{2}}\boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = n^{-\frac{1}{2}}\frac{\partial \boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ in $\mathbb{R}^q$. Applying Taylor expansion on $\boldsymbol{U}_c(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ yields $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \boldsymbol{I}_c(\boldsymbol{\theta}^*)^{-1}\boldsymbol{U}_c(\boldsymbol{\theta}_0)$, where $\boldsymbol{I}_c(\boldsymbol{\theta}) = -n^{-1}\frac{\partial \boldsymbol{U}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

Since $n^{-1}\frac{\partial \boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \xrightarrow{p} \boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ and $\boldsymbol{I}_c(\boldsymbol{\theta}^*) \xrightarrow{p} -E\{n^{-1}\frac{\partial \boldsymbol{U}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{I}(\boldsymbol{\theta}_0)$, we can re-express $n^{-\frac{1}{2}}\boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}})$ as

$$n^{-\frac{1}{2}}\boldsymbol{U_{w_2}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = n^{-\frac{1}{2}}\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\sum_{i=1}^{n}\boldsymbol{U}_{c_i}(\boldsymbol{\theta}_0) + o_p(1).$$

Combining the terms above, we now have the expression

$$n^{-\frac{1}{2}}\boldsymbol{U_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\phi}_i(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) + o_p(1),$$

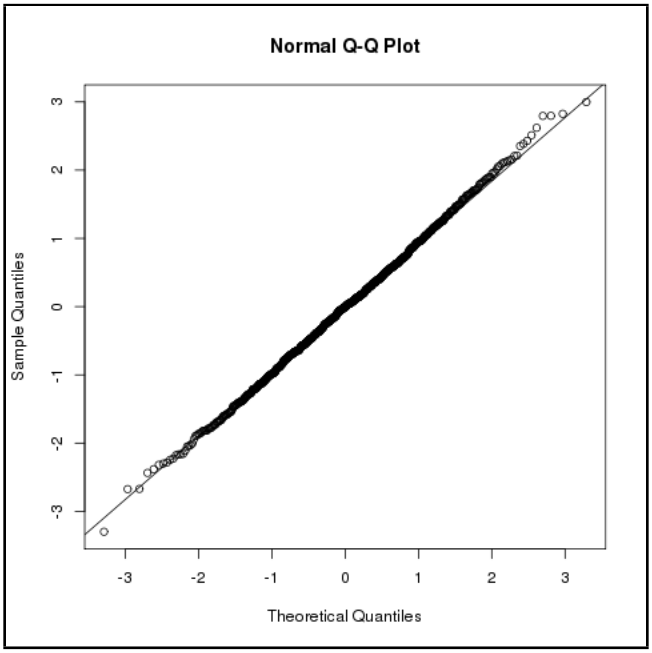
which is asymptotically equivalent to a sum of independent and identically distributed random vectors. Using the multivariate central limit theorem (van der Vaart 2000) yields $n^{-\frac{1}{2}}\boldsymbol{U_w}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\theta}}) \xrightarrow{D} N\{\boldsymbol{0}, \boldsymbol{V_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\}$. Finally, we apply the result that $n^{-1}\frac{\partial \boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}} \xrightarrow{p} \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ along with Slutsky's theorem to conclude $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}} - \boldsymbol{\beta}_0) \xrightarrow{D} N\{\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\}$, where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1}\boldsymbol{V_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\boldsymbol{A_w}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1}$. The covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ can be consistently estimated by $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{\widehat{\boldsymbol{\theta}}}}, \widehat{\boldsymbol{\theta}})$. ■.

The nature of $\widehat{\boldsymbol{\pi}}^{np}$ (e.g. no closed form) complicates the establishment of asymptotic normality for $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$. Thus we apply the bootstrap technique to get estimates of the standard error for $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ and corresponding confidence intervals. While asymptotic normality and the theoretical validity of the bootstrap are not formally established in this paper, our empirical evidence suggests that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ is asymptotically normal and that the bootstrap estimators are valid. The evidence for asymptotic normality is based on the Q-Q plot of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ from our simulation studies, shown in 2.2. Furthermore, these simulation studies show that the bootstrap standard errors of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}_{np}}$ are close to the observed sample standard deviations, and that the 95% confidence intervals based on the (bootstrap) percentile method result in coverage probabilities that are close to the nominal level of 0.95 (Table 2.1). In addition, previous simulations have shown the bootstrap confidence intervals match those based

on assuming normality.

Figure 2.2: Normal Q-Q plot of $T = \frac{\widehat{\beta}_{w_{np}} - \beta_0}{\widehat{\sigma}}$ from 1000 simulations under the truncation scenario for the second model described in Table 2.1

.



Here θ_1 = 0.40 and θ_2 = 0.25, and n=100. Here $\widehat{\sigma}$ is the standard error estimate of $\widehat{\beta}_{w_{np}}$, and is estimated using the simple bootstrap method.

2.4. Simulations

In this section we examine the performance of the proposed weighted estimators and compare them to the naïve unweighted estimator which ignores truncation. In all simulations, the survival times were generated from a proportional hazards model with hazard function $\lambda(t|Z) = \lambda_0(t)e^{\beta_0 Z}$, and follow a Weibull distribution with scale parameter $\rho = 0.1$ and shape parameter $\kappa = 1.2$. We set β_0 = 1, and generated the explanatory variable Z from a Unif[0,1] distribution. We simulated the left truncation time from a $c_1 Beta(\theta_1, 1)$ distribution and the right truncation time from a $c_2 Beta(1, \theta_2)$ distribution, with $c_1 = c_2 = 30$. We chose these distributions based on our data example. The assumption of the beta distribution for the truncation times in our data example was validated by a goodness-of-fit test (Section 2.5).

We conducted 1000 simulation repetitions with sample sizes of n = 50, 100, and 250. To obtain n observations after truncation, we simulated $N = \frac{n}{1-q}$ observations, where q is the proportion of

truncated data. For each simulation, we estimated the hazard ratio using the naïve unweighted estimator which ignores truncation ($\widehat{\beta}_{uw}$), the parametric weighted estimator ($\widehat{\beta}_{w_{\hat{\theta}}}$), the nonparametric weighted estimator ($\widehat{\beta}_{w_{np}}$), and the complete case estimator ($\widehat{\beta}_{cc}$) based on the full (truncated and non-truncated) sample. For these estimators, we calculated the estimated bias ($\widehat{\beta} - \beta_0$), observed sample standard deviations (SD), estimated standard errors ($\widehat{\mathrm{SE}}$), and the average empirical coverage probability of the 95% confidence intervals (Cov). We used 2000 bootstrap resamples to estimate the standard error and confidence interval of $\widehat{\beta}_{w_{np}}$.

Table 2.1 shows the results of the simulations described above. In the first model we set $\theta_1 = 0.06$ and $\theta_2 = 0.60$, which produced mild left and right truncation and a total of 20% of the observations truncated. In the second model we set $\theta_1 = 0.15$ and $\theta_2 = 1$, which produced moderate truncation from the left and right and a total of 40% of the observations truncated. In the third model we set $\theta_1 = 0.40$ and $\theta_2 = 0.25$, which produced heavy left truncation and mild right truncation and a total of 60% of the observations truncated. In the fourth model we set $\theta_1 = 0.50$ and $\theta_2 = 2.5$, which produced both heavy left and right truncation and a total of 80% of the observations truncated.

In all models, the weighted estimators $\widehat{\beta}_{w_{\hat{\theta}}}$ and $\widehat{\beta}_{w_{np}}$ had little bias, while the unweighted estimator $\widehat{\beta}_{uw}$ was biased. The observed sample standard deviations of $\widehat{\beta}_{w_{\hat{\theta}}}$ corresponded well with the standard error estimates based on asymptotic theory. The observed sample standard deviations of $\widehat{\beta}_{w_{np}}$ were accurately estimated by the bootstrap technique, and were slightly greater than those of $\widehat{\beta}_{w_{\hat{\theta}}}$. Both weighted estimators had coverage probabilities that were close to the nominal level of 0.95. All of these results held for both smaller (n=50) and larger (n=250) sample sizes. We note that the high coverage probabilities of $\widehat{\beta}_{uw}$ are an artifact of its large standard error relative to its bias, which led to wider confidence intervals for $\widehat{\beta}_{uw}$. In simulations where the standard error of $\widehat{\beta}_{uw}$ was small relative to its bias, the coverage probabilities of $\widehat{\beta}_{uw}$ did not come close to the nominal level (e.g. Table 2.2).

We now examine the bias of $\widehat{\beta}_{w_{np}}$ and $\widehat{\beta}_{uw}$ as a function of left and right truncation proportion (Figure 2.3). For the purpose of clarity we do not include $\widehat{\beta}_{w_{\hat{\theta}}}$ in Figure 2.3, but note that its bias was nearly identical to that of $\widehat{\beta}_{w_{np}}$. Even under mild truncation, $\widehat{\beta}_{uw}$ was biased, and this bias increased drastically as the proportion of right truncation increased. Here $\widehat{\beta}_{w_{np}}$ had little bias, regardless of truncation proportion.

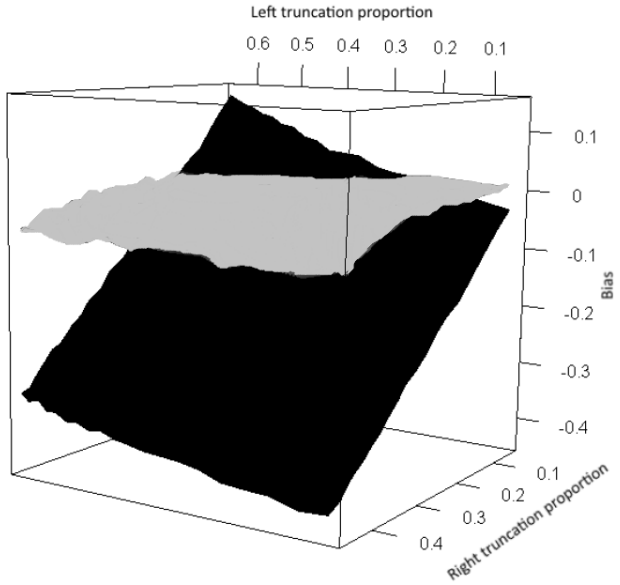
Table 2.1: Simulation results

q	n	Estimator	Bias	SD	$\widehat{\mathrm{SE}}$	Cov
0.20	50	$\widehat{\beta}_{uw}$	-0.081	0.574	0.545	0.943
	50	$\widehat{\beta}_{w_{\widehat{\theta}}}$	-0.011	0.616	0.552	0.927
	50	$\widehat{\beta}_{w_{np}}$	-0.015	0.616	0.620	0.937
	63	$\widehat{\beta}_{cc}$	0.003	0.504	0.475	0.943
	100	$\widehat{\beta}_{uw}$	-0.071	0.375	0.371	0.945
	100	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.003	0.405	0.374	0.940
	100	$\widehat{\beta}_{w_{np}}$	0.000	0.406	0.408	0.943
	125	$\widehat{\beta}_{cc}$	-0.005	0.340	0.328	0.941
	250	$\widehat{\beta}_{uw}$	-0.066	0.235	0.231	0.938
	250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.007	0.254	0.232	0.925
	250	$\widehat{\beta}_{w_{np}}$	0.004	0.254	0.250	0.945
	313	$\widehat{\beta}_{cc}$	0.011	0.205	0.205	0.951
0.40	50	$\widehat{\beta}_{uw}$	-0.031	0.548	0.536	0.957
	50	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.053	0.593	0.551	0.935
	50	$\widehat{\beta}_{w_{np}}$	0.045	0.605	0.626	0.934
	83	$\widehat{\beta}_{cc}$	0.047	0.423	0.404	0.949
	100	$\widehat{\beta}_{uw}$	-0.092	0.381	0.370	0.939
	100	$\widehat{\beta}_{w_{\widehat{\theta}}}$	-0.006	0.424	0.381	0.936
	100	$\widehat{\beta}_{w_{np}}$	-0.009	0.426	0.419	0.938
	167	$\widehat{\beta}_{cc}$	0.008	0.274	0.282	0.958
	250	$\widehat{\beta}_{uw}$	-0.084	0.235	0.231	0.927
	250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.005	0.263	0.235	0.922
	250	$\widehat{\beta}_{w_{np}}$	0.004	0.266	0.258	0.944
	417	$\widehat{\beta}_{cc}$	0.008	0.180	0.177	0.948
0.60	50	$\widehat{\beta}_{uw}$	0.139	0.562	0.542	0.937
	50	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.041	0.547	0.561	0.950
	50	$\widehat{\beta}_{w_{np}}$	0.034	0.555	0.580	0.939
	125	$\widehat{\beta}_{cc}$	0.005	0.338	0.326	0.947
	100	$\widehat{\beta}_{uw}$	0.122	0.374	0.372	0.949
	100	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.014	0.361	0.392	0.970
	100	$\widehat{\beta}_{w_{np}}$	0.011	0.363	0.382	0.955
	250	$\widehat{\beta}_{cc}$	-0.004	0.234	0.228	0.936
	250	$\widehat{\beta}_{uw}$	0.111	0.244	0.232	0.911
	250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.013	0.234	0.249	0.964
	250	$\widehat{\beta}_{w_{np}}$	0.005	0.237	0.234	0.937
	625	$\widehat{\beta}_{cc}$	0.006	0.150	0.144	0.947
0.80	50	$\widehat{\beta}_{uw}$	-0.127	0.560	0.538	0.937
	50	$\widehat{\beta}_{w_{\widehat{\theta}}}$	-0.015	0.666	0.633	0.940
	50	$\widehat{\beta}_{w_{np}}$	-0.004	0.724	0.701	0.947
	250	$\widehat{\beta}_{cc}$	0.008	0.226	0.233	0.961
	100	$\widehat{\beta}_{uw}$	-0.122	0.373	0.367	0.940
	100	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.013	0.472	0.456	0.924
	100	$\widehat{\beta}_{w_{np}}$	0.016	0.493	0.472	0.949
	500	$\widehat{\beta}_{cc}$	0.006	0.162	0.164	0.955
	250	$\widehat{\beta}_{uw}$	-0.163	0.236	0.228	0.878
	250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	-0.021	0.316	0.328	0.913
	250	$\widehat{\beta}_{w_{np}}$	-0.019	0.315	0.294	0.927
	1250	$\widehat{\beta}_{cc}$	0.000	0.104	0.103	0.949

q is proportion of truncated observations, n is size of observed sample. $\widehat{\beta}_{uw}$ denotes naïve unweighted estimator, $\widehat{\beta}_{w_{\widehat{\theta}}}$ denotes proposed parametric weighted estimator, $\widehat{\beta}_{w_{np}}$ denotes proposed nonparametric weighted estimator, $\widehat{\beta}_{cc}$ denotes unattainable complete case estimator based on both truncated and non-truncated observations. SD is empirical standard deviation of estimates across simulations, $\widehat{SE}$ is average of estimated standard errors, Cov is coverage of 95% confidence intervals. True value of β is 1.

We also examined the robustness of $\widehat{\beta}_{w_{\widehat{\theta}}}$ under misspecification of the truncation distribution in Table 2.2. In this setting, $\widehat{\beta}_{w_{\widehat{\theta}}}$ was biased. Here $\widehat{\beta}_{w_{np}}$ still had little bias, as $\widehat{\beta}_{w_{np}}$ makes no distributional

Figure 2.3: Comparing bias and MSE (mean-squared error) of estimators



Bias of the unweighted estimator $\widehat{\beta}_{uw}$ (black) and nonparametric weighted estimator $\widehat{\beta}_{w_{np}}$ (gray). Left truncation time simulated from a $c_1 Beta(\theta_1, 1)$ distribution, right truncation time simulated from a $c_2 Beta(1, \theta_2)$ distribution, with $c_1 = c_2 = 30$. Here θ_1 ranges from 0.025 to 0.50 which results in a range of 5% to 65% truncation from the left, and θ_2 ranges from 0.25 to 5 which results in a range of 5% to 45% truncation from the right. The remaining settings are kept the same as in Table 2.1, with $n = 250$.

assumptions for the truncation times.

Table 2.2: Simulation results under misspecification of the truncation distribution

q	n	Estimator	Bias	SD	$\widehat{\text{SE}}$	Cov
0.50	250	$\widehat{\beta}_{uw}$	-0.198	0.233	0.229	0.849
	250	$\widehat{\beta}_{w_{\hat{\theta}}}$	-0.053	0.296	0.241	0.876
	250	$\widehat{\beta}_{w_{np}}$	-0.029	0.360	0.318	0.930
	500	$\widehat{\beta}_{cc}$	0.003	0.168	0.164	0.939
0.40	250	$\widehat{\beta}_{uw}$	-0.095	0.235	0.230	0.923
	250	$\widehat{\beta}_{w_{\hat{\theta}}}$	-0.165	0.237	0.258	0.919
	250	$\widehat{\beta}_{w_{np}}$	-0.002	0.306	0.288	0.938
	417	$\widehat{\beta}_{cc}$	-0.002	0.173	0.176	0.960
0.35	250	$\widehat{\beta}_{uw}$	-0.245	0.235	0.229	0.795
	250	$\widehat{\beta}_{w_{\hat{\theta}}}$	-0.175	0.267	0.249	0.866
	250	$\widehat{\beta}_{w_{np}}$	-0.034	0.427	0.356	0.920
	385	$\widehat{\beta}_{cc}$	0.002	0.182	0.184	0.953

q is the proportion of observations missing due to truncation and n is the size of the observed sample. $\widehat{\beta}_{uw}$ denotes the naïve unweighted estimator, $\widehat{\beta}_{w_{\hat{\theta}}}$ denotes the proposed parametric weighted estimator, $\widehat{\beta}_{w_{np}}$ denotes the proposed nonparametric weighted estimator, and $\widehat{\beta}_{cc}$ denotes the unattainable complete case estimator based on both truncated and non-truncated observations. SD is the empirical standard deviation of estimates across simulations, $\widehat{SE}$ is the average of the estimated standard errors, Cov is the coverage of 95% confidence intervals. The true value of β is 1.

The simulations above assumed U and V are independent. In some cases, V can be expressed

as $V = U + d_0$, where d_0 can be random or constant. To assess the performance of our proposed estimators under this dependent truncation structure, we conducted a simulation study in Table 2.3. The results are similar to those presented in Table 2.1.

Table 2.3: Simulation results under dependent truncation structure $V = U + d_0$.

θ_U	γ_{d_0}	q	n	Estimator	Bias	SD	$\widehat{\text{SE}}$	Cov
0.15	30	0.33	250	$\widehat{\beta}_{uw}$	-0.051	0.235	0.230	0.937
			250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.012	0.251	0.245	0.946
			250	$\widehat{\beta}_{w_{np}}$	0.009	0.253	0.251	0.941
			374	$\widehat{\beta}_{cc}$	0.002	0.187	0.187	0.956
0.25	20	0.47	250	$\widehat{\beta}_{uw}$	-0.088	0.231	0.231	0.932
			250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.019	0.268	0.305	0.964
			250	$\widehat{\beta}_{w_{np}}$	0.017	0.271	0.280	0.957
			472	$\widehat{\beta}_{cc}$	0.004	0.168	0.167	0.951
0.35	20	0.56	250	$\widehat{\beta}_{uw}$	-0.037	0.248	0.232	0.930
			250	$\widehat{\beta}_{w_{\widehat{\theta}}}$	0.018	0.273	0.261	0.930
			250	$\widehat{\beta}_{w_{np}}$	0.011	0.276	0.270	0.931
			569	$\widehat{\beta}_{cc}$	0.004	0.155	0.154	0.947

$U \sim 20 \times Beta(\theta_U, 1)$, $d_0 \sim Unif[0, \gamma_{d_0}]$. *The remaining settings were kept the same as in the simulations in Section 2.4 of the paper.* q *is the proportion of observations missing due to truncation and* n *is the size of the observed sample.* $\widehat{\beta}_{uw}$ *denotes the naïve unweighted estimator,* $\widehat{\beta}_{w_{\widehat{\theta}}}$ *denotes the proposed parametric weighted estimator,* $\widehat{\beta}_{w_{np}}$ *denotes the proposed nonparametric weighted estimator, and* $\widehat{\beta}_{cc}$ *denotes the unattainable complete case estimator based on both truncated and non-truncated observations. SD is the empirical standard deviation of estimates across simulations,* $\widehat{SE}$ *is the average of the estimated standard errors, Cov is the coverage of 95% confidence intervals. The true value of* β *is 1.*

2.5. Application to Alzheimer's Disease Study

We illustrate our method by considering an autopsy-confirmed AD study conducted by the Center for Neurodegenerative Disease Research at the University of Pennsylvania. The target population for the research purposes of this study consists of all subjects with AD symptom onset before 2012 that met the study criteria and therefore would have been eligible to enter the center. Our observed sample contains all subjects who entered the center between 1995 and 2012, and had an autopsy performed before 2012. Thus one criterion for a subject to be included in our sample is that they did not succumb to AD before they entered the study, yielding left truncated data. In addition, our sample only contains subjects who had an autopsy-confirmed diagnosis of AD, and therefore we have no knowledge of subjects who live past the end of the study. Thus our data is also right truncated. Our data consists of n=47 subjects, all of whom have event times. The event time of interest is the survival time (T) from AD symptom onset. The left truncation time (U) is the time between the onset of AD symptoms and entry into the study (i.e. initial clinic visit). The right truncation time (V) is the time between the onset of AD symptoms and the end of the study, which is taken to be July 15, 2012. Due to double truncation, we only observe subjects with $U \leq T \leq V$.

Our motivation for studying the effect of education on survival in AD is that education serves as a proxy for cognitive reserve (CR). CR theorizes that individuals develop cognitive strategies and neuronal connections throughout their lives through experiences such as education and other forms of mental engagement (Valenzuela and Sachdev, 2007). For example, CR may have a protective role in the brain, and therefore lengthen survival during the course of the disease (Ientile et al., 2013). Paradise et al. (2009) and Meng and D'Arcy (2012) failed to detect an effect of education on survival from AD symptom onset. However the studies included in their meta-analyses did not consist of populations with autopsy-confirmed AD.

Here we assess the effect of education on survival time in our autopsy-confirmed cohort, where education is measured by years of schooling. The median years of education in this cohort is 16 years. Comparing the low education group ($<$ 16 years) and high education group ($\geq$ 16 years) on the variables of interest revealed no significant differences (Table 2.4).

Table 2.4: Comparing low education ($<$ 16 years) and high education ($\geq$ 16 years) groups

Variable	Low education (n=15) mean (sd)	High education (n=32) mean (sd)	Test statistic	p-value
Age Onset	61.8 (10.5)	63.2 (12.9)	$t_{45} = -0.37$	0.712
Survival time	8.7 (3.4)	7.9 (3.2)	$t_{45} = 0.80$	0.430
Time to study entry	3.4 (1.71)	2.7 (1.5)	$t_{45} = 1.37$	0.177
Time to end of study	13.3 (2.8)	12.6 (4.7)	$t_{45} = 0.58$	0.563
Male (%)	53	72	$\chi^2_1 = 1.56$	0.211

Survival time, time to study entry, and time to end of study are measured in years from AD symptom onset.

Since our data is doubly truncated, we apply the Cox regression model using the proposed weighted estimating equation approach. We check the assumption of independence between the truncation and survival times in the observable region $U \leq T \leq V$ using the conditional Kendall's tau proposed by Martin and Betensky (2005). The resulting p-value is 0.10, and therefore we do not have enough evidence to reject the null hypothesis that the observed survival and truncation times are independent. We justify the identifiability constraints, $a_{H_U} < a_F \leq a_{H_V}$ and $b_{H_U} \leq b_F < b_{H_V}$, in Section 2.4.1 below.

We adjust for double truncation using both parametric and nonparametric weights. The parametric weights are estimated under the assumption that $U \sim c_1 Beta(\alpha_1, \beta_1)$ and $V \sim c_2 Beta(\alpha_2, \beta_2)$, where $c_1 = 20$ and $c_2 = 40$. Under these parametric assumptions, we have $\widehat{\alpha}_1 = 2.6, \widehat{\beta}_1 = 13.8$ and

$\widehat{\alpha}_2 = 3.0, \widehat{\beta}_2 = 9.7$. To check our assumption of the beta distribution, we test the null hypothesis H_0 : $K(u,v) = K_{\boldsymbol{\theta}}(u,v)$, where $\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2)$. Here the parametric joint cumulative distribution function $K_{\boldsymbol{\theta}}(u,v)$ = $I_{u/c_1}(\alpha_1,\beta_1) \times I_{v/c_2}(\alpha_2,\beta_2)$, where $I_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$. As described by Moreira, de Ũna-Álvarez, and Van Keilegom (2014), we can test H_0 using a Kolmogorov-Smirnov type test statistic $D_n = sup_{u,v\in\mathbb{R}}|K_n(u,v) - K_{\widehat{\boldsymbol{\theta}}}(u,v)|$, where $K_n(u,v)$ is the NPMLE of $K(u,v)$ (Shen, 2010a). This yields a p-value of 0.60, and therefore we do not have enough evidence against the beta distribution assumption for the truncation times.

Table 2.5 displays the results from the Cox regression model using no weights, parametric weights, and nonparametric weights. The effects of age at AD symptom onset and male on survival are nearly twice as large in the weighted models relative to the unweighted model, but these effects are only significant under parametric assumptions. When we do not account for double truncation, there is no effect of education on survival ($\widehat{\beta}_{uw}$ = 0; 95% CI: [-0.11,0.12]). When we account for double truncation, higher education is associated with increased survival under parametric weights ($\widehat{\beta}_{\boldsymbol{w}_{\widehat{\theta}}}$ = -0.07; 95% CI: [-0.20,0.06]) and nonparametric weights ($\widehat{\beta}_{\boldsymbol{w}_{np}}$ = -0.06; 95% CI: [-0.29,0.19]). However the confidence intervals for both $\widehat{\beta}_{\boldsymbol{w}_{\widehat{\theta}}}$ and $\widehat{\beta}_{\boldsymbol{w}_{np}}$ contain 0.

Table 2.5: Application: Education on survival in AD

	Unweighted		Parametric weights		Nonparametric weights	
Predictor	$\widehat{\beta}_{uw}$ ($\widehat{\text{SE}}$)	95% CI	$\widehat{\beta}_{\boldsymbol{w}_{\widehat{\theta}}}$ ($\widehat{\text{SE}}$)	95% CI	$\widehat{\beta}_{\boldsymbol{w}_{np}}$ ($\widehat{\text{SE}}$)	95%CI
Age Onset	0.03 (0.03)	(-0.01, 0.06)	0.05 (0.02)	(0.00, 0.09)	0.05 (0.03)	(-0.02, 0.12)
Male	0.45 (0.34)	(-0.21, 1.11)	1.01 (0.49)	(0.06, 1.97)	0.95 (0.61)	(-0.36, 2.18)
Education	0.00 (0.06)	(-0.11, 0.12)	-0.07 (0.07)	(-0.20, 0.06)	-0.06 (0.11)	(-0.29, 0.19)

2.5.1. Justification of identifiability constraints

Here we justify that the identifiability constraints given in Section 2.2.1, $a_{H_U} < a_F \leq a_{H_V}$ and $b_{H_U} \leq b_F < b_{H_V}$, hold in our data example. First we introduce some notation. Denote τ as the end of study date, τ_E as the study entry date, and τ_A as the date of symptom onset. Note that τ is the same for all subjects, while τ_E and τ_A can differ among subjects. The left truncation time is defined as $U = \tau_E - \tau_A$, and the right truncation time is defined as $V = \tau - \tau_A$.

Subjects can theoretically enter the center at the time of AD symptom onset, but not before. Therefore the smallest possible left truncation time is U=0, and thus $a_{H_U} = 0$. Since recruitment of subjects into the center stops one week prior to the end of the study, the smallest possible right truncation time is V = 1 week (or $V \approx 0.019$ years). Therefore $a_{H_V} \geq 0.019$. Since subjects can die

from AD within a week of symptom onset, and we assume that subjects cannot die on the day of symptom onset, $P(T \leq t) > 0$ for some $t \in (0, 0.019)$, where t is measured in years. We therefore have that $0 < a_F < 0.019$, and thus the constraint $a_{H_U} < a_F \leq a_{H_V}$ is satisfied.

Because subjects are not expected to enter the study more than 20 years after symptom onset, the assumption $b_{H_U} \leq 20$ is reasonable in practice. Since subjects with AD can live past 20 years after symptom onset, $P(T \geq 20) > 0$, and thus $b_F \geq 20$. Our study recruited subjects from 1995 to 2012, and subjects with AD symptom onset before 1995 were included in the study. Since the study recruited subjects over a 17 year period, we have that $b_{H_V} = b_F + 17$ and thus $b_F < b_{H_V}$. To see why this is so, note that a subject with a survival time $T = b_F$ could have theoretically had symptom onset in the year 1995 - b_F. For example, if $b_F = 20$, a subject could have entered the study in 1975, in which case their right truncation time would be 37 years. Therefore it is reasonable to assume that the constraint $b_{H_U} \leq b_F < b_{H_V}$ is satisfied.

A violation of these assumptions implies that we cannot observe a particular subset $S \subseteq [a_F, b_F]$ of the survival times, which violates the positivity assumption and may lead to unstable estimators. In other words, a violation of the identifiability constraints implies that $\pi(t) = 0$ for a particular survival time $t \in S$. For example, when $a_{H_U} < a_F$ is violated, we have that for all $t \in S = [a_F, a_{H_U}]$, $\pi(t) = P(U \leq t \leq V) = \int_t^{b_{H_V}} \int_{a_{H_U}}^{t} K(du, dv) = K(t, b_{H_V}) - K(t, t) = 0$ (note that the terms $K(a_{H_U}, b_{H_V})$ and $K(a_{H_U}, t)$ are 0 by the definition of a_{H_U}). To see why $\pi(t) = 0$, recall that $t \in S$ implies that $t < a_{H_U}$, and thus $P(U \leq t) = 0$. Since $K(u, v) = P(U \leq u, V \leq v)$, we have that $K(t, b_{H_V}) = K(t, t) = 0$ when $t < a_{H_U}$.

In practice, a violation of the identifiability constraints could happen if there is a mediating event that must occur before a subject enters a clinic or study. Suppose that this mediating event occurs only after the onset of symptoms, say δ units of time, so that $a_{H_U} \geq \delta$. If there is a non-zero probability that subjects can die between the onset of symptoms and this mediating event, then $a_F < \delta$ and thus $a_F < a_{H_U}$. In this case $\pi(t) = 0$ for all $t \in [a_F, \delta)$ and we would never observe these subjects, which will lead to invalid inference on the target population.

The justification of the identifiability constraints in our data, along with the study design, provides evidence that the positivity assumption holds for our observed sample. To demonstrate this, we will show that if the identifiability constraints hold, then $\pi(t) = 0$ if and only if $P(U \leq t | V \geq t) = 0$ (or

equivalently $P(V \geq t|U \leq t) = 0$) for some $t \in [a_F, b_F]$. By Bayes rule,

$$\pi(t) = P(U \leq t \leq V) = P(U \leq t, V \geq t) = P(V \geq t|U \leq t) \cdot P(U \leq t) = P(U \leq t|V \geq t) \cdot P(V \geq t).$$

The constraint $a_{H_U} < a_F$ implies that $P(U \leq t) > 0$ for all $t \geq a_F$, and the constraint $b_F < b_{H_V}$ implies that $P(V \geq t) > 0$ for all $t \leq b_F$. Therefore when the identifiability constraints hold, the positivity assumption can only be violated if $P(U \leq t|V \geq t) = 0$ (or equivalently $P(V \geq t|U \leq t) = 0$) for some $t \in [a_F, b_F]$. If such t did exist, say t', such that $\pi(t') = 0$, then $P(U \leq t'|V \geq t') = 0$ would imply that all subjects who have a right truncation time that exceeds t' years could not have entered the study within t' years of symptom onset. For example, if $t' = 10$, then all subjects with $V \geq 10$ must have had AD symptom onset before 2002 (recall that V is the time from symptom onset to the year 2012). If $P(U \leq 10|V \geq 10) = 0$, then subjects with AD symptom onset after 2002 would be unable to enter the study within 10 years after symptom onset. However this is not possible under our study design, since the criteria for entry did not change throughout the course of the study.

2.6. Discussion

We proposed a weighted estimating equation approach to adjust the Cox regression model under double truncation, by weighting the subjects in the score equation of the Cox partial likelihood by the inverse of the probability that they were observed (i.e. *not* truncated). The probability of being observed was estimated both parametrically and nonparametrically by methods introduced in Shen (2010; 2010) and Moreira and de Ũna-Álvarez (2010), and did not require any contribution from missing subjects. The proposed hazard ratio estimators are consistent. The simulation studies confirmed that the proposed estimators have little bias, while the naïve estimator which ignores truncation is biased. The parametric weighted estimator is asymptotically normal, and a consistent estimator of its asymptotic variance is provided. Our simulations showed that the bootstrap estimate of the standard error for the nonparametric weighted estimator matched the observed sample standard deviation.

The proposed estimators have little bias in practical settings, which has useful implications in observational studies. One example is AD - a severe neurodegenerative disorder which has devastating effects for patients and their caregivers. Thus any knowledge of factors associated with extending

survival from AD symptom onset can have a great impact on society. In this paper, we assessed the effect of education on survival in subjects with autopsy-confirmed AD. Our method is critical for analyzing data of this sort, since autopsy confirmation leads to doubly truncated survival times, which can result in biased hazard ratio estimators. While AD studies that do not use autopsy confirmation avoid double truncation, the conclusions based on these studies may be unreliable due to the inaccuracy of clinical diagnosis. This may explain the inconclusive findings of the two meta-analyses conducted by Paradise et al. (2009) and Meng and D'Arcy (2012), who used studies with clinically diagnosed AD subjects to examine the effect of education on survival. Using our proposed method on an autopsy-confirmed AD study found that higher education was associated with increased survival. However, these effects were not statistically significant. This may be due to our small sample size and the fact that our sample was highly educated (range = 12 - 20 years). When double truncation was ignored, we found no effect of education on survival.

The consistency of the estimated selection probabilities used in our proposed method rests on the assumption of independence between the survival and truncation times in the observable region. A violation of this assumption may lead to biased hazard ratio estimators. Currently, we are not aware of any methods to adjust for violations of this assumption. Because the estimation procedure for the selection probabilities does not make use of the assumed relationship between the survival time and covariates, this independence assumption cannot be relaxed simply by covariate adjustment in the Cox model. However, when conditional independence on discrete covariates holds, we can stratify the data based on the levels of the covariates, and then estimate the weights independently within each stratum. In this situation, conditional independence can be tested by applying the conditional Kendall's tau (Martin and Betensky, 2005) within each stratum. However, this approach may not be practical if the number of strata is large. Future work is thus needed to relax the independence assumption.

Currently there are no closed form estimates for the nonparametric selection probabilities, which complicate the development of asymptotic properties for the nonparametric weighted estimator. While our simulations show that the nonparametric weighted estimator appears to satisfy asymptotic normality, an extension to our method is to formally prove this result. Furthermore, the theoretical validity of the bootstrap estimators needs to be established. Finally, the proposed method assumes that no censoring is present in the data. While right censoring is uncommon under dou-

ble truncation, interval censoring could be present in the data (Bilker and Wang, 1996; Martin and Betensky, 2005). Future work would thus be needed to extend our methods in the presence of interval censored data.

While weighting leads to consistent estimators, it may also lead to an increase in the variance of these estimators in certain cases. In practice, an investigator may wonder whether it is worth adjusting for double truncation. We recommend using the proposed weighted estimators since they are consistent and perform well in finite samples, while the naïve estimator can be biased even in cases of mild truncation. When the truncation is severe, the naïve estimator can be heavily biased. However, severe truncation may produce large weights which can lead to an increase in the standard error of the weighted estimators. Therefore if the estimated weights are large, we recommend performing a sensitivity analysis by truncating the weights as described in Seaman and White (2013).

CHAPTER 3

COX REGRESSION MODEL UNDER DEPENDENT TRUNCATION

3.1. Introduction

Truncation is a statistical phenomenon that has been shown to occur in a wide range of applications, including survival analysis, epidemiology, economics, and astronomy. Individuals who are subject to truncation provide no information to the investigator. *Left truncation* occurs when data is only recorded for individuals whose event time exceeds a random time (i.e. left truncation time). Under left truncation, individuals with smaller event times are less likely to be observed, resulting in a study sample that is biased towards larger event times and risk factors associated with larger event times. *Right truncation* occurs when data is only recorded for individuals whose event time proceeds a random time (i.e. right truncation time). Under right truncation, individuals with larger event times are less likely to be observed, resulting in a study sample that is biased towards smaller event times and risk factors associated with smaller event times. When both left and right truncation are present, this is known as *double truncation*.

Double truncation is inherent in autopsy-confirmed studies of neurodegenerative diseases Rennert and Xie, 2017. Left truncation occurs because individuals enter the study after the onset of the disease, and therefore those who succumb to the disease before they enter the study are unobserved. The right truncation occurs because individuals who live past the end of the study date do not receive a pathological diagnosis of the disease. Since these subjects cannot be definitively diagnosed with a particular disease, they are excluded from the autopsy-confirmed study sample and therefore provide no information to the investigator. This is contrary to censored individuals, who provide partial information about their survival time. We note, however, that right censoring is not possible in autopsy-confirmed studies, since any individual who has an autopsy performed will also have a known survival time. This truncation scheme is illustrated in Figure 2.1, where only individuals whose time of death falls between the study entry time and end of study time are observed.

The aim of our data analysis is to get accurate estimates of the effect of risk factors on survival from disease symptom onset in subjects with autopsy-confirmed Alzheimer's disease (AD), the

most common neurodegenerative disease. Because individuals with shorter survival times are less likely to enter the study, left truncation leads to a study sample that is biased towards larger survival times and risk factors associated with larger survival times. Similarly, individuals with longer survival times are more likely to live past the end of the study, and therefore right truncation leads to a study sample that is biased towards smaller survival times and risk factors associated with smaller survival times. If double truncation is not accounted for, then the regression coefficient estimators from the Cox regression model Cox, 1972 will be biased.

Methods to handle double truncation have recently started gaining traction in the literature. In 2017, three methods were published to adjust the Cox model under double truncation (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017). The estimation procedure for all three methods rely on estimating the joint distribution of the left and right truncation times, which is used to compute the probability that a subject is observed (i.e. not truncated). These probabilities are then used as weights or offsets in the Cox model. However, the estimation of the truncation distribution relies on the assumption of independence between the observed survival and truncation times, which may not a reasonable assumption in practice. For example, according to the Alzheimer's association and discussions with our clinical investigators, factors such as lower age of symptom onset, depression, and stress are associated with delayed study entry. Since these factors are associated with survival, this induces a dependence between the left truncation times and survival times. As shown in the simulation studies in Section 3.3, the regression coefficient estimators from (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017) are sensitive to violations of this independence assumption. Therefore, the existing literature is unable to address the unique challenges present in our study.

In this paper, we propose a novel method to relax the assumption of independence between the observed survival and truncation times in the Cox proportional hazards model under left, right, and double truncation. Specifically, by conditioning on the observed truncation times, our method relaxes the independence assumption to an assumption of *conditional independence*. Treating the truncated survival times as missing, we introduce an expectation-maximization (EM) algorithm to estimate the regression coefficients and baseline hazard rates. This approach, which completely avoids the estimation of the truncation distribution, yields consistent and asymptotically normal regression coefficient estimators. We show through extensive simulation studies that our proposed

estimators have little bias in small samples, while the estimators based on the methods introduced in (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017), and the standard model which ignores double truncation, can be heavily biased under violations of the independence assumption. We show that even if the independence assumption is satisfied, our proposed method performs as well as the existing approaches. We illustrate our method by analyzing the effect of cognitive reserve on survival in an autopsy-confirmed AD cohort.

Cognitive reserve (CR) is a widely used hypothetical construct intended to account for individual differences in cognitive decline and clinical manifestations of dementia among individuals with AD (Meng and DArcy, 2012; Stern, 2012). CR hypothesizes that individuals develop cognitive strategies and neural connections throughout their life times through experience such as occupation, education, and other forms of mental engagement (Valenzuela and Sachdev, 2007). This may modulate the effects of AD because of compensatory strategies obtained from a higher level of professional performance or a good education (Sanchez et al., 2011). For example, CR may have a protective role in the brain and therefore lengthen survival from disease symptom onset (Ientile et al., 2013).

Occupation, often used as a proxy for CR, has been shown to modulate survival in healthy aging and AD (Massimo et al., 2015). Several studies in healthy aging have examined the possible protective influence of higher occupational attainment on survival (Andel, Silverstein, and Kareholt, 2014; Correa Ribeiro, Lopes, and Loureno, 2013; Enroth et al., 2014). However other studies have shown that for individuals with AD, those with a higher occupational attainment had a higher mortality rate than those with a lower occupation attainment (Stern et al., 1999, 1995). The caveat to previous studies assessing the effect of occupation on survival is that most consisted of populations with clinically diagnosed AD subjects, which can be unreliable (Beach et al., 2012). Due to the inaccuracy of clinical diagnosis of AD, autopsy-confirmation is used for a definitive diagnosis (Grossman and Irwin, 2016). Without an accurate diagnosis of AD, any estimates of factors affecting survival are not reliable. In this paper, we aim to get improved estimates of the effect of occupation on survival from an autopsy-confirmed AD sample, adjusting for both truncation and dependence.

The remainder of this paper is organized as follows. In Section 3.2 we introduce the proposed EM method, including the estimation procedure and the large sample properties of the resulting estimators. In Section 3.3, we conduct a simulation study to assess the finite sample performance

of the proposed estimators under dependent truncation. In Section 3.4, we apply the proposed method to estimate the effect of occupation on survival in individuals with autopsy-confirmed AD. Discussion and concluding remarks are given in Section 3.5. Proofs of the large sample results are outlined in the Appendix.

3.2. Methods

We first introduce notation and assumptions. Let T denote the survival time of interest (e.g. survival time from disease symptom onset), L denote the left truncation time (e.g. time from disease symptom onset to entry into the study), R denote the right truncation time (e.g. time from disease symptom onset to the end of study date), and $\boldsymbol{Z}$ denote a $p \times 1$ vector of covariates. Let N denote the size of the target population – the population that would have been observed had there been no truncation present in the study. Due to double truncation, we only observe $(T_i, L_i, R_i, \boldsymbol{Z}_i)$ for $i = 1, ..., n \leq N$ individuals who live long enough to enter the study (i.e. $T \geq L$) and do not live past the end of the study (i.e. $T \leq R$). Here we have denoted the population random variables from the target population without subscripts, and the sampling random variables from the observed sample with subscripts.

The proportional hazards model (Cox, 1972) is considered the standard regression model for analyzing traditional right-censored survival data. The model assumes that the covariate-specific hazard function is given by $\lambda_{\boldsymbol{Z}}(t) = \lambda(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z})$, where $\boldsymbol{\beta}$ is a $p \times 1$ regression parameter vector, and $\lambda(t)$ is the baseline hazard function and is unspecified. When the survival data are subject to selection bias, Cox's partial likelihood approach (Cox, 1975) cannot be directly applied. This is because the observed data are not a representative sample of the target population, and therefore the observed, biased data do not follow the model that is assumed for the unbiased data from the target population. When the data is biased due to double truncation, the distribution of the observed survival time T_i is given by:

$$P(T_i \leq t|\boldsymbol{Z}_i) = P(T \leq t|\boldsymbol{Z}_i, L \leq T \leq R) = \frac{P(T \leq t, L \leq T \leq R|\boldsymbol{Z}_i)}{P(L \leq T \leq R|\boldsymbol{Z}_i)} \neq P(T \leq t|\boldsymbol{Z}_i),$$

which differs from the distribution of the survival time T from the target population. Therefore the resulting estimates of the regression coefficients based on data from the observed sample will be biased estimators of the regression coefficients from the target population.

Under the assumption of independence between the survival and truncation times, (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017) adjust for double truncation by estimating the probability that a subject with survival time T_i is observed, defined by $\widehat{\pi}_i = \widehat{P}(L \leq T \leq R|T = T_i)$, $i = 1, ..., n$. These probabilities are then used as weights or offsets in the Cox regression model. For example, under double truncation and independence between the survival times and truncation times, Rennert and Xie (2017) consistently estimate the true $p \times 1$ regression coefficient vector $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}}_w$, the solution to

$$\boldsymbol{U_w}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}) = \sum_{i=1}^{n} \int_0^{\tau} \widehat{\pi}_i^{-1} \left\{ \boldsymbol{Z}_i(t) - \frac{\sum_{j=1}^{n} \widehat{\pi}_j^{-1} Y_j(t) \exp\{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)\} \boldsymbol{Z}_j(t)}{\sum_{j=1}^{n} \widehat{\pi}_j^{-1} Y_j(t) \exp\{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)\}} \right\} dN_i(t) = \mathbf{0}, \tag{3.1}$$

where $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, ..., \widehat{\pi}_n)$, $Y_i(t) = I(T_i \geq t)$, $N_i(t) = I(T_i \leq t)$, and I is the indicator function. Here τ is the maximum of the observed event times. The standard Cox regression estimator (Cox, 1975) which ignores double truncation, $\widehat{\boldsymbol{\beta}}_s$, is the solution to $\boldsymbol{U_w}(\boldsymbol{\beta}, \mathbf{1}) = \mathbf{0}$, where $\boldsymbol{U_w}(\boldsymbol{\beta}, \mathbf{1})$ is the score equation from the standard Cox model.

The caveat of the approaches which adjust for double truncation is that they require estimating the distribution of the truncation times, which is needed to obtain the estimator of the selection probabilities $\widehat{\boldsymbol{\pi}}$. Existing methods to estimate the truncation distribution require independence between the survival and truncation times. When this independence assumption is violated, the estimator of the truncation distribution will be biased, and therefore the estimator of the selection probabilities $\widehat{\boldsymbol{\pi}}$ will be biased. Because the methods in (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017) depend on $\widehat{\boldsymbol{\pi}}$, the resulting regression coefficient estimators will also be biased. The severity of this bias is demonstrated by the simulation studies in the next section.

When the survival times are conditionally independent of the truncation times given the covariate $\boldsymbol{Z}$, the likelihood of the observed survival times, conditional on the truncation times and covariates, is given by

$$L_n(\boldsymbol{\beta}, \Lambda) = \prod_{i=1}^{n} \frac{\lambda(T_i) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \exp\{-\Lambda(T_i) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)\}}{\alpha_i(\boldsymbol{\beta}, \lambda)},$$

where $\alpha_i(\boldsymbol{\beta}, \lambda) = \exp\{-\Lambda(L_i) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)\} - \exp\{-\Lambda(R_i) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)\}$ and $\Lambda(t) = \int_0^t \lambda(u) du$. That is, $\alpha_i(\boldsymbol{\beta}, \lambda) = P(L \leq T \leq R|\boldsymbol{Z}_i, L_i, R_i; \boldsymbol{\beta}, \lambda)$ is the probability of observing a random subject

from the target sample with covariate $\boldsymbol{Z}_i$ and truncation times L_i and R_i. Conditioning on the truncation times allows us to utilize the information in the covariates $\boldsymbol{Z}$ to relax the assumption of independence to an assumption of conditional independence. Furthermore, this conditioning completely avoids the need to estimate the distribution of the truncation times.

The log-likelihood function, $\log L_n(\boldsymbol{\beta}, \Lambda)$, can be expressed as

$$l_n(\boldsymbol{\beta}, \Lambda) = n^{-1} \sum_{i=1}^{n} \Big[\int_0^{\tau} \{\log \lambda(t) + \boldsymbol{\beta}' \boldsymbol{Z}_i - \Lambda(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)\} dN_i(t) - \log \alpha_i(\boldsymbol{\beta}, \lambda) \Big]. \tag{3.2}$$

Due to the difficulties of maximizing the log-likelihood (3.2) over all absolutely continuous cumulative hazard functions, we allow the estimator of λ to be discrete. Because the maximum likelihood estimation (MLE) of $\boldsymbol{\beta}$ and λ may be computationally intractable if directly solving the score equations for (3.2), we estimate $\boldsymbol{\beta}_0$ and λ using an EM algorithm. This has the advantage that its maximization step (M-step) only involves the complete-data likelihood. Based on the EM algorithm, we provide a convenient estimation approach to obtain estimators of the regression coefficients and baseline hazard function under left, right, or double truncation. This approach allows the survival and truncation times to be dependent through the covariate vector $\boldsymbol{Z}$. Furthermore, it does not require the estimation of the truncation time distribution. The estimation approach given here can easily be implemented using standard software for the Cox regression model.

3.2.1. Proposed EM Algorithm

Motivated by the approach in (Qin et al., 2011), who proposed EM algorithms for length-biased and right-censored data, Shen and Liu (Shen and Liu, 2017) proposed an EM algorithm to obtain pseudo MLEs of the regression coefficients from the Cox model under independent left and right truncation. They referred to their MLEs as pseudo because their proposed likelihood included the plug-in value of the estimator of the selection probabilities $\widehat{\boldsymbol{\pi}}$. However, as the authors point out, the estimated selection probabilities will be biased if the truncation times depend on the covariates **Z**. Hence, the resulting pseudo MLEs of the regression coefficients from the Cox model will also be biased.

We propose an EM algorithm for obtaining the MLE of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ based on (3.2). This allows us to relax the assumption of independence required by the methods in (Mandel et al., 2017; Rennert

and Xie, 2017; Shen and Liu, 2017) to an assumption of conditional independence by avoiding the estimation of the truncation distribution (and corresponding selection probabilities). Similar to the approaches of (Shen and Liu, 2017) and (Qin et al., 2011), we let $t_1 < ... < t_d$ denote the ordered, distinct failure times for $\{T_1, ..., T_n\}$. We develop the EM algorithm based on the discrete version of Λ, which we redefine as a step function only taking jumps at $t_1, ..., t_d$. Specifically, we set $\Lambda(t) = \sum_{t_j \leq t} \lambda_j$, were λ_j is the positive jump at time t_j for $j = 1, ..., n$.

Our observed data consists of $\boldsymbol{O} = \{\boldsymbol{O}_1, ..., \boldsymbol{O}_n\}$, where $\boldsymbol{O}_i \equiv (T_i, L_i, R_i, \boldsymbol{Z}_i)$ for $i = 1, ..., n$. Let $\boldsymbol{O}^* = \{T^*_{ir};\ i = 1, ..., n, r = 1, ..., m_i\}$ denote the truncated latent data, where T^*_{ir} is the missing survival time for a subject with truncation times (L_i, R_i) and covariate vector $\boldsymbol{Z}_i$ for $i = 1, ..., n$ and $r = 1, ..., m_i$. For notational convenience, we set $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ and define the density of T at time t_j, given $\boldsymbol{Z}_i$, as $f_i(t_j; \boldsymbol{\theta}) = \lambda_j \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \exp\{-\sum_{s=1}^{j} \lambda_s \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)\}$, where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$. Assuming the latent survival times T^*_{ir} take their values in $\{t_1, ..., t_d\}$, the complete data log-likelihood is given by

$$l_{full}(\boldsymbol{\theta}; \boldsymbol{O}, \boldsymbol{O}^*) = \sum_{j=1}^{d} \sum_{i=1}^{n} \left[I(T_i = t_j) + \sum_{r=1}^{m_i} I(T^*_{ir} = t_j) \right] \log f_i(t_j; \boldsymbol{\beta}, \boldsymbol{\lambda})$$

To estimate the parameter $\boldsymbol{\theta}$, the EM algorithm begins by choosing an initial value for $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$. In our setting, we can choose $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}_s, \boldsymbol{\lambda}_s)$, which are the estimates from the standard Cox model. For $k = 0, 1, 2, ...$, the expectation step (E-step) consists of calculating the expected value of the complete data log likelihood function $l_{full}(\boldsymbol{\theta}; \boldsymbol{O}, \boldsymbol{O}^*)$ with respect to the missing data T^*_{ir}, $i = 1, ..., n, r = 1, ..., m_i$, conditional on the observed data $(T_i, L_i, R_i, \boldsymbol{Z}_i)$, $i = 1, ..., n$, under the current estimate $\boldsymbol{\theta}^{(k)}$. That is, we compute:

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}} \left[l_{full}(\boldsymbol{\theta}; \boldsymbol{O}, \boldsymbol{O}^*) \middle| \boldsymbol{O} \right]$$

In the maximization step (M-step), we choose $\boldsymbol{\theta}^{(k+1)}$ to maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. That is, we set

$$\boldsymbol{\theta}^{(k+1)} = \text{arg max}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$$

The E- and M-steps are carried out again, but this time with $\boldsymbol{\theta}^{(k)}$ replaced by $\boldsymbol{\theta}^{(k+1)}$. The E- and M-steps are then alternated repeatedly until $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| < \epsilon$ for some prespecified error $\epsilon > 0$.

The EM algorithm described here is under the double truncation setting. If only left truncation is present, the algorithm is easily adjusted by setting $R_i = \infty$ for $i = 1, ..., n$. When only right truncation is present, we set $L_i = -\infty$ for $i = 1, ..., n$. Note that when only left truncation is present, the standard Cox regression estimator can account for dependent left truncation by adjusting the risk set at a given time point to include all individuals who are alive and in the study at that time (Klein and Moeschberger, 2003). We denote this estimator by $\widehat{\boldsymbol{\beta}}_{s,l}$. We show through simulations in the next section that when only left truncation is present, our proposed estimator and $\widehat{\boldsymbol{\beta}}_{s,l}$ yield nearly identical results.

3.2.2. E-step

At the k^{th} iteration, define $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)})$. Then,

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= E_{\boldsymbol{\theta}^{(k)}}\Big[\sum_{j=1}^{d}\sum_{i=1}^{n}\big[I(T_i = t_j) + \sum_{r=1}^{m_i} I(T_{ir}^* = t_j)\big]\log f_i(t_j;\boldsymbol{\theta})\Big|\boldsymbol{O}\Big] \\ &= \sum_{j=1}^{d}\sum_{i=1}^{n}\Big\{I(T_i = t_j) + \sum_{j=1}^{d}\sum_{i=1}^{n} E_{\boldsymbol{\theta}^{(k)}}\Big[\sum_{r=1}^{m_i} I(T_{ir}^* = t_j)\Big|\boldsymbol{O}_i\Big]\Big\}\log f_i(t_j;\boldsymbol{\theta}) \\ &= \sum_{j=1}^{d}\sum_{i=1}^{n}\big\{I(T_i = t_j) + \sum_{j=1}^{d}\sum_{i=1}^{n} E_{m_i}\big[m_i \times E_{\boldsymbol{\theta}^{(k)}}[I(T_{ir}^* = t_j)|\boldsymbol{O}_i]\big]\big\}\log f_i(t_j;\boldsymbol{\theta}), \end{aligned}$$

where

$$\begin{aligned} E_{\boldsymbol{\theta}^{(k)}}[I(T_{ir}^* = t_j)|\boldsymbol{O}_i] &= P_{\boldsymbol{\theta}^{(k)}}(T_{ir}^* = t_j|L_i, R_i, \boldsymbol{Z}_i) = P_{\boldsymbol{\theta}^{(k)}}(T = t_j|L_i, R_i, \boldsymbol{Z}_i, \{T < L\}\cup\{T > R\}) \\ &= \frac{P_{\boldsymbol{\theta}^{(k)}}(T = t_j, \{T < L\}\cup\{T > R\}|L_i, R_i, \boldsymbol{Z}_i)}{P_{\boldsymbol{\theta}^{(k)}}(\{T < L\}\cup\{T > R\}|L_i, R_i, \boldsymbol{Z}_i)} \\ &= \frac{f_i(t_j;\boldsymbol{\theta}^{(k)}) \times [I(t_j < L_i) + I(t_j > R_i)]}{1 - \alpha_i(\boldsymbol{\theta}^{(k)})}. \end{aligned}$$

Since m_i is the number of missing/truncated subjects with covariate values $\boldsymbol{Z}_i$ and truncation times L_i and R_i, m_i follows a geometric distribution with success rate $\alpha_i(\boldsymbol{\theta})$. Therefore when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, $E[m_i] = \frac{1-\alpha_i(\boldsymbol{\theta}^{(k)})}{\alpha_i(\boldsymbol{\theta}^{(k)})}$.

The complete data log likelihood is then given by

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{d}\sum_{i=1}^{n}\big\{I(T_i = t_j) + \frac{I(t_j < L_i) + I(t_j > R_i)}{\alpha_i(\boldsymbol{\theta}^{(k)})} f_i(t_j;\boldsymbol{\theta}^{(k)})\big\}\log f_i(t_j;\boldsymbol{\theta})$$

3.2.3. M-step

Let $w_{ij}^{(k)} = I(T_i = t_j) + \frac{I(t_j<L_i)+I(t_j>R_i)}{\alpha_i(\boldsymbol{\theta}^{(k)})} f_i(t_j;\boldsymbol{\theta}^{(k)})$. The complete data log likelihood can be written as

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{d}\sum_{i=1}^{n} w_{ij}^{(k)} \log f_i(t_j\boldsymbol{\theta}) = \sum_{j=1}^{d} w_{+j}^{(k)}\lambda_j + \sum_{i=1}^{n} w_{i+}^{(k)}\boldsymbol{\beta}'\boldsymbol{Z}_i + \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{s=1}^{j} w_{ij}^{(k)} \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\lambda_s,$$

where $w_{+j}^{(k)} = \sum_{i=1}^{n} w_{ij}^{(k)}$ and $w_{i+}^{(k)} = \sum_{j=1}^{d} w_{ij}^{(k)}$.

Treating $w_{ij}^{(k)}$ as constant, we set $\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})}{\partial\lambda_j} = 0$ to get a closed form solution to λ_j as a function of $\boldsymbol{\beta}$:

$$\lambda_j = \frac{w_{+j}^{(k)}}{\sum_{s=j}^{d}\sum_{i=1}^{n} w_{is}^{(k)} \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}, \quad j = 1, ..., d. \tag{3.3}$$

Differentiating $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\beta}$ yields

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} w_{i+}^{(k)}\boldsymbol{Z}_i + \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{s=1}^{j} w_{ij}^{(k)}\boldsymbol{Z}_i \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\lambda_s.$$

Setting the equation above equal to 0 and inserting the equation for λ_j yields

$$\sum_{i=1}^{n} w_{i+}^{(k)}\boldsymbol{Z}_i - \sum_{s=1}^{d} w_{+s}^{(k)}\left\{\frac{\sum_{i=1}^{n}\sum_{j=s}^{d} w_{ij}^{(k)}\boldsymbol{Z}_i \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}{\sum_{i=1}^{n}\sum_{j=s}^{d} w_{ij}^{(k)} \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}\right\} = \mathbf{0}. \tag{3.4}$$

The estimating equation (3.4) can be solved by specifying the "weights" option in the "coxph" function in R. First, a weight vector of length nd must be created: $\boldsymbol{w}_{nd}^{(k)} = (w_{11}^{(k)}, ..., w_{1d}^{(k)}, ..., w_{n1}^{(k)}, ..., w_{nd}^{(k)})$. The corresponding failure time data and covariate vectors are also created with length nd as follows: $\boldsymbol{T}_{nd} = (t_1, ..., t_d, ..., t_1, ..., t_d)$ and $\boldsymbol{Z}_{nd} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_1, ..., \boldsymbol{Z}_n, ..., \boldsymbol{Z}_n)$. Letting $\boldsymbol{\Delta}_{nd}$ be the identity vector of length nd, the solution to (3.4), which we denote by $\boldsymbol{\beta}^{(k+1)}$, can be obtained with the following command:

$$\text{coxph(Surv}(\boldsymbol{T}_{nd}, \boldsymbol{\Delta}_{nd}) \sim \boldsymbol{Z}_{nd}, \text{ weights = } \boldsymbol{w}_{nd}^{(k)}, \text{ subset = which}(\boldsymbol{w}_{nd}^{(k)} > 0)).$$

Plugging $\boldsymbol{\beta}^{(k+1)}$ into (3.3) yields an updated estimator for $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^{(k+1)}$. We then set $\boldsymbol{\theta}^{(k+1)} = (\boldsymbol{\beta}^{(k+1)}$,

$\boldsymbol{\lambda}^{(k+1)}$), and repeat the E- and M-steps. We continue to alternate between the E- and-M steps until $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| < \epsilon$, for some prespecified error $\epsilon > 0$. The MLE of the hazard ratio is then given by $\widehat{\boldsymbol{\beta}}_{em} = \boldsymbol{\beta}^{(k+1)}$. We denote the corresponding baseline hazard by $\widehat{\boldsymbol{\lambda}}_{em} = \boldsymbol{\lambda}^{(k+1)}$, and the cumulative baseline hazard function by $\widehat{\Lambda}_{em}(t) = \sum_{t_j \leq t} \lambda_j^{(k+1)}$.

The EM algorithm presented here falls into the general scheme of the ECM algorithm, and therefore its convergence to the local maximizer is guaranteed by the same conditions required for convergence of the ECM algorithm (Qin et al., 2011). The uniqueness of the resulting estimators are guaranteed by the regularity conditions in Appendix A. The R code implementing the EM algorithm described is provided in Appendix D.

3.2.4. Asymptotic Properties

In this section, we establish the strong consistency and asymptotic normality of the proposed EM estimators. Here we denote the proposed estimators by $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_{em}, \widehat{\Lambda}_{em})$, and denote the true regression coefficients and cumulative baseline hazard function $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \Lambda_0)$. The asymptotic properties of the proposed estimators refer to the situation when the total number of observed (non-truncated) subjects $n \to \infty$. The following theorems assume that the regularity assumptions in Appendix A hold.

Theorem 3.1: Under the regularity assumptions given in Appendix A, $\widehat{\boldsymbol{\theta}}$ is consistent: As $n \to \infty$, $\widehat{\boldsymbol{\beta}}_{em}$ converges to $\boldsymbol{\beta}_0$, and $\widehat{\Lambda}_{em}(t)$ converges to $\Lambda_0(t)$ almost surely and uniformly in t for $t \in [0, \tau]$.

The existence and uniqueness of the MLE can be proved based on the log-likelihood function

$$l_n(\boldsymbol{\beta}, \boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^{n} \Big[\int_0^{\tau} \boldsymbol{\beta}' \boldsymbol{Z}_i dN_i(t) + \sum_{s=1}^{d} \log \lambda_s \times I(T_i = t_s) - \sum_{t_s \leq t_i} \lambda_s \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \\ - \log \Big\{ \exp\Big(- \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \sum_{t_s < L_i} \lambda_s \Big) - \exp\Big(- \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \sum_{t_s \leq R_i} \lambda_s \Big) \Big\} \Big].$$

Theorem 3.1 can then be proved by applying the classical Kullback-Leibler information approach as in (Qin et al., 2011).

Theorem 3.2: Under the regularity assumptions given in Appendix A, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges weakly to a tight mean-zero Gaussian process.

Theorem 3.2 is proved using the Z-theorem for infinite dimensional equations (Vaart and Wellner, 2000). The proofs of Theorem 3.1 and Theorem 3.2 are outlined in Appendix B and C, respectively.

To obtain an estimate of the standard deviation of $\widehat{\boldsymbol{\beta}}_{em}$, we apply the simple bootstrap technique. In our setting, the bootstrap sample is obtained by drawing n independent vectors $(T_j^b, L_j^b, R_j^b, \boldsymbol{Z}_j^b)$, $j = 1, ..., n$, from the observed data vectors $(T_i, L_i, R_i, \boldsymbol{Z}_i)$, $i = 1, ..., n$, with replacement. These data vectors are then used to obtain an estimate of regression coefficients, denoted by $\widehat{\boldsymbol{\beta}}_{em}^{(b)}$. This process is repeated B times to obtain the B estimators $\widehat{\boldsymbol{\beta}}_{em}^{(1)}, ..., \widehat{\boldsymbol{\beta}}_{em}^{(B)}$. The estimate of the standard deviation of $\widehat{\boldsymbol{\beta}}_{em}$ is computed by taking the standard deviation of the $\widehat{\boldsymbol{\beta}}_{em}^{(b)}$, $b = 1, ..., B$. We denote this estimate by $\widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\beta}}_{em}}$. We show through simulation studies in the next section that the standard deviation of $\widehat{\boldsymbol{\beta}}_{em}$ is accurately estimated by $\widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\beta}}_{em}}$.

3.3. Simulations

In this section we examine the performance of the proposed estimator under dependent truncation. We compare our proposed estimator to the weighted estimator which adjust for double truncation but assumes independence between the survival and truncation times. We also compare the proposed estimator to the estimator from the standard Cox regression model. In all simulations, the survival times were generated from a proportional hazards model with hazard function $\lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$, and follow a Weibull distribution with scale parameter $\nu = 0.001$ and shape parameter $\kappa = 5$. We set $\beta_1 = \beta_2 = 1$, and generated the risk factors Z_1 and Z_2 from independent Unif[0,5] distributions. The truncation times were also simulated from Weibull distributions with scale parameter $\nu = 0.001$ and shape parameter $\kappa = 5$. The left truncation times were generated from a proportional hazards model with hazard function $\lambda_L(l)\exp(\beta_{L1} Z_1 + \beta_{L2} X)$, and the right truncation times were simulated from a proportional hazards model with hazard function $\lambda_R(r)\exp(\beta_{R1} Z_1 + \beta_{R2} Y)$. Here X and Y were generated from independent Unif[0,5] distributions, with $\beta_{L2} = \beta_{R2} = 1$. To adjust the proportion of missing data due to left and right truncation, the truncation times were multiplied by constants c_l and c_r, respectively. A higher value of c_l induced a higher proportion of missing data due to left truncation, while a lower value of c_r induced a higher proportion of missing data due to right truncation. Because the survival, left, and right truncation times are all functions of Z_1 for $\beta_1 \neq 0$, $\beta_{L1} \neq 0$, and $\beta_{R1} \neq 0$, they are dependent. However, the survival and truncation times are *conditionally independent* given Z_1. To adjust the degree of

dependence between T and L, and T and R, we varied the regression coefficients β_{L1} and β_{R1}, respectively.

We conducted 1000 simulation repetitions with sample sizes of n = 100 and 250. To obtain n observations after truncation, we simulated $N = \frac{n}{1-q}$ observations, where q is the proportion of truncated data. For each simulation, we estimated $\boldsymbol{\beta} = (\beta_1, \beta_2)$, using the proposed EM estimator $\widehat{\boldsymbol{\beta}}_{em} = (\widehat{\beta}_{em,1}, \widehat{\beta}_{em,2})$, the weighted Cox regression estimator $\widehat{\boldsymbol{\beta}}_{w} = (\widehat{\beta}_{w,1}, \widehat{\beta}_{w,2})$, and the standard Cox regression estimator $\widehat{\boldsymbol{\beta}}_{s} = (\widehat{\beta}_{s,1}, \widehat{\beta}_{s,2})$. Of the estimators which adjust for double truncation under the independence assumption, we only focus on $\widehat{\boldsymbol{\beta}}_{w}$ from (Rennert and Xie, 2017), as previous simulations (not shown here) have concluded that this estimator and that in (Mandel et al., 2017) are nearly identical, and both outperform the estimators in (Shen and Liu, 2017). For each estimator, we calculated the estimated bias, observed sample standard deviations (SD), estimated standard errors ($\widehat{\mathrm{SE}}$), and the average empirical coverage probability of the 95% confidence intervals (Cov). To compare the efficiency of the estimators which adjust for double truncation to the efficiency of the standard estimator, we calculated the relative mean-squared error (MSE) of $\widehat{\beta}_j$ to $\widehat{\beta}_{s,j}$, $j = 1, 2$. That is, we computed rMSE$(\widehat{\beta}_j) = \frac{\mathrm{MSE}(\widehat{\beta}_j)}{\mathrm{MSE}(\widehat{\beta}_{s,j})}$ for $\widehat{\beta}_j = \widehat{\beta}_{em,j}$ and $\widehat{\beta}_j = \widehat{\beta}_{w,j}$. We used 200 bootstrap resamples to estimate the standard error of $\widehat{\beta}_{em,j}$ and $\widehat{\beta}_{w,j}$, $j = 1, 2$.

Table 3.1 shows the results of the simulations described above. In the first model, we set $\beta_{L1} = -1$ to induce a negative dependence between the survival times and left truncation times, which resulted in a correlation of -0.35. In the second model, we set $\beta_{L1} = 0$ to induce independence between the survival and left truncation times. In the third model, we set $\beta_{L1} = 1$ to induce a positive dependence between the survival times and left truncation times, which resulted in a correlation of 0.35. Here c_l and c_r were chosen such that 25% of the survival times were left truncated and 25% of the survival times were right truncated, which resulted in $q \approx$ 0.50. The parameter β_{R1} was set to 1 in all models, which resulted in a correlation of 0.35 between the survival times and right truncation times.

In all models, the proposed EM estimators $\widehat{\beta}_{em,1}$ and $\widehat{\beta}_{em,2}$ had little bias, while the standard estimators $\widehat{\beta}_{s,1}$ and $\widehat{\beta}_{s,2}$ were biased. The weighted estimator $\widehat{\beta}_{w,1}$ was heavily biased in all models, while $\widehat{\beta}_{w,2}$ was biased in the first set of models ($\rho_{LT} = -0.35$). The observed sample standard deviations of the proposed estimators were accurately estimated by the bootstrap technique, and the coverage probabilities of the proposed estimators were all close to the nominal level of 0.95.

Table 3.1: Simulation results

Here ρ_{LT} is the correlation between the left truncation and survival time. The correlation between the right truncation and survival time is fixed at 0.35. The EM method produces the proposed estimator $\widehat{\beta}_{em}$, which adjusts for double truncation and dependence. The weighted method produces the estimator $\widehat{\beta}_w$, which adjusts for double truncation, but assumes independence between the survival and truncation times. The standard method assumes no truncation and produces the estimator $\widehat{\beta}_s$, the solution to the standard Cox score equation. Here SD is the empirical standard deviation of estimates across simulations, $\widehat{SE}$ is the average of the estimated standard errors. For an estimator $\widehat{\beta}$, rMSE = $\frac{MSE(\widehat{\beta})}{MSE(\widehat{\beta}_s)}$, where MSE is the mean-squared error. Cov is the coverage of 95% confidence intervals. Survival times generated from hazard function $\lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$, with $\beta_1 = \beta_2 = 1$. Survival times conditionally independent of left and right truncation times given Z_1.

ρ_{LT}	Method	n	Bias($\widehat{\beta}_1$)	SD($\widehat{\beta}_1$)	$\widehat{\text{SE}}(\widehat{\beta}_1)$	rMSE($\widehat{\beta}_1$)	Cov($\widehat{\beta}_1$)	Bias($\widehat{\beta}_2$)	SD($\widehat{\beta}_2$)	$\widehat{\text{SE}}(\widehat{\beta}_2)$	rMSE($\widehat{\beta}_2$)	Cov($\widehat{\beta}_2$)
-0.35	EM	100	0.01	0.13	0.14	1.05	0.96	-0.00	0.14	0.13	0.82	0.94
	weighted	100	0.10	0.15	0.16	1.94	0.95	0.06	0.15	0.15	1.11	0.95
	standard	100	-0.05	0.12	0.12	1.00	0.91	-0.10	0.12	0.11	1.00	0.82
	EM	250	-0.01	0.08	0.08	0.64	0.95	-0.02	0.08	0.08	0.38	0.94
	weighted	250	0.08	0.09	0.09	1.54	0.89	0.04	0.09	0.09	0.55	0.92
	standard	250	-0.07	0.07	0.07	1.00	0.83	-0.11	0.07	0.07	1.00	0.60
0.00	EM	100	-0.01	0.12	0.13	0.67	0.96	0.00	0.13	0.13	1.07	0.96
	weighted	100	-0.05	0.12	0.13	0.75	0.94	0.02	0.12	0.13	0.99	0.96
	standard	100	-0.10	0.11	0.12	1.00	0.84	-0.04	0.11	0.11	1.00	0.92
	EM	250	-0.01	0.08	0.08	0.38	0.95	-0.01	0.08	0.08	0.80	0.95
	weighted	250	-0.05	0.08	0.08	0.56	0.89	0.01	0.07	0.08	0.76	0.95
	standard	250	-0.10	0.07	0.07	1.00	0.68	-0.05	0.07	0.07	1.00	0.88
0.35	EM	100	0.03	0.16	0.16	0.72	0.95	0.01	0.14	0.14	1.15	0.96
	weighted	100	0.20	0.15	0.15	1.74	0.78	0.00	0.14	0.15	1.23	0.96
	standard	100	0.13	0.13	0.12	1.00	0.82	-0.05	0.12	0.12	1.00	0.92
	EM	250	0.00	0.09	0.09	0.45	0.94	-0.01	0.08	0.08	0.73	0.94
	weighted	250	0.18	0.09	0.09	2.22	0.45	-0.01	0.09	0.09	0.87	0.95
	standard	250	0.11	0.08	0.07	1.00	0.67	-0.06	0.07	0.07	1.00	0.84

The coverage probabilities of $\widehat{\beta}_{s,1}$ and $\widehat{\beta}_{s,2}$ were well below the nominal level, as were the coverage probabilities for $\widehat{\beta}_{w,1}$. Furthermore, the mean-squared errors of the proposed estimators were lower than those of the weighted and standard estimators in almost all settings, indicating that the proposed EM method is more efficient.

We further explored the bias and MSE of these estimators as a function of left and right truncation proportion (Figure 3.1). We set $\beta_{L1} = \beta_{R1} = 1$ and $n = 250$, which corresponded to the setting of the last model in Table 3.1, inducing a positive dependency between the survival times and both left and right truncation times. The proposed estimators had little bias, regardless of truncation proportion. Even under mild truncation, the weighted estimator $\widehat{\beta}_{w,1}$ of the regression coefficient corresponding to Z_1, which is correlated with the truncation times, was biased. This bias increased drastically as the proportion of right truncation increased. The bias was relatively small for both the proposed and weighted estimator of the regression coefficient corresponding to Z_2, which is

uncorrelated with the truncation times. Both standard estimators $\widehat{\beta}_{s,1}$ and $\widehat{\beta}_{s,2}$ were heavily biased in this setting. The MSE of $\widehat{\beta}_{em,j}$ was significantly lower than the MSE of $\widehat{\beta}_{w,j}$ for $j = 1, 2$. Furthermore, in most cases the MSE of $\widehat{\beta}_{em,j}$ was lower than the MSE of the standard estimator $\widehat{\beta}_{s,j}$, i.e. rMSE($\widehat{\beta}_{em,j}$) < 1 for $j = 1, 2$.

In Figure 3.2, we compared the bias and MSE of these estimators under varying truncation proportions, when the assumption of independence holds (i.e. $\beta_{L1} = \beta_{R1} = 0$). The proposed EM estimators $\widehat{\beta}_{em,j}$ and weighted estimators $\widehat{\beta}_{w,j}$ had little bias, while the standard estimators $\widehat{\beta}_{s,j}$ were biased for $j = 1, 2$. We also compared the rMSE of $\widehat{\beta}_{em,j}$ and $\widehat{\beta}_{w,j}$ to $\widehat{\beta}_{s,j}$, $j = 1, 2$. As indicated by the bottom row of Figure 3.2, the proposed EM estimators had similar efficiency to the weighted estimators when the independence assumption holds. When the proportion of missing data due to left and right truncation were approximately equal, the standard estimator was more efficient than the proposed EM estimator and the weighted estimator. This is because the bias due to left truncation canceled out with the bias due to right truncation when these proportions were equal, which yielded a lower MSE.

The standard Cox regression model can accommodate left truncation when the left truncation time is conditionally independent of the survival times given the observed risk factors. We compare the estimator from this model to our proposed estimator under dependent left truncation only. To adjust the correlation between the left truncation times and survival times, we varied the parameter β_{L1} between -1 and 1. In this setting, a value of $\beta_{L1} = 0$, which yields a correlation of 0, indicates independence between the left truncation times and survival times. We denote the standard regression coefficient estimator which adjusts for dependent left truncation as $\widehat{\boldsymbol{\beta}}_{s,l} = (\widehat{\beta}_{s,l,1}, \widehat{\beta}_{s,l,2})$. As shown in Figure 3.3, $\widehat{\beta}_{em,j}$ and $\widehat{\beta}_{s,l,j}$ had little bias, while the weighted estimators $\widehat{\beta}_{w,j}$ were biased for $j = 1, 2$. As indicated by the bottom row of Figure 3.3, the proposed EM estimators had similar efficiency to the standard estimators which accounted for dependent left truncation, and both estimators were more efficient than the weighted estimators.

3.4. Application to Alzheimer's Disease

We illustrate our method by considering an autopsy-confirmed AD study conducted by the Center for Neurodegenerative Disease Research at the University of Pennsylvania. The target population for the research purposes of this study consists of all subjects with AD symptom onset before 2012

that met the study criteria and therefore would have been eligible to enter the center. Our observed sample contains all subjects who entered the center between 1995 and 2012, and had an autopsy performed before July 1, 2012. Thus one criterion for a subject to be included in our sample is that they did not succumb to AD before they entered the study, yielding left truncated data. In addition, our sample only contains subjects who had an autopsy-confirmed diagnosis of AD, and therefore we have no knowledge of subjects who live past the end of the study. Thus our data is also right truncated. Our data consists of n=91 subjects, all of whom have event times. The event time of interest is the survival time (T) from AD symptom onset. The left truncation time (L) is the time between the onset of AD symptoms and entry into the study (i.e. initial clinic visit). The right truncation time (R) is the time between the onset of AD symptoms and the end of the study, which is taken to be July 1, 2012. Due to double truncation, we only observe subjects with $L \leq T \leq R$.

We are interested in assessing the effect of occupation on survival in AD. Occupation is often used as a proxy for cognitive reserve (CR), which hypothesizes that individuals develop cognitive strategies and neural connections throughout their life times through experience such as occupation, education, and other forms of mental engagement (Valenzuela and Sachdev, 2007). A common hypothesis in the literature is that CR has protective role in the brain and modulates the effects of AD because of compensatory strategies obtained from a higher level of professional performance and therefore lengthens survival during the course of the disease (Ientile et al., 2013; Sanchez et al., 2011).

However some studies have shown a higher mortality rate in AD individuals with higher occupational attainment (Stern et al., 1999, 1995). This supports an alternative theory of CR; individuals with higher CR tolerate more pathology which delays the onset of the disease. Because higher age of AD symptom onset is associated with an increased risk of mortality, this would support the hypothesis that those with higher CR would have an increased risk of mortality. There are two caveats to the studies described above. The first is that these studies consisted of populations with clinically diagnosed AD subjects, which can be unreliable. The second caveat is that the statistical analyses were subject to confounding, since age of AD symptom onset was not recorded nor adjusted for.

Here we are interested in obtaining improved estimates of the effect of occupation on survival from an autopsy-confirmed cohort of individuals with AD who have a known age of disease symptom on-

set. We use the highest occupational attainment for a given subject as a proxy for their CR. Primary occupation was classified and ranked based on the US census categories. In the following analyses, subjects who were classified as manager, business/government, and professional/technical workers were labeled as having *high occupational attainment* in our study. Subjects classified as unskilled/semiskilled, skilled trade or craft, and clerical/office workers were classified as having *low occupational attainment*. This classification is consistent with previous studies (Massimo et al., 2015, 2018; Stern et al., 1995). Age at AD symptom onset was estimated based on a family report at first contact with the individual.

We first check the assumption of independence between the observed survival and truncation times using the conditional Kendall's tau proposed by Martin and Betensky (Martin and Betensky, 2005). The resulting p-value is 0.038, and therefore we reject this independence assumption. The corresponding Kendall's tau statistic is $\tau_K = (0.20, 0.16)$, indicating positive dependence between the survival times and truncation times. The positive dependence between the left truncation times and survival times is clinically plausible because doctors often attribution the symptoms of early onset AD (onset of AD before 65 years of age) to other causes such as depression and stress, hence delaying the study entry time. Since younger age at onset is also associated with higher survival, this induces a positive dependence between the left truncation times and survival times.

Due to the dependence between the survival and truncation times, we apply the proposed method to estimate the effect of occupation on survival, adjusting for age at AD symptom onset and sex. Table 3.2 displays the results from the Cox regression model using the proposed EM estimators, weighted estimators, and the standard estimators. Using the proposed method, the estimated log hazard ratio for age at AD symptom onset is 0.029 (p-value = 0.016), indicating that AD individuals who have symptom onset one year later are roughly 3% more likely to die than subjects who have symptom onset a year earlier ($e^{0.029} = 1.03$). The estimated effect of female is -0.636 (p-value = 0.023), indicating that males are almost twice as likely to die than females ($e^{0.636} = 1.89$). These effects are nearly doubled using the weighted method which assumes independence, however the effects are not statistically significant (p-values = 0.117 and 0.088, respectively).

High occupational attainment is associated with increased survival in all models. Under the proposed method, the effect of high occupational attainment on survival is -0.673 (p-value = 0.009), indicating that those with a low occupational attainment are approximately twice as likely to die

than those with a high occupational attainment ($e^{0.673} = 1.96$). This effect is attenuated under the weighted and standard methods, and neither method yielded statistically significant estimates (p-values are 0.186 and 0.158, respectively).

Table 3.2: Application: Occupational attainment on survival in AD.

	EM		Weighted		Unweighted	
Predictor	$\widehat{\beta}_{em}(\widehat{\mathrm{SE}})$	p-value	$\widehat{\beta}_{w}(\widehat{\mathrm{SE}})$	p-value	$\widehat{\beta}_{s}(\widehat{\mathrm{SE}})$	p-value
Age onset	0.029 (0.012)	0.016	0.047 (0.030)	0.117	0.035 (0.013)	0.013
Female	-0.636 (0.280)	0.023	-1.026 (0.602)	0.088	-0.532 (0.223)	0.017
High occupation	-0.673 (0.257)	0.009	-0.464 (0.351)	0.186	-0.487 (0.345)	0.158

3.5. Discussion

We proposed a novel method which relaxes the independence assumption between the observed survival and truncation times in the Cox model under left, right, or double truncation to an assumption of conditional independence between the observed survival and truncation times. We obtained consistent and asymptotically normal estimators of the regression coefficients and baseline hazard function by maximizing the conditional likelihood of the observed survival times using an EM algorithm. The simulation studies confirmed that the proposed estimators had little bias in small samples, while the naïve estimators from the Cox models which ignore truncation or assume independence were biased. The existing methods which adjust for truncation but assume independence resulted in heavily biased estimators of the regression coefficients for risk factors of survival that were also correlated with the truncation times. Furthermore, the proposed estimators were more efficient than the naïve estimators in most of the simulation settings.

We applied our proposed method to an autopsy-confirmed sample of individuals with Alzheimer's disease (AD). AD is a major neurodegenerative disease which currently affects 5.3 million people in the United States according to the Alzheimer's Association. In 2017 alone, AD and other dementias will have cost the nation an estimated $259 billion. Autopsy-confirmation is needed for a definitive diagnosis of AD, and a definitive diagnosis is necessary to accurately estimate the effect of potential risk factors associated with a given neurodegenerative disease. However, autopsy-confirmed samples of neurodegenerative diseases are subject to an inherent selection bias due to double truncation. Existing methods which adjust the Cox model in the presence of double truncation assume that the observed survival and truncation times are independent. This assumption may not be reasonable for studies of neurodegenerative diseases. In our data example, this independence

assumption was rejected. Therefore, previous methods are not appropriate for our setting.

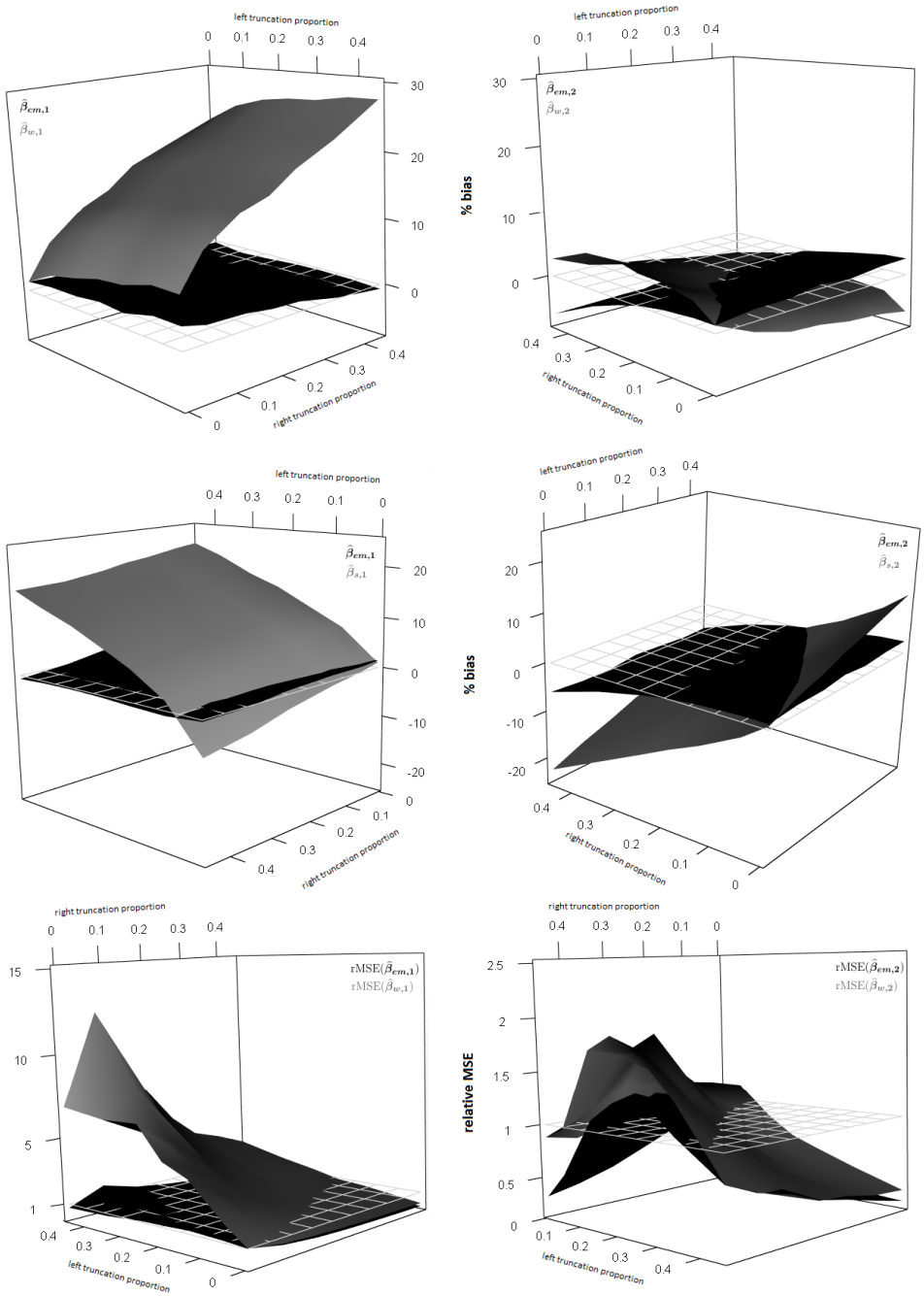
Given the severity of Alzheimer's disease on patients, their caregivers, and society, accurate estimation of the effects of risk factors on survival is crucial. One such factor, cognitive reserve (CR), is hypothesized to lengthen survival during the course of the disease. Using occupation as a proxy for CR, we estimated the effect of CR on survival in an autopsy-confirmed AD sample. Using our proposed method to adjust for both left and right truncation and dependence between the survival and truncation times, we found that a low occupational attainment was associated with shortened survival. Compared to existing methods, the estimated hazard ratios for occupation on survival were larger under our proposed method. This is consistent with many studies concluding that an individual's occupation may provide a protective effect and lengthen survival in AD. These findings suggest the importance of incorporating occupation in treatment trials and prognostic considerations in individuals with AD.

A limitation of our proposed method is that it in its current form, it cannot properly handle time-varying covariates measured after study entry, such as cognitive test scores. This is a consequence of the estimation procedure, which uses an expectation-maximization algorithm to estimate the latent survival times conditional on the observed truncation times and risk factors. This leads to predicting survival times based on risk factors measured after death for those missing subjects whose survival time is less than their left truncation time, which may yield biased regression coefficient estimators.

The proposed method has useful implications for observational studies. Double truncation has been shown to be present in a variety of studies, such as studies of clinically diagnosed Parkinson's disease (Mandel et al., 2017), childhood cancer (Moreira and Una-Alvarez, 2010), astronomy data Efron and Petrosian, 1999, and studies based on registry data (Bilker and Wang, 1996; Shen and Liu, 2017). In fact, any data pulled from a disease registry will be subject to inherent right truncation, since data is only recorded for subjects who have the disease and are entered in the registry by the time the data is extracted (Bilker and Wang, 1996). In certain cases, the data will also be subject to left truncation (Bilker and Wang, 1996; Shen and Liu, 2017). In a similar fashion, studies which only include data from individuals whose event times fall within the time course of the study are subject to double truncation (Moreira and Una-Alvarez, 2010). Therefore careful consideration of the study design must be taken into account when fitting the Cox proportional hazards model.

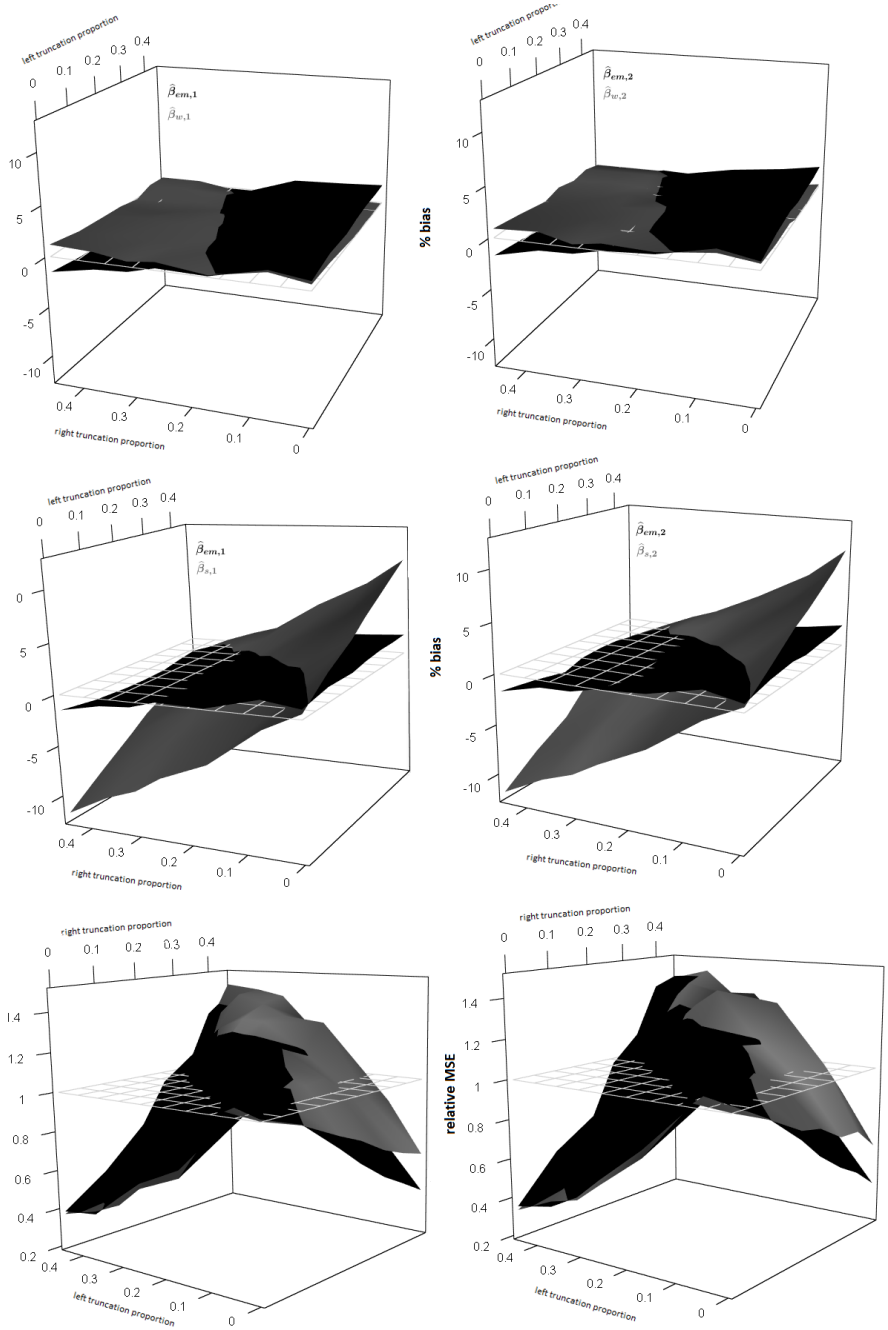
Furthermore, the assumption of independence should always be tested, given the high sensitivity of existing methods to this assumption. For example, a quick application of a Kendall's conditional Tau test (Martin and Betensky, 2005) revealed this independence assumption is violated in the AIDS data used in Shen and Liu, 2017. We therefore recommend using the proposed estimators in most practical settings, since they have little bias, and in most situations, have a lower mean-squared error compared to existing estimators under left, right, or double truncation, under a wide range of dependence structures.

Figure 3.1: Comparing bias and MSE (mean-squared error) of estimators across different left and right truncation proportions, under *dependent* survival and truncation times.



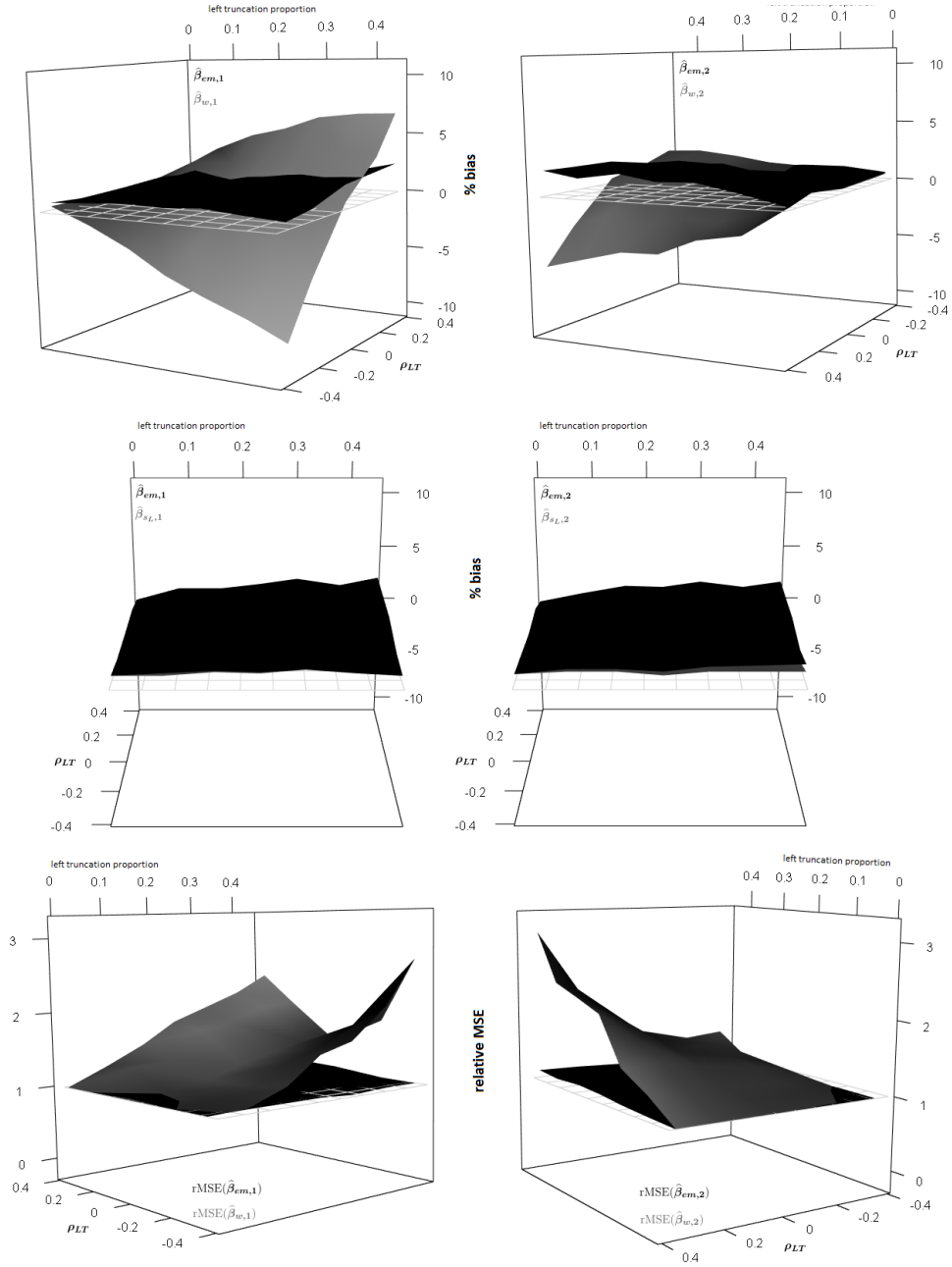
Survival times generated from proportional hazards model with hazard function $\lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$*, with* $\beta_1 = \beta_2 = 1$*. Survival times conditionally independent of left and right truncation times given* Z_1*. For* $j = 1$ *(left column) and* $j = 2$ *(right column): Top row compares bias of proposed EM estimator* $\widehat{\beta}_{em,j}$ ***(black)*** *to weighted estimator* $\widehat{\beta}_{w,j}$ ***(gray)****, which does not account for dependent truncation. Middle row compares bias of* $\widehat{\beta}_{em,j}$ ***(black)*** *to the standard estimator* $\widehat{\beta}_{s,j}$ ***(gray)****, which ignores truncation completely. Bottom row compares rMSE(*$\widehat{\beta}_{em,j}$*) (**black**) to rMSE(*$\widehat{\beta}_{w,j}$*) (**gray**). Here rMSE(*$\widehat{\beta}$*) =* $\frac{MSE(\widehat{\beta}_j)}{MSE(\widehat{\beta}_{s,j})}$ *is the relative MSE of the estimator* $\widehat{\beta}_j$ *to the standard estimator* $\widehat{\beta}_{s,j}$*.*

Figure 3.2: Comparing bias and MSE (mean-squared error) of estimators across different left and right truncation proportions, under *independent* survival and truncation times.



Survival times generated from proportional hazards model with hazard function $\lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$*, with* $\beta_1 = \beta_2 = 1$*. For* $j = 1$ *(left column) and* $j = 2$ *(right column): Top row compares bias of proposed EM estimator* $\widehat{\beta}_{em,j}$ ***(black)*** *to weighted estimator* $\widehat{\beta}_{w,j}$ ***(gray)****, which does not account for dependent truncation. Middle row compares bias of* $\widehat{\beta}_{em,j}$ ***(black)*** *to the standard estimator* $\widehat{\beta}_{s,j}$ ***(gray)****, which ignores truncation completely. Bottom row compares rMSE(*$\widehat{\beta}_{em,j}$*)* ***(black)*** *to rMSE(*$\widehat{\beta}_{w,j}$*)* ***(gray)****. Here rMSE(*$\widehat{\beta}$*) =* $\frac{\text{MSE}(\widehat{\beta}_j)}{\text{MSE}(\widehat{\beta}_{s,j})}$ *is the relative MSE of the estimator* $\widehat{\beta}_j$ *to the standard estimator* $\widehat{\beta}_{s,j}$*.*

Figure 3.3: Comparing bias and MSE (mean-squared error) of estimators under *dependent left truncation*.



*Here ρ_{LT} is the correlation between the left truncation and survival time. Survival times generated from proportional hazards model with hazard function $\lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$, with $\beta_1 = \beta_2 = 1$. For $j = 1$ (left column) and $j = 2$ (right column): Top row compares bias of proposed EM estimator $\widehat{\beta}_{em,j}$ (**black**) to weighted estimator $\widehat{\beta}_{w,j}$ (**gray**), which does not account for dependent truncation. Middle row compares bias of $\widehat{\beta}_{em,j}$ (**black**) to the standard estimator under left truncation $\widehat{\beta}_{s_L,j}$ (**gray**), which accounts for dependent left truncation. Bottom row compares rMSE($\widehat{\beta}_{em,j}$) (**black**) to rMSE($\widehat{\beta}_{w,j}$) (**gray**). Here rMSE($\widehat{\beta}$) = $\frac{MSE(\widehat{\beta}_j)}{MSE(\widehat{\beta}_{s_L,j})}$ is the relative MSE of the estimator $\widehat{\beta}_j$ to the standard estimator under left truncation $\widehat{\beta}_{s_L,j}$.*

CHAPTER 4

BIAS IN THE SURVIVAL DISTRIBUTION FUNCTION ESTIMATOR UNDER DOUBLE TRUNCATION: A CASE STUDY OF NEURODEGENERATIVE DISEASES

4.1. Introduction

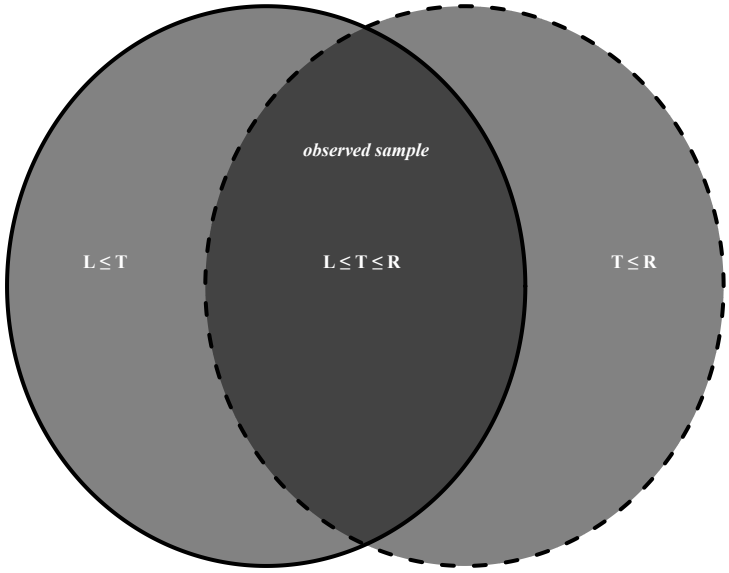
Neurodegenerative diseases, such as Alzheimer's disease (AD) and frontotemporal lobar degeneration (FTLD), require an autopsy for a definitive diagnosis (Grossman and Irwin, 2016). Without an autopsy-confirmed diagnosis, it is uncertain which disease a given individual may have. Hence this individual cannot be included in an autopsy-confirmed study sample pertaining to a particular disease. Therefore when the event of interest is death, studies which include only autopsy-confirmed subjects result in pure right truncation, since individuals who have the disease of interest and live past the end of study date do not receive a pathological diagnosis. Since these individuals cannot be included in the autopsy-confirmed study sample, they are treated as unobserved. Furthermore, studies that recruit individuals after the onset of the disease has occurred may result in left truncation, since individuals who succumb to the disease before they enter the study are unobserved. This simultaneous presence of left and right truncation, also known as double truncation, is therefore inherent in autopsy-confirmed studies of neurodegenerative disease.

Double truncation occurs in these studies as follows: Subjects are only observed if their time of death, t_{death}, occurs after the time of study entry, t_{entry}, and before the study end time t_{end}. In other words, only subjects with $t_{entry} \leq t_{death} \leq t_{end}$ are observed. The survival time T in individuals with neurodegenerative diseases is typically measured as the time from disease symptom onset to death. That is, $T = t_{death} - t_{onset}$, where t_{onset} is defined as the time in which disease symptom onset occurs. We therefore define the left truncation time as the time from disease symptom onset to study entry, $L = t_{entry} - t_{onset}$, and the right truncation time as the time from disease symptom onset to study end, $R = t_{end} - t_{onset}$. The truncation scheme $t_{entry} \leq t_{death} \leq t_{end}$ is therefore equivalent to $L \leq T \leq R$.

This truncation scheme is illustrated in Figure 4.1. Unlike a censored individual who provides partial information about their survival time, a truncated individual is completely unobserved and provides

no information to the investigator, resulting in a biased sampling scheme. Right truncation in this setting yields an observed sample that is biased towards smaller survival times, since individuals with longer survival times are more likely to live past the end of the study. The left truncation simultaneously leads to an observed sample that is biased towards larger survival times, since individuals with shorter survival times are more likely to succumb to the disease before they enter the study. Therefore any estimation procedure of the survival time distribution which does not account for the double truncation will be biased. In this paper, we focus on autopsy-confirmed studies of neurodegenerative diseases, but note that double truncation can be present in other studies (Bilker and Wang, 1996).

Figure 4.1: Schematic depiction of doubly truncated neurodegenerative disease data



Here L, T, and R denote the time from disease symptom onset to study entry, death, and the end of study, respectively. The solid circle (left) consists of all subjects who entered the study and are therefore not left truncated. The light grey region of the solid circle is right truncated, and consists of all subjects who entered but lived past the end of the study, i.e. $\{L \leq T\} \cap \{T > R\}$. The dotted circle (right) consists of all subjects who had an autopsy performed by the end of the study and are therefore not right truncated. The light grey region of the dotted circle is left truncated, and consists of all subjects who never entered the study but died before the end of study date, i.e. $\{T < L\} \cap \{T \leq R\}$. The observed sample is represented by the intersection of the two circles (dark grey region), and consists of all subjects who entered the study and had an autopsy performed $\{L \leq T \leq R\}$.

The bias introduced in autopsy-confirmed survival studies is briefly discussed in Rennert and Xie (2017) in the context of Cox regression models. One of the goals of our paper is to further em-

phasize and explore this important issue by examining the bias introduced in autopsy-confirmed survival studies in the context of survival distribution estimation, thus avoiding any assumptions about the survival time. Survival distribution estimation is useful in time event analysis as it serves as the first step of evaluating the disease risk. It is a useful exploratory tool before any regression modeling. It is particularly suited for graphical display which is an essential part of disease risk modeling.

There are a few papers devoted to the estimation of the survival time distribution in the presence of double truncation. Bilker and Wang (1996) were one of the first to motivate the problem of double truncation by noticing that it was present in certain retrospective studies of survival from HIV infection to AIDS. Motivated by doubly truncated quasar data, Efron and Petrosian (1999) introduced a nonparametric maximum likelihood estimator (NPMLE) of the survival time distribution under double truncation. Shen (2010) established the asymptotic properties of the NPMLE, and introduced a nonparametric estimator of the truncation distribution. Under the assumption that the joint distribution function of the truncation times comes from a parametric family, Shen (2010) and Moreira and de Ũna-Álvarez (2010) introduced a semiparametric maximum likelihood estimator (SPMLE) for the survival time distribution function under double truncation. The NPMLE and SPMLE both assume independence between survival and truncation times. A version of a conditional Kendall's Tau was introduced by Martin and Betensky (2005) to test for dependence between survival and both left and right truncation times.

There are several new contributions of this paper, which we summarize below. Our first contribution is to inform the reader about the inherent double truncation present in autopsy-confirmed survival studies of neurodegenerative diseases and highlight the importance of accounting for it. Our second contribution is to compare the SPMLE and NPMLE to the naïve empirical distribution function which ignores double truncation. This fills a void in the literature on survival distribution estimation, which lacks formal comparisons of approaches that adjust for double truncation to approaches that ignore it. Our third contribution is examining the robustness of the SPMLE and NPMLE to violations of independence between the survival and truncation times, which have not been previously studied. Our fourth contribution is that we discover through simulations that the SPMLE is robust to model misspecification when a gamma distribution with two unknown parameters is assumed for the truncation times. This discovery is contrary to previous literature which has concluded that

the SPMLE can be heavily biased under misspecification of the truncation distribution (Moreira and Una-Alvarez, 2010; Shen, 2010b). Our fifth contribution is to demonstrate how to appropriately estimate and compare survival distribution functions in the context of autopsy-confirmed survival studies of AD and FTLD.

In Section 4.2, we introduce notation and the SPMLE and NPMLE of the survival distribution function, as well as formal tests to compare distribution functions in the presence of double truncation. The simulations to evaluate and compare the performance of these estimators are presented in Section 4.3. In Section 4.4 we apply the SPMLE and NPMLE to the neurodegenerative disease study to estimate and compare the survival curves for subjects diagnosed with AD or FTLD. Concluding remarks and limitations of these methods are discussed in Section 4.5.

4.2. Existing methods to adjust for double truncation

We state the problem in statistical terms as follows. Let T denote the survival time of interest (e.g. survival time from disease symptom onset), L denote the left truncation time (e.g. time from disease symptom onset to entry into the study), and R denote the right truncation time (e.g. time from disease symptom onset to the end of study date). Let N denote the size of the target sample – the sample that would have been observed had there been no truncation present in the study. We denote the observed data as (T_i, L_i, R_i) for $i = 1, ..., n$. Due to double truncation, we only observe (T_i, L_i, R_i) for $n \leq N$ individuals who live long enough to enter the study (i.e. $T \geq L$) and do not live past the end of the study (i.e. $T \leq R$). Here we have denoted the population random variables from the target population without subscripts, and the sampling random variables from the observed sample with subscripts.

We are interested in estimating the cumulative distribution function F of T, where $F(t) = P(T \leq t)$ for a given time t. The survival distribution function is given by $S(t) = 1 - F(t)$. We note that right censoring is not present in autopsy-confirmed studies of neurodegenerative diseases. This is because individuals who live past the end of the study are undiagnosed (since an autopsy is never performed) and not included in the study sample. Therefore no information is available on the survival time of these individuals. With no censoring, the standard estimator of the cumulative distribution function of the survival times is just the empirical cumulative distribution function (eCDF) $\hat{F}_{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[T_i \leq t]}$ for a given time t, where I is the indicator function. We show through

simulations in the next section that this estimator, which does not take into account that the data are doubly truncated, is biased. Throughout the paper, we refer to the eCDF as the naïve estimator and denote it by $\hat{F}_{emp}$. We note that the eCDF is equivalent to the Kaplan-Meier estimator when no censoring is present.

The methods to estimate F, described below, assume that the survival times are independent of the left and right truncation times, that no censoring is present, and that (L_i, T_i, R_i) are independent and identically distributed for $i = 1, ..., n$. For any cumulative distribution function Q, we define the left endpoint of its support by $a_Q = inf\{x : Q(x) > 0\}$ and the right endpoint of its support by $b_Q = inf\{x : Q(x) = 1\}$. Let K denote the joint cumulative distribution function of the left and right truncation times. Let $H_L(l) = K(l, \infty)$ and $H_R(r) = K(\infty, r)$ denote the marginal cumulative distribution functions of L and R, respectively. The methods described below assume that $a_{H_L} < a_F \leq a_{H_R}$ and $b_{H_L} \leq b_F < b_{H_R}$. These conditions are needed for identifiability of the cumulative distribution function estimators (Woodroofe, 1985).

The two existing methods for estimating the cumulative distribution function under double truncation are the SPMLE and the NPMLE. Both make no assumptions about the distribution of the survival times, but the SPMLE assumes that the truncation times L and R have a joint cumulative distribution function, $K(\cdot, \cdot; \theta)$, that depends on a parameter θ. As described in (Shen, 2010b) and (Moreira and Una-Alvarez, 2010), an estimate $\hat{\theta}$ of θ can be obtained and then used to compute $W_{\hat{\theta}}(T_i)$, the estimated likelihood of observing a subject with survival time T_i in the sampled population relative to the target population. Specifically, $W_{\hat{\theta}}(T_i) = P_{\hat{\theta}}(L \leq T \leq R | T = T_i)$, the inverse of the estimated probability (under parametric assumptions) of observing a subject in the study sample with survival time $T = T_i$.

The SPMLE is then a weighted sum of the elements $I(T_i \leq t)$ of the eCDF and is given by

$$\hat{F}_{SP}(t) = \frac{1}{n} \sum_{i=1}^{n} W_{\hat{\theta}}(T_i) \times I(T_i \leq t). \tag{4.1}$$

Under the regularity conditions given in Shen (2010), namely that $K(l, r; \theta)$ is continuous in (l, r) for each θ in a compact set Θ, and $K(l, r; \theta)$ is continuously differentiable in θ for each fixed (l, r), we have that $\sqrt{n}\big(\hat{F}_{SP}(t) - F(t)\big) \rightarrow N\big(0, \sigma^2(t)\big)$. This result also rests on the assumption that the truncation distribution is correctly specified. Details can be found in (Moreira and Una-Alvarez,

2010; Shen, 2010b). The distributional assumptions for the truncation times can be checked using the test statistics introduced in (Moreira, Ivarez, and Van Keilegom, 2014).

The NPMLE makes no distributional assumptions about the truncation times. Similar to the SPMLE, the NPMLE is weighted by $\widehat{W}(T_i)$, the nonparametric estimate of the likelihood of observing a subject with survival time T_i in the sampled population relative to the target population. Here $\widehat{W}(T_i) = \widehat{P}(L \leq T \leq R|T = T_i)^{-1}$, the inverse of the estimated probability (under no parametric assumptions) of observing a subject in the study sample with survival time $T = T_i$. The NPMLE is then given by

$$\hat{F}_{NP}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}(T_i) \times I(T_i \leq t). \tag{4.2}$$

Details of this estimation procedure are given in (Shen, 2010a). We note that there is no closed form variance estimator for $\hat{F}_{NP}(t)$. We therefore apply the simple bootstrap technique to estimate the variance of $\hat{F}_{NP}(t)$.

Often we would like to test whether two survival distributions are equal. Under double truncation, this can be done using the semiparametric extension of the Mann-Whitney test introduced in Bilker and Wang, 1996. This estimator also makes use of the parametric distribution of the truncation times. Let $(L_{1_i}, T_{1_i}, R_{1_i})$, $1 \leq i \leq n_1$ be the observed data from group 1 and $(L_{2_j}, T_{2_j}, R_{2_j})$, $1 \leq j \leq n_2$ be the observed data from group 2. Here it is assumed that (L_1, R_1) have a parametric joint cumulative distribution function K_θ and (L_2, R_2) have a parametric joint cumulative distribution function H_γ. The two-sample U-statistic is of the form

$$U(\hat{\theta}, \hat{\gamma}) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} sign(T_{1_i} - T_{2_j}) \times W_{1\hat{\theta}}(T_{1_i}) \times W_{2\hat{\gamma}}(T_{2_j}).$$

Similar to the definition of $W_{\widehat{\theta}}(T_i)$, $W_{1\hat{\theta}}(T_{1_i})$ is the inverse of the estimated probability of observing a subject from group 1 in our study sample with survival time T_{1_i}, and $W_{2\hat{\gamma}}(T_{2_j})$ is the inverse of the estimated probability of observing a subject from group 2 in our study sample with survival time T_{2_j}. See (Bilker and Wang, 1996) for more details.

Bilker and Wang's U-statistic tests whether two survival distributions are equal across all time points. To test whether the probability of survival between two independent groups are equal at a single

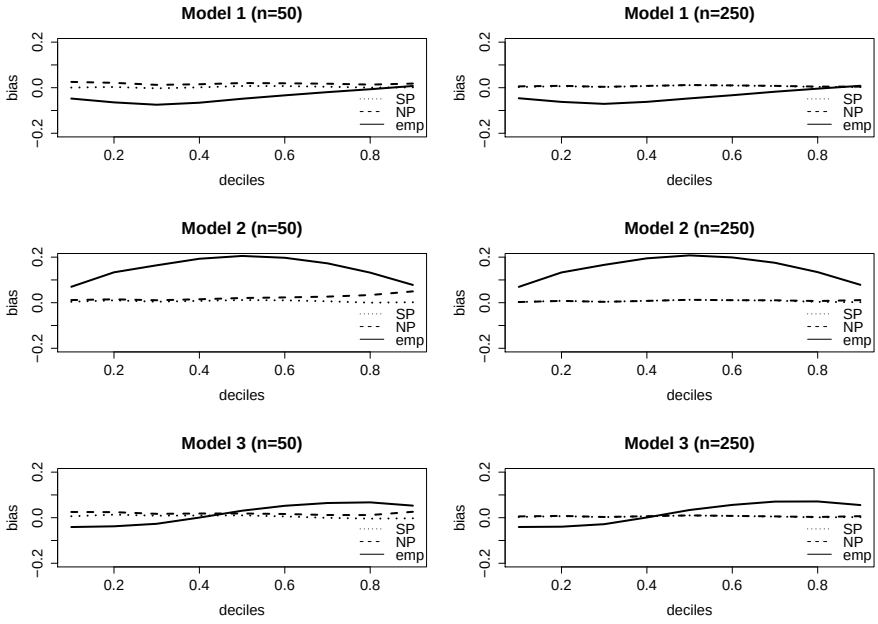
time point t, we introduce the Wald statistic $W_t = \frac{[\hat{F}_1(t)-\hat{F}_2(t)]^2}{\hat{\sigma}_1^2(t)/n_1+\hat{\sigma}_2^2(t)/n_2}$, where $W_t \sim \chi_1^2$. Here $\hat{F}_j(t)$ and $\hat{\sigma}_j^2(t)$ are any estimates of the cumulative distribution function and standard error at time t for $j = 1, 2$.

4.3. Simulation study

We conducted a simulation study to further investigate the impact of ignoring double truncation in autopsy-confirmed survival studies of neurodegenerative disease and to assess the performance of the SPMLE and NPMLE under different truncation schemes. Specifically, we compared the SPMLE ($\hat{F}_{SP}$) and NPMLE ($\hat{F}_{NP}$) to the eCDF ($\hat{F}_{emp}$) on bias $(\hat{F} - F_0)$, where F_0 is the true distribution function, observed sample standard deviations (SD), estimated standard errors ($\widehat{\mathrm{SE}}$), mean squared errors (MSE), and the average empirical coverage probability of the 95% confidence intervals (Cov). We also compared the bias and observed sample standard deviation of the estimated median survival time $\hat{t}_{0.5}$ across these estimators. We conducted 1000 simulation repetitions with a target sample size of n=50 and n=250. In order to get to the desired sample size n, we simulated n/p_0 observations to account for truncation, where p_0 is the true probability of observing a randomly selected subject from the target sample.

For these simulations, we generated the survival time from disease symptom onset, T, as $gamma(10, 1)$. The time from disease symptom onset to study entry, L, was generated as $gamma(\alpha_1, \beta_1)$, and the time from disease symptom onset to the end of study, R, was generated as $gamma(\alpha_2, \beta_2)$. These distributions were chosen to emulate the AD data described in Section 4.4. In the following models, we changed the values of (α_1, β_1) and (α_2, β_2) to adjust the percentage of truncated observations. In model 1, we set $(\alpha_1, \beta_1) = (4.5, 1.5)$ and $(\alpha_2, \beta_2) = (8, 2.5)$, which resulted in mild left and right truncation and a total of 30% of the observations truncated. In model 2, we reduced the left truncation and increased the right truncation by setting $(\alpha_1, \beta_1) = (3, 1)$ and $(\alpha_2, \beta_2) = (5, 2)$, which resulted in 55% of truncated observations. Here the values of (α_1, β_1) and (α_2, β_2) were the resulting parameter estimates for the AD truncation distribution in Section 4.4. In model 3, we set $(\alpha_1, \beta_1) = (5, 2)$ and kept $(\alpha_2, \beta_2) = (5, 2)$. This resulted in heavy left and right truncation and a total of 80% of the observations truncated. Figure 4.2 displays the bias of $\hat{F}_{SP}$, $\hat{F}_{NP}$, and $\hat{F}_{emp}$ across the 1st through 9th deciles of F_0 for the three models. Here $\hat{F}_{SP}$ has little bias regardless of sample size or truncation proportion, and $\hat{F}_{NP}$ is slightly biased in the right

Figure 4.2: Bias of distribution function estimators.



Bias of $\hat{F}_{SP}$ ($\cdots\cdots$), $\hat{F}_{NP}$ (— — —), and $\hat{F}_{emp}$ (———) at $t_{0.1}, ..., t_{0.9}$, which are the deciles of the true survival time distribution F_0. Here $F_0(t_{0.1})$ = 0.1, $F_0(t_{0.2})$ = 0.2, etc.

tail of the distribution under smaller sample sizes and heavy right truncation, and has little bias otherwise. The naïve estimator, $\hat{F}_{emp}$, is biased in all three models. The bias of $\hat{F}_{emp}$ in model 1 is negative since the proportion of missing observations due to left truncation is slightly greater than the proportion missing due to right truncation, and thus we are under sampling the smaller survival times. In model 2, this bias is both positive and larger in magnitude relative to model 1, since we are severely under sampling the larger survival times due to the heavy right truncation. In model 3, this bias is negative across the 1st through 4th deciles of F_0, and positive across the 5th through 9th deciles of F_0. The bias here is smaller in magnitude relative to model 2, since we are (almost) equally under sampling the smaller and larger survival times, and therefore the bias due to left truncation is canceling out some of the bias due to right truncation.

Table 4.1 compares (absolute) bias($\hat{F}$), SD($\hat{F}$), $\widehat{SE}(\hat{F})$, MSE($\hat{F}$), cov($\hat{F}$), bias($\hat{t}_{0.5}$), and SD($\hat{t}_{0.5}$) for $\hat{F}$ = $\hat{F}_{SP}$, $\hat{F}$ = $\hat{F}_{NP}$, and $\hat{F}$ = $\hat{F}_{emp}$. With the exception of bias($\hat{t}_{0.5}$) and SD($\hat{t}_{0.5}$), these statistics were averaged across the 1st through 9th deciles of F_0. For example, bias($\hat{F}$) in the first line of Table 4.1 represents the average absolute value of the bias corresponding to $\hat{F}_{SP}$ in the top left panel of Figure 4.2. For $\hat{F} = \hat{F}_{NP}$, $\widehat{SE}(\hat{F})$ was based on 200 bootstrap resamples. The median

Table 4.1: Simulation results

Model	q	n	Estimator	Bias($\hat{F}$)	SD($\hat{F}$)	$\widehat{\mathrm{SE}}(\hat{F})$	MSE($\hat{F}$)	Cov($\hat{F}$)	Bias($\hat{t}_{0.5}$)	SD($\hat{t}_{0.5}$)
1	0.30	50	$\hat{F}_{SP}$	0.005	0.069	0.075	0.005	0.927	-0.028	0.622
			$\hat{F}_{NP}$	0.021	0.070	0.069	0.005	0.910	-0.050	0.664
			$\hat{F}_{emp}$	0.039	0.057	0.045	0.005	0.453	0.339	0.485
1	0.30	250	$\hat{F}_{SP}$	0.005	0.031	0.031	0.001	0.941	-0.024	0.280
			$\hat{F}_{NP}$	0.006	0.031	0.031	0.001	0.940	-0.021	0.284
			$\hat{F}_{emp}$	0.040	0.025	0.020	0.003	0.326	0.382	0.230
2	0.55	50	$\hat{F}_{SP}$	0.010	0.088	0.101	0.008	0.880	0.042	0.977
			$\hat{F}_{NP}$	0.028	0.091	0.077	0.010	0.818	0.070	0.984
			$\hat{F}_{emp}$	0.151	0.054	0.042	0.028	0.438	-1.364	0.431
2	0.55	250	$\hat{F}_{SP}$	0.005	0.039	0.038	0.002	0.935	-0.017	0.345
			$\hat{F}_{NP}$	0.007	0.041	0.039	0.002	0.930	-0.003	0.404
			$\hat{F}_{emp}$	0.149	0.024	0.019	0.025	0.332	-1.338	0.196
3	0.80	50	$\hat{F}_{SP}$	0.005	0.092	0.118	0.009	0.900	0.034	1.055
			$\hat{F}_{NP}$	0.017	0.098	0.088	0.010	0.871	0.062	1.138
			$\hat{F}_{emp}$	0.043	0.055	0.043	0.006	0.426	-0.175	0.424
3	0.80	250	$\hat{F}_{SP}$	0.006	0.040	0.036	0.002	0.897	-0.030	0.348
			$\hat{F}_{NP}$	0.007	0.042	0.041	0.002	0.924	-0.024	0.377
			$\hat{F}_{emp}$	0.044	0.024	0.020	0.003	0.327	-0.161	0.184

Survival times simulated from a $gamma(10,1)$ distribution. Left and right truncation times correctly assumed to come from a $gamma(\alpha_1,\beta_1)$ and $gamma(\alpha_2,\beta_2)$ distribution, respectively. Model 1 corresponds to $(\alpha_1,\beta_1)=(4.5,1.5)$ and $(\alpha_2,\beta_2)=(8,2.5)$. Model 2 corresponds to $(\alpha_1,\beta_1)=(3,1)$ and $(\alpha_2,\beta_2)=(5,2)$. Model 3 corresponds to $(\alpha_1,\beta_1)=(5,2)$ and $(\alpha_2,\beta_2)=(5,2)$. Here q is the proportion of observations missing due to truncation and n is the size of the observed sample. $\hat{F}_{SP}$ denotes the SPMLE, $\hat{F}_{NP}$ denotes the NPMLE, and $\hat{F}_{emp}$ denotes the naïve empirical CDF which ignores double truncation. These estimators were all computed at $t_{0.1},...,t_{0.9}$, the 1st through 9th deciles of the true survival distribution F_0. For a given estimator $\hat{F}$, Bias($\hat{F}$) is the (absolute) difference between $\hat{F}$ and F_0, averaged across the 9 deciles. Here SD($\hat{F}$) is standard deviation of $\hat{F}$ across simulations, $\widehat{SE}(\hat{F})$ is estimated standard error of $\hat{F}$, MSE($\hat{F}$) is mean squared error of $\hat{F}$, and Cov($\hat{F}$) is 95% coverage, all averaged across the 9 deciles. Here $\hat{t}_{0.5}$ is the estimated median value based on $\hat{F}$. The true median value based on F_0 is $t_{0.5}=9.7$. Here Bias($\hat{t}_{0.5}$) = $\hat{t}_{0.5}$ - $t_{0.5}$ and SD($\hat{t}_{0.5}$) is the standard deviation of $\hat{t}_{0.5}$ across simulations.

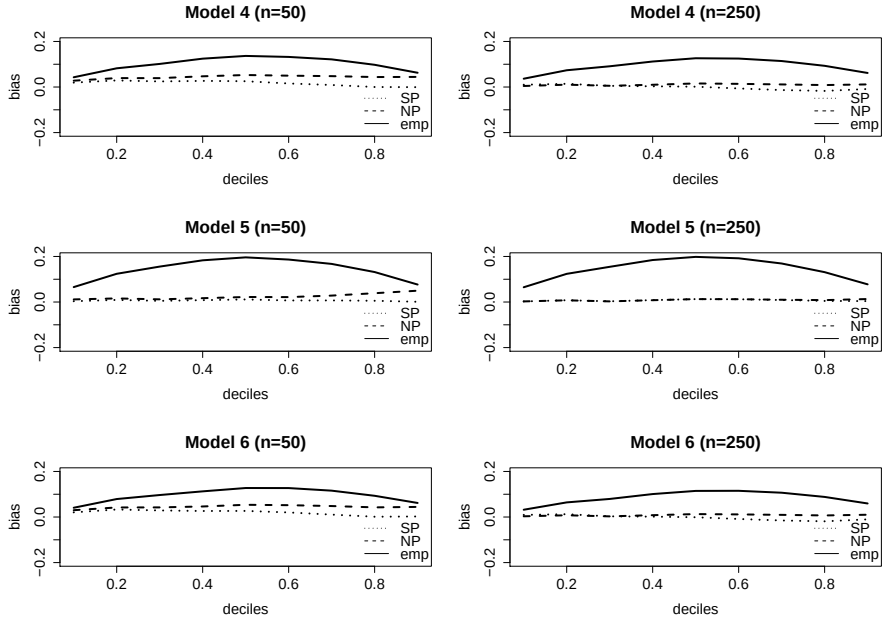
survival time ($t_{0.5}$) of the $gamma(10,1)$ distribution is 9.7. From Table 4.1, we see that $\hat{F}_{SP}$ and $\hat{F}_{NP}$ greatly outperform $\hat{F}_{emp}$ in terms of bias. Furthermore, with the exception of model 3 for n=50, $\hat{F}_{emp}$ has a greater MSE than $\hat{F}_{SP}$ and $\hat{F}_{NP}$ and is therefore less efficient. When the sample size is small, $\hat{F}_{SP}$ has a slightly lower MSE than $\hat{F}_{NP}$. The average coverage probabilities of the 95% confidence intervals for $\hat{F}_{SP}$ and $\hat{F}_{NP}$ are close to the nominal level of 0.95 when the sample size is large. This is not the case for $\hat{F}_{emp}$, where the coverage probabilities are not even close to the nominal level, even under mild truncation. The bias of the survival distribution and median survival time based on $\hat{F}_{emp}$ were much greater in model 2, since the truncation scheme in models 1 and 3

resulted in a sampling scheme that (almost) equally under sampled the smaller and larger survival times, and therefore the bias due to left truncation canceled out a large amount of the bias due to right truncation.

4.3.1. Robustness to misspecification of truncation distribution

Since $\hat{F}_{SP}$ requires distributional assumptions on the truncation times, we examine the impact of misspecification of the truncation distribution. We again assume $L \sim gamma(\alpha_1, \beta_1)$, $R \sim gamma(\alpha_2, \beta_2)$, and $T \sim gamma(10, 1)$. However, we now incorrectly specify the right truncation distribution by simulating $R \sim Unif[0, 20]$, and correctly specify the left truncation distribution by simulating $L \sim gamma(3, 1)$ in model 4. In model 5, we correctly specify the right truncation distribution by simulating $R \sim gamma(5, 2)$, and incorrectly specify the left truncation distribution by simulating $L \sim Weibull(1, 3)$. In model 6, we incorrectly specify both the left and right truncation distributions by simulating $L \sim Weibull(1, 3)$ and $R \sim Unif[0, 20]$.

Figure 4.3: Bias of distribution function estimators under misspecification of truncation distribution.



Bias of $\hat{F}_{SP}$ ($\cdots\cdots$), $\hat{F}_{NP}$ (— — —), and $\hat{F}_{emp}$ (———) at $t_{0.1}, ..., t_{0.9}$, under misspecification of the truncation distribution. Here $t_{0.1}, ..., t_{0.9}$ are the deciles of the true survival time distribution F_0, where $F_0(t_{0.1})$ = 0.1, $F_0(t_{0.2})$ = 0.2, etc.

Figure 4.3 displays the bias of $\hat{F}_{SP}$, $\hat{F}_{NP}$, and $\hat{F}_{emp}$ across the 1st through 9th deciles of F_0 for

models 4, 5, and 6. The bias of $\hat{F}_{SP}$ was still small in this setting. Table 4.2 shows that $\hat{F}_{SP}$ still performed as well as $\hat{F}_{NP}$ in terms bias and MSE. Furthermore, misspecification of the truncation distribution only resulted in a slight bias of the median survival time. However the standard error estimates for $\hat{F}_{SP}$ were biased when the right truncation distribution was misspecified. As expected, $\hat{F}_{emp}$ was heavily biased while $\hat{F}_{NP}$ remained unbiased, since neither of these estimators make distributional assumptions about the truncation times. We note that in (Moreira and Una-Alvarez, 2010; Shen, 2010b), the bias of $\hat{F}_{SP}$ was not robust to misspecification of the truncation distribution. However the simulations were based on an assumed beta distribution for the truncation times with only one parameter estimated. Here we assumed a gamma distribution with both parameters estimated, which allows more flexibility in estimating different distributions.

4.3.2. Robustness to independence violation between survival and truncation times

Both $\hat{F}_{SP}$ and $\hat{F}_{NP}$ assume that the survival and truncation times are independent. However this might not always be the case in practice. We therefore examine the robustness of these estimators when this independence assumption is violated. We simulate the survival and truncation times from a normal copula. The marginal distributions for the survival, left, and right truncation times are set to $gamma(10, 1)$, $gamma(3, 1)$, and $gamma(5, 2)$ distributions, respectively. Let ρ_{XY} denote the correlation between random variables X and Y. In model 7, we set $\rho_{LT} = 0.5$, $\rho_{LR} = 0.1$, and $\rho_{TR} = 0.1$. In model 8, we set $\rho_{LT} = -0.5$, $\rho_{LR} = 0.1$, and $\rho_{TR} = -0.1$. These correlations lead to a strong positive dependence (model 7) and strong negative dependence (model 8) between the left truncation times and survival times. We set $\rho_{LT} = -0.1$, $\rho_{LR} = 0.1$, and $\rho_{TR} = -0.5$ in model 9, which leads to a strong negative dependence between the survival times and right truncation times. In model 10, we set $\rho_{LT} = -0.5$, $\rho_{LR} = 0.1$, and $\rho_{TR} = -0.5$, which leads to a strong negative dependence between both the survival times and left truncation times as well as the survival times and right truncation times.

Figure 4.4 displays the bias of $\hat{F}_{SP}$, $\hat{F}_{NP}$, and $\hat{F}_{emp}$ across the 1st through 9th deciles of F_0 for models 7 through 10. The bias of $\hat{F}_{SP}$ and $\hat{F}_{NP}$ is relatively small when there is only a strong dependence between the left truncation and survival time (i.e. models 7 and 8). However the bias of these estimators become much more severe when there is a strong dependence with the right truncation time (i.e. models 9 and 10). As Table 4.3 shows, the coverage probabilities in this setting are extremely poor.

Table 4.2: Simulation results under misspecification of the truncation distribution

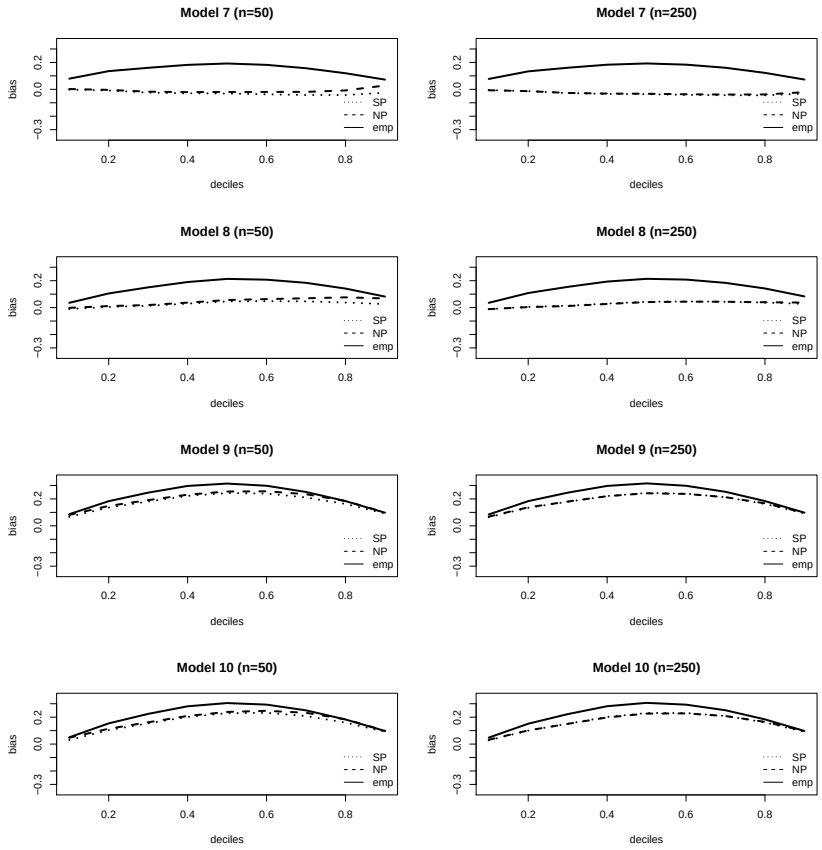
Model	q	n	Estimator	Bias($\hat{F}$)	SD($\hat{F}$)	$\widehat{\mathrm{SE}}(\hat{F})$	MSE($\hat{F}$)	Cov($\hat{F}$)	Bias($\hat{t}_{0.5}$)	SD($\hat{t}_{0.5}$)
4	0.52	50	$\hat{F}_{SP}$	0.008	0.074	0.160	0.006	0.968	0.091	0.725
			$\hat{F}_{NP}$	0.023	0.075	0.068	0.006	0.878	0.017	0.734
			$\hat{F}_{emp}$	0.092	0.056	0.044	0.013	0.455	-0.848	0.461
4	0.52	250	$\hat{F}_{SP}$	0.009	0.033	0.062	0.001	0.995	0.062	0.310
			$\hat{F}_{NP}$	0.007	0.033	0.033	0.001	0.932	-0.014	0.299
			$\hat{F}_{emp}$	0.091	0.025	0.020	0.010	0.301	-0.816	0.207
5	0.56	50	$\hat{F}_{SP}$	0.011	0.084	0.098	0.008	0.886	0.032	0.841
			$\hat{F}_{NP}$	0.029	0.087	0.077	0.009	0.834	0.073	0.954
			$\hat{F}_{emp}$	0.144	0.054	0.042	0.026	0.441	-1.297	0.423
5	0.56	250	$\hat{F}_{SP}$	0.007	0.038	0.037	0.002	0.921	-0.032	0.339
			$\hat{F}_{NP}$	0.008	0.040	0.039	0.002	0.921	-0.021	0.353
			$\hat{F}_{emp}$	0.144	0.024	0.019	0.024	0.331	-1.288	0.201
6	0.54	50	$\hat{F}_{SP}$	0.009	0.073	0.168	0.006	0.961	0.047	0.698
			$\hat{F}_{NP}$	0.027	0.073	0.068	0.006	0.873	-0.030	0.659
			$\hat{F}_{emp}$	0.088	0.056	0.044	0.012	0.455	-0.826	0.476
6	0.54	250	$\hat{F}_{SP}$	0.010	0.034	0.062	0.001	0.992	0.035	0.320
			$\hat{F}_{NP}$	0.009	0.035	0.032	0.001	0.924	-0.039	0.365
			$\hat{F}_{emp}$	0.086	0.025	0.020	0.009	0.301	-0.771	0.210

Survival times simulated from a $gamma(10,1)$ distribution. Left and right truncation times assumed to come from a $gamma(\alpha_1,\beta_1)$ and $gamma(\alpha_2,\beta_2)$ distribution, respectively. Model 4 corresponds to misspecification of the right truncation time by simulating it as $Unif(0,20)$, and the left truncation time as $gamma(3,1)$. Model 5 corresponds to misspecification of the left truncation time by simulating it as $Weibull(1,3)$, and the right truncation time as $gamma(5,2)$. Model 6 corresponds to misspecification both truncation times by simulating the left truncation time as $Weibull(1,3)$ and the right truncation time as $Unif(0,20)$. Here q is the proportion of observations missing due to truncation and n is the size of the observed sample. $\hat{F}_{SP}$ denotes the SPMLE, $\hat{F}_{NP}$ denotes the NPMLE, and $\hat{F}_{emp}$ denotes the naïve empirical CDF which ignores double truncation. These estimators were all computed at $t_{0.1},...,t_{0.9}$, the 1st through 9th deciles of the true survival distribution F_0. For a given estimator $\hat{F}$, Bias($\hat{F}$) is the (absolute) difference between $\hat{F}$ and F_0, averaged across the 9 deciles. Here SD($\hat{F}$) is standard deviation of $\hat{F}$ across simulations, $\widehat{SE}(\hat{F})$ is estimated standard error of $\hat{F}$, MSE($\hat{F}$) is mean squared error of $\hat{F}$, and Cov($\hat{F}$) is 95% coverage, all averaged across the 9 deciles. Here $\hat{t}_{0.5}$ is the estimated median value based on $\hat{F}$. The true median value based on F_0 is $t_{0.5}=9.7$. Here Bias($\hat{t}_{0.5}$) = $\hat{t}_{0.5}$ - $t_{0.5}$ and SD($\hat{t}_{0.5}$) is the standard deviation of $\hat{t}_{0.5}$ across simulations.

4.4. Example: Autopsy-confirmed Alzheimer's disease and frontotemporal lobar degeneration

Our motivating example comes from autopsy-confirmed data on individuals with either AD or FTLD retrieved from The Center for Neurodegenerative Disease Research at the University of Pennsylvania between 1995 and 2012. The target sample for the research purposes of the study consists of all individuals with either AD or FTLD onset before 2012, who either entered the center between

Figure 4.4: Bias of distribution function estimators under violation of independence assumption



Bias of $\hat{F}_{SP}$ $(\cdots\cdots)$*,* $\hat{F}_{NP}$ $(---)$*, and* $\hat{F}_{emp}$ (———) *at* $t_{0.1}, ..., t_{0.9}$*, under violation of independence between the survival and truncation times. Here* $t_{0.1}, ..., t_{0.9}$ *are the deciles of the true survival time distribution* F_0*, where* $F_0(t_{0.1})$ *= 0.1,* $F_0(t_{0.2})$ *= 0.2, etc.*

1995 and 2012, or *would* have entered the center between 1995 and 2012, had they not succumbed to the disease beforehand. Our observed sample contains all individuals who entered the center between 1995 and 2012, and had an autopsy-confirmed diagnosis of AD or FTLD before 2012. Individuals with AD or FTLD who met the study criteria but died before entering the center were not observed, yielding left truncated data. Furthermore, observations were only obtained from individuals who had an autopsy-confirmed diagnosis of AD or FTLD. Individuals who lived past the end of study date were not diagnosed, and therefore not included in our sample. Thus our data is also right truncated. Our data consists of 47 autopsy-confirmed AD subjects and 31 autopsy-confirmed FTLD subjects. The survival time of interest (T) is the time between disease symptom onset and death. The left truncation time (L) is the time between disease symptom onset and entry into the study (i.e. initial clinic visit). The right truncation time (R) is the time between disease symptom

Table 4.3: Simulation results under violation of the independence assumption

Model	ρ_{LT}	ρ_{LR}	ρ_{TR}	Estimator	Bias($\hat{F}$)	SD($\hat{F}$)	$\widehat{\mathrm{SE}}(\hat{F})$	MSE($\hat{F}$)	Cov($\hat{F}$)	Bias($\hat{t}_{0.5}$)	SD($\hat{t}_{0.5}$)
				$\hat{F}_{SP}$	0.031	0.042	0.039	0.003	0.865	0.373	0.422
7	0.5	0.1	0.1	$\hat{F}_{NP}$	0.027	0.043	0.040	0.003	0.883	0.347	0.454
				$\hat{F}_{emp}$	0.142	0.025	0.019	0.023	0.331	-1.294	0.202
				$\hat{F}_{SP}$	0.028	0.038	0.038	0.003	0.847	-0.236	0.311
8	-0.5	0.1	-0.1	$\hat{F}_{NP}$	0.028	0.039	0.037	0.003	0.832	-0.211	0.319
				$\hat{F}_{emp}$	0.146	0.024	0.019	0.025	0.331	-1.295	0.182
				$\hat{F}_{SP}$	0.173	0.028	0.034	0.034	0.048	-1.436	0.199
9	-0.1	0.1	-0.5	$\hat{F}_{NP}$	0.174	0.028	0.027	0.035	0.021	-1.437	0.204
				$\hat{F}_{emp}$	0.218	0.021	0.016	0.054	0.235	-1.786	0.163
				$\hat{F}_{SP}$	0.157	0.027	0.035	0.030	0.134	-1.301	0.185
10	-0.5	0.1	-0.5	$\hat{F}_{NP}$	0.157	0.027	0.027	0.029	0.089	-1.289	0.190
				$\hat{F}_{emp}$	0.204	0.020	0.017	0.050	0.246	-1.663	0.147

Survival and truncation times simulated from a normal copula with correlations ρ_{LT}, ρ_{LR}, and ρ_{TR}, where ρ_{XY} denotes the correlation between random variables X and Y. The marginal distributions for the survival, left, and right truncation times are set to $gamma(10,1)$, $gamma(3,1)$, and $gamma(5,2)$ distributions, respectively. This resulted in roughly 55% of truncated observations in all models. Here n is the size of the observed sample. $\hat{F}_{SP}$ denotes the SPMLE, $\hat{F}_{NP}$ denotes the NPMLE, and $\hat{F}_{emp}$ denotes the naïve empirical CDF which ignores double truncation. These estimators were all computed at $t_{0.1}, ..., t_{0.9}$, the 1st through 9th deciles of the true survival distribution F_0. For a given estimator $\hat{F}$, Bias($\hat{F}$) is the (absolute) difference between $\hat{F}$ and F_0, averaged across the 9 deciles. Here SD($\hat{F}$) is standard deviation of $\hat{F}$ across simulations, $\widehat{SE}(\hat{F})$ is estimated standard error of $\hat{F}$, MSE($\hat{F}$) is mean squared error of $\hat{F}$, and Cov($\hat{F}$) is 95% coverage, all averaged across the 9 deciles. Here $\hat{t}_{0.5}$ is the estimated median value based on $\hat{F}$. The true median value based on F_0 is $t_{0.5} = 9.7$. Here Bias($\hat{t}_{0.5}$) = $\hat{t}_{0.5}$ - $t_{0.5}$ and SD($\hat{t}_{0.5}$) is the standard deviation of $\hat{t}_{0.5}$ across simulations.

onset and the end of the study, which is taken to be July 1st, 2012. Due to double truncation, we only observe individuals with $L \leq T \leq R$.

The study and comparison of AD and FTLD are of importance because it gives us insight towards developing disease modifying therapies in the future. Our goal here is to estimate and compare the survival distributions for these two groups. Before we apply the SPMLE or NPMLE to estimate the survival distributions for AD and FTLD, we must test whether the survival times are independent of the truncation times for each group. We test this assumption using the test statistic introduced in (Martin and Betensky, 2005). The resulting tests did not reject the null hypothesis of independence at the $\alpha = 0.05$ level for either the AD or FTLD group. We can therefore proceed to apply the methods described in Section 4.2 to our data.

The NPMLE and eCDF were computed without any parametric assumptions on the survival or truncation times. For the AD group, the SPMLE was computed by assuming that the left truncation

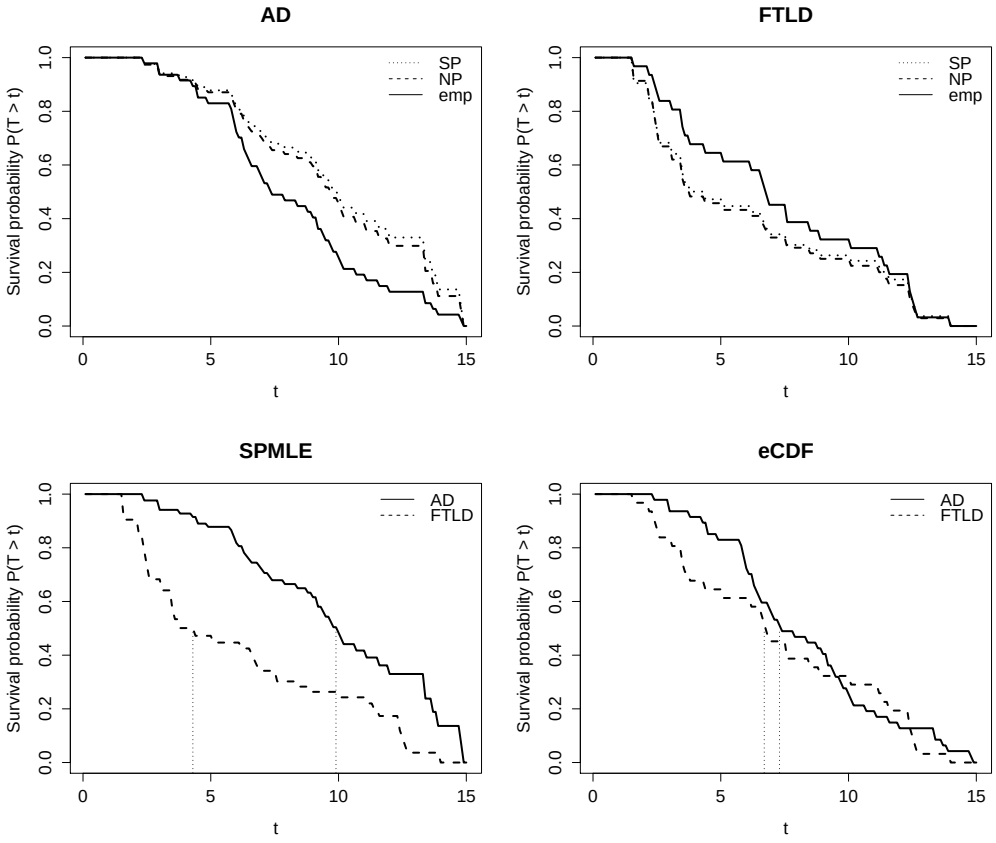
time has a $gamma(\alpha_1, \beta_1)$ distribution and the right truncation time has a $gamma(\alpha_2, \beta_2)$ distribution. The SPMLE for the FTLD group was estimated independently of the AD group, and assumed that the left truncation time has a $gamma(\theta_1, \gamma_1)$ distribution and the right truncation time has a $gamma(\theta_2, \gamma_2)$. The distribution of the truncation times were chosen by examining an external data set of individuals with clinically diagnosed AD and FTLD. Under these parametric assumptions, we have $(\hat{\alpha_1} = 2.9, \hat{\beta}_1 = 1.1)$, $(\hat{\alpha_2} = 5.2, \hat{\beta}_2 = 1.9)$, $(\hat{\theta_1} = 1.7, \hat{\gamma}_1 = 3.1)$, and $(\hat{\theta_2} = 12.7, \hat{\gamma}_2 = 1.1)$. Based on these results, the probability of observing an individual with AD or FTLD was estimated to be 0.42 and 0.46, respectively.

To check whether the choice of the gamma distribution is appropriate, we test the null hypothesis $H_0 : K = K_\theta$, independently for the AD and FTLD group, using a Kolmogorov-Smirnov type test statistic introduced in (Moreira, Ivarez, and Van Keilegom, 2014). The resulting test did not reject H_0 at the $\alpha = 0.05$ level for either AD or FTLD, and therefore we do not have enough evidence against the gamma distribution assumptions for the truncation times in either group.

The estimated survival curves $\hat{S}(t) = 1 - \hat{F}(t)$ based on the SPMLE, NPMLE, and eCDF are plotted in Figure 4.5. In the top left panel, we compare these three estimators for the AD group. The estimated survival probabilities based on the SPMLE and NPMLE are similar, and are greater than those based on the eCDF. This implies that right truncation had a greater impact than left truncation in the AD group. In other words, a greater proportion of larger survival times were unobserved relative to smaller survival times. The top right panel compares these estimators for the FTLD group. Here the estimated survival probabilities based on the SPMLE and NPMLE are also similar, but are less than those based on the eCDF. This implies that left truncation had a greater impact than right truncation in the FTLD group. In other words, a greater proportion of smaller survival times were unobserved relative to larger survival times.

The bottom row of Figure 4.5 compares the AD and FTLD survival probabilities based on the SPMLE (left) and the eCDF (right). When we do not adjust for double truncation, the eCDF concludes that the survival curves of AD and FTLD are nearly identical, with median survival times less than 1 year apart (AD = 7.3 years, FTLD = 6.7 years). This is not consistent with previous literature (Rascovsky et al., 2005). When we adjust for the double truncation, the survival probabilities for AD are greater than those of FTLD. Furthermore, the difference in median survival time is now greater than 5 years (AD = 9.9 years, FTLD = 4.3 years).

Figure 4.5: Estimated distribution functions for AD and FTLD



Top row: Estimated survival curves for AD (top left panel) and FTLD (top right panel) based on $\hat{F}_{SP}$ $(\cdots\cdots)$, $\hat{F}_{NP}$ $(---)$, *and* $\hat{F}_{emp}$ $(———)$.
Bottom row: Comparing AD $(———)$ *and FTLD* $(---)$ *survival curves based on the SPMLE* $\hat{F}_{SP}$ *(bottom left panel) and eCDF* $\hat{F}_{emp}$ *(bottom right panel). Vertical dotted lines represent median survival times for each group.*

We test for equality of the distribution functions of AD and FTLD using Bilker and Wang's semiparametric extension of the Mann-Whitney test. The resulting U-statistic is $\hat{U}$=2.62 with variance $\hat{V}_{\hat{U}}$ = 2.81. $\hat{U} > 0$ gives evidence that the survival curve for AD is greater than that for FTLD. However this result is not statistically significant (p-value = 0.12). We note that the standard log-rank test (ignoring truncation) resulted in a p-value of 0.46.

The Mann-Whitney test above tests whether two survival curves are equal. We now test for a difference in survival probabilities between AD and FTLD at specific time points. The results are provided in Table 4.4. When we adjust for double truncation, we conclude that the AD group has a greater survival probability than the FTLD group at years 3, 6 and 9. While the probability of

survival at 12 years is also greater for the AD group, the resulting test is not statistically significant (p=0.221). When we do not account for double truncation, we find no significant difference in the survival probabilities.

Table 4.4: Testing equality of survival probabilities between AD and FTLD

		AD	FTLD		
t	Estimator	$\hat{S}(t)$ $(\widehat{\mathrm{SE}}_t)$	$\hat{S}(t)$ $(\widehat{\mathrm{SE}}_t)$	W_t	p-value
3	SPMLE	0.94 (0.04)	0.64 (0.11)	6.80	0.009
	NPMLE	0.93 (0.03)	0.62 (0.12)	6.45	0.011
	eCDF	0.94 (0.04)	0.81 (0.07)	2.67	0.102
6	SPMLE	0.81 (0.06)	0.45 (0.10)	9.58	0.002
	NPMLE	0.79 (0.06)	0.43 (0.12)	7.05	0.008
	eCDF	0.70 (0.07)	0.61 (0.09)	0.66	0.417
9	SPMLE	0.62 (0.09)	0.26 (0.08)	8.56	0.003
	NPMLE	0.60 (0.08)	0.25 (0.09)	7.50	0.006
	eCDF	0.40 (0.07)	0.32 (0.08)	0.55	0.459
12	SPMLE	0.33 (0.11)	0.17 (0.07)	1.50	0.221
	NPMLE	0.30 (0.08)	0.15 (0.07)	1.28	0.258
	eCDF	0.13 (0.05)	0.19 (0.07)	0.59	0.444

$\hat{S}(t) = 1 - \hat{F}(t)$ *is survival probability at time* t*,* $\widehat{SE}_t$ *is the estimated standard error at time* t*,* W_t *is the Wald statistic comparing the survival probability between AD and FTLD at time t, for* t *=3, 6, 9, 12*

4.5. Discussion and Recommendations

Due to the inaccuracy of clinical diagnoses and a lack of available biomarkers, many studies of neurodegenerative diseases rely on autopsy-confirmed diagnoses. The purpose of this paper was to raise awareness of the selection bias in these studies and to highlight appropriate methods to account for it. We described how the selection bias arises due to the double truncation inherent in these studies, and showed that ignoring double truncation leads to biased estimators of the survival time distribution. To adjust for double truncation, we applied semiparametric and nonparametric maximum likelihood estimators of the survival time distribution. We conducted a simulation study to evaluate the performance of these estimators in a variety of settings, and applied these estimators to a data set consisting of autopsy-confirmed AD and FTLD individuals.

The simulation study confirmed that the SPMLE and NPMLE had little bias in small samples, while the naïve empirical CDF which ignores double truncation was heavily biased. We also found that the empirical CDF had a much larger mean squared error relative to the SPMLE and NPMLE under moderate to severe truncation. Furthermore, the 95% confidence intervals of the empirical CDF

were well below the nominal level, while those corresponding to the SPMLE and NPMLE were close to the nominal level under larger sample sizes.

When applied to our autopsy-confirmed data set, the survival probabilities based on the SPMLE and NPMLE were significantly greater for the AD group relative to the FTLD group at almost all time points. Furthermore, the difference in median survival time between AD and FTLD was over 5 years. However we did not have enough evidence to conclude that the survival curves were significantly different between the two groups. Application of the empirical CDF to the AD and FTLD groups found that the survival probabilities were similar between the two groups, with median survival time less than one year apart. This is contrary to the previous literature hypothesizing that survival in AD is greater than that of FTLD (Rascovsky et al., 2005).

We recommend the approach taken in our data example when estimating the survival time distribution of an autopsy-confirmed neurodegenerative disease, since this approach leads to consistent and more efficient estimators. Our approach consisted of first testing whether the truncation and survival times are independent, and then applying the SPMLE and NPMLE of the survival distribution function to the data. Based on our simulations and (Moreira and Una-Alvarez, 2010; Shen, 2010b), the SPMLE has a lower standard error and MSE than the NPMLE, and is therefore a more efficient estimator. However the SPMLE requires the correct distribution of the truncation times. Although incorrectly specifying the truncation distribution did not result in biased estimators of the survival time distribution in our simulation study, this is not always the case (Moreira and Una-Alvarez, 2010; Shen, 2010b). We therefore recommend testing the parametric assumptions of the SPMLE using the test statistics provided in (Moreira, Ivarez, and Van Keilegom, 2014)

The main limitation with the methods described in this paper is that they require independence of the truncation and survival times. This is not always a realistic assumption in individuals with neurodegenerative diseases. Our simulation studies showed that the estimators which adjust for double truncation are sensitive to this independence assumption. Therefore these estimators must be used with caution. While methods exist to test this independence assumption (Martin and Betensky, 2005), an extension of these methods is needed to adjust for dependent truncation and survival times in the presence of double truncation.

The double truncation inherent in autopsy-confirmed studies of neurodegenerative diseases and

methods to correct for it have so far received little attention in the literature. In this paper, we showed that ignoring double truncation leads to biased estimators of the survival time distribution, and outlined methods to adjust for it. The effects of ignoring double truncation in these studies was highlighted in our data example, where the estimated survival curves for AD and FTLD were not consistent with previous literature. Given the devastating effects of neurodegenerative diseases on patients, their caregivers, and society, it is imperative to adjust for double truncation in order to have accurate knowledge of the survival time distribution.

CHAPTER 5

A PACKAGE FOR ANALYZING TRUNCATED DATA IN R

5.1. Introduction

Truncation is a statistical phenomenon that has been shown to occur in a wide range of applications, including survival analysis, epidemiology, economics, and astronomy. Individuals who are subject to truncation provide no information to the investigator. *Left truncation* occurs when individuals who experience the terminating event before they are recruited into the study are unobserved. For example, when individuals are recruited into a study after some initiating event (e.g. age at disease onset), then individuals who experience the terminating event (e.g. death) before they enter the study will not be observed. *Right truncation* occurs when data is only recorded for individuals whose terminating event occurs before some specified time. For example, data retroactively pulled from a disease registry will only include individuals who experienced the event of interest by the study end date or the date of data extraction. Individuals who do not experience the event by this time will be unobserved (Bilker and Wang, 1996).

Double truncation, the simultaneous presence of left *and* right truncation, refers to the situation when observations are only record for data that fall within a subject-specific random interval. Bilker and Wang, 1996 noticed this issue in an epidemiological study of AIDS incubation times from HIV infection. Because the database only reported information for individuals who were diagnosed before a specific date, the data is subject to right truncation. This data is also subject to left truncation because data was not recorded for individuals who developed AIDS before 1982, as AIDS was unknown before then. Because smaller incubation times are less likely due be observed due to left truncation and large incubation times are less likely to be observed due to right truncation, double truncation results in a complex observational bias.

Prior to this decade, there have only been a few papers devoted to double truncation, and all of them dealt with the estimation of the survival distribution function for the event times. Efron and Petrosian, 1999 first introduced an iterative algorithm to compute the nonparametric maximum likelihood estimators (NPMLE) for the distribution of event times that are subject to double truncation. Shen, 2010a developed the asymptotic properties of this estimator, and also introduced an iterative

algorithm to jointly estimate both the NPMLE of the event time distribution and the truncation time distribution. Both of these methods have been implemented in the **DTDA** package in R (Moreira, Una-Alvarez, and Crujeiras, 2010).

In recent years double truncation has started gaining traction in the literature. In 2017, three methods were introduced to adjust the Cox regression model for doubly truncated data (Mandel et al., 2017; Rennert and Xie, 2017; Shen and Liu, 2017). However, there is no existing software to adjust the Cox regression model for doubly truncated data. In this paper, we introduce the R package **SurvTruncation**, which contains functions to fit the Cox regression model and compute the NPMLE of the distribution function for the event time and truncation times using the methods introduced in (Rennert and Xie, 2017) and (Shen, 2010a), respectively. In addition to obtaining estimates of the regression coefficients from the Cox regression model, this package also provides estimates of the standard error and confidence intervals, as well as p-values, which are all based on the simple bootstrap technique.

The remainder of this paper is organized as follows. In Section 5.2, we provide a brief review of the existing algorithms to compute the NPMLE of the event time distribution and truncation time distribution under double truncation, as well as the algorithm to adjust the Cox regression model under double truncation. In Section 5.3 we describe the **SurvTruncation** package and illustrate its use through the analysis of the AIDS data example. Concluding remarks and a discussion of future extensions are provided in Section 5.4.

5.2. Statistical methodology

We refer to *population random variables* as random variables from the target population and denote them without subscripts. We refer to *sampling random variables* as random variables from the observed sample and denote them with subscripts. Let T_i denote the observed survival time and $\boldsymbol{Z}_i(t)$ denote the observed $p \times 1$ covariate vector at time t for subject $i = 1, ..., n$, where n is the size of the observed sample. The left and right truncation times are denoted by L and R, respectively. Due to truncation, we observe the data vector $\{T, L, R, \boldsymbol{Z}(t)\}$ if and only if $L \leq T \leq R$.

Let $F(t)$ denote the cumulative distribution functions of T. Let $K(l, r)$ denote the joint cumulative distribution function of (L, R). For any cumulative distribution function H, define the left endpoint

of its support by $a_H = inf\{x : H(x) > 0\}$ and the right endpoint of its support by $b_H = inf\{x : H(x) = 1\}$. Let $H_L(l) = K(l, \infty)$ and $H_R(r) = K(\infty, r)$ denote the marginal cumulative distribution functions of L and R, respectively. The following methods assume that $a_{H_L} < a_F \leq a_{H_R}$ and $b_{H_L} \leq b_F < b_{H_R}$, which are required for identifiability of the estimators presented here (Rennert and Xie, 2017; Shen, 2010a; Woodroofe, 1985).

These methods also assume that the survival times are independent of the truncation times in the observed region $L \leq T \leq R$. This independence assumption is needed to estimate the probability that a subject with survival time T_i is not truncated and thus observed. These are referred to as selection probabilities, and are denoted by π_i, $i = 1, ..., n$. Here $\pi_i = P(L \leq T \leq R|T = T_i)$. Under the independence assumption, π_i is simply $P(L \leq T_i \leq R)$.

5.2.1. NPMLE of survival and truncation distribution functions

Here we present a slightly modified version of the algorithm described in (Shen, 2010a). Let $\varphi_i = F(R_i) - F(L_i)$, $i = 1, ..., n$. The NPMLE's of φ_i and π_i can be found using the following iterative algorithm:

Step 0) Set $\widehat{\varphi}_i^{(0)} = n^{-1} \sum_{j=1}^{n} 1_{\{L_i \leq T_j \leq R_i\}}$, for $i = 1, ..., n$.

Step 1) Set $\widehat{\pi}_i^{(1)} = \left(\sum_{j=1}^{n} \frac{1}{\widehat{\varphi}_j^{(0)}} \right)^{-1} \sum_{j=1}^{n} \frac{1_{\{L_j \leq T_i \leq R_j\}}}{\widehat{\varphi}_j^{(0)}}$, for $i = 1, ..., n$.

Step 2) Set $\widehat{\varphi}_i^{(1)} = \left(\sum_{j=1}^{n} \frac{1}{\widehat{\pi}_j^{(1)}} \right)^{-1} \sum_{j=1}^{n} \frac{1_{\{L_i \leq T_j \leq R_i\}}}{\widehat{\pi}_j^{(1)}}$, for $i = 1, ..., n$.

Step 3) For a prespecified error e, repeat steps 1 and 2 until $\sum_{i=1}^{n} |\widehat{\pi}_i^{(s)} - \widehat{\pi}_i^{(s-1)}| < e$.

The NPMLE of π_i and φ_i are given by $\widehat{\pi}_i = \widehat{\pi}_i^{(s)}$ and $\widehat{\varphi}_i = \widehat{\varphi}_i^{(s)}$, respectively. The NPMLE of the event time distribution at time t, $\widehat{F}_{np}(t)$, and the NPMLE of the joint distribution function for the truncation times at (l, r), $\widehat{K}_{np}(l, r)$, are then given by

$$\widehat{F}_{np}(t) = \big[\sum_{j=1}^{n} 1/\widehat{\pi}_j\big]^{-1} \sum_{j=1}^{n} \frac{1_{\{T_j \le t\}}}{\widehat{\pi}_j},$$
$$\widehat{K}_{np}(l,r) = \big[\sum_{j=1}^{n} 1/\widehat{\varphi}_j\big]^{-1} \sum_{j=1}^{n} \frac{1_{\{L_j \le l, R_j \le r\}}}{\widehat{\varphi}_j}.$$

More details can be found in (Shen, 2010a).

5.2.2. Estimating the regression coefficients from the Cox regression model

For a given time t, define $Y_i(t) = 1_{\{T_i \ge t\}}$ and $N_i(t) = 1_{\{T_i \le t\}}$. Let τ be a constant set to the largest observed survival time. The Cox regression model assumes that for a given subject with $p \times 1$ covariate vector $\boldsymbol{Z}_i(t)$, the hazard function at time t is given by $\lambda_i(t) = \lambda_0(t) e^{\boldsymbol{\beta}_0' \boldsymbol{Z}_i(t)}$, where $\lambda_0(t)$ is the true baseline hazard function and is unspecified. Here $\boldsymbol{\beta}_0$ is the true regression coefficient.

When subjects have unequal probabilities of selection, then the study sample will not be a representative sample of the underlying target population. In this situation the standard Cox regression coefficient estimator will be a biased estimator of $\boldsymbol{\beta}_0$. To adjust for biased samples due to double truncation, Rennert and Xie (2017) maximize the weighted Cox score function (Binder, 1992) using the estimated selection probabilities $\widehat{\pi}_i$ for $i = 1, ..., n$. The weighted Cox score function is given by

$$\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{1}{\pi_i} \left\{ \boldsymbol{Z}_i(t) - \frac{\sum_{j=1}^{n} \frac{1}{\pi_j} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)} \boldsymbol{Z}_j(t)}{\sum_{j=1}^{n} \frac{1}{\pi_j} Y_j(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j(t)}} \right\} dN_i(t) = \boldsymbol{0}. \tag{5.1}$$

Letting $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, ..., \widehat{\pi}_n)$, a consistent estimator of $\boldsymbol{\beta}_0$ is obtained by solving $\boldsymbol{U}_{\boldsymbol{w}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}) = \boldsymbol{0}$. We denote the resulting estimator by $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{w}}}$.

5.3. Overview of the package **SurvTruncation**

The package **SurvTruncation** contains a function to compute the NPMLE of the event time distribution and truncation time distribution when the event time is subject to double truncation. In addition, the **SurvTruncation** package includes a function to fit the Cox regression model to doubly truncated data. This section shows the usage of the **SurvTruncation** package by analyzing the

AIDS data set using both functions.

This package incorporates the methods introduced in Shen, 2010a and Rennert and Xie, 2017 described in Section 5.2. The package is composed of the following two functions that allow the user to fit these methods. These functions are:

```
cdfDT()  computes the NPMLE of the event time distribution and
truncation time distribution, when the event times
are subject to left and/or right truncation.


coxDT()  fits a Cox proportional hazards regression model when the
event times are subject to left and/or right truncation.
```

Tables 5.1 and 5.2 show a summary of the arguments for the functions `cdfDT` and `coxDT`, respectively. For the function `cdfDT`, the argument `boot` has a default setting of FALSE. If true, the standard error and 95% pointwise upper and lower confidence limits will be computed. Note that only the unique event times, number of events at each unique event time, and the event time cumulative distribution function and corresponding survival function are displayed (assuming display=TRUE). The remaining values must be called from the saved output (see Section 5.3.1). The R code for the functions `coxDT` and `cdfDT` are provided in Appendix E and F, respectively.

5.3.1. Data example

The AIDS Blood Transfusion Data were collected by the Center for Disease Control and retrieved from their registry database. The data set `AIDS`, included in the **SurvTruncation** package, consists of individuals who were infected with HIV from a contaminated blood transfusion on April 1, 1978. The infection time is the months from April 1, 1978 to HIV infection. The event of interest here is the induction time, which is the time from HIV infection to the development of AIDS. The data, taken from Klein and Moeschberger (2003), contains 295 infection and induction times for 258 adults and 37 children. The pediatric population was either infected in utero or at birth via the parent who received the contaminated blood transfusion. The infection time for the pediatric population is months from April 1, 1978 to birth.

Let t_{AIDS} denote the calendar time of the AIDS virus. Because AIDS was unknown prior to 1982,

any individual who developed AIDS before τ_{start} = January 1982 would not have been included in the data registry. Therefore we only observe cases with $\tau_{start} \leq t_{AIDS}$. Because cases reported after τ_{end} = July 1986 are not included to avoid inconsistent data and bias from reporting delay, we only observe cases with $t_{AIDS} \leq \tau_{end}$. Therefore the data is doubly truncated, as we only observe cases with $\tau_{start} \leq t_{AIDS} \leq \tau_{end}$.

Let T denote the induction time from HIV infection to the development of AIDS (`Induction.time` in `AIDS` data set). Let U denote the time from the contaminated blood transfusion (April, 1978) to HIV infection (`Infection.time` in `AIDS` data set). Due to double truncation, it can be shown that we only observe individuals with $L \leq T \leq R$ (Shen and Liu, 2017). Here L is the left truncation time (`L.time` in `AIDS` data set) and is equal to 45 months - months from contaminated blood transfusion (i.e. 45 - U). Here R is the right truncation time (`R.time` in `AIDS` data set) and is equal to L + 54 months.

5.3.2. Estimating the event time distribution using `cdfDT`

We apply the function `cdfDT` to estimate the distribution function of the time from HIV infection to the development of AIDS (i.e. T). In the `AIDS` data set, this variable is denoted by `Induction.time`. The

Arguments	Description
y	vector of event times
l	vector of left truncation times
r	vector of right truncation times
n.iter	maximum number of iterations
boot	Logical. Default=FALSE. If TRUE, the simple bootstrap method is applied to estimate the standard error and pointwise confidence intervals of the event time distribution
B.boot	Numeric value for number of bootstrap resamples. Default is 200.
joint	Logical. Default=FALSE. If TRUE, computes joint and marginal distributions of the truncation times
plot.cdf	Logical. Default is FALSE. If TRUE, the estimated cumulative distribution and survival functions of the event times are plotted. If boot=TRUE, confidence intervals are also plotted.
plot.joint	Logical. Default is FALSE. If TRUE, the estimated marginal distribution functions of the truncation times and the joint distribution of the truncation times, are plotted. Note: Plot will only be displayed if both plot.joint=TRUE and joint=TRUE.
display	Logical. Default is TRUE. If FALSE, output will not be displayed upon execution of function.

Table 5.1: Summary of the arguments of the function `cdfDT`.

Arguments	Description
formula	a formula object, with the response on the left of a operator, and the terms on the right. The response must be a survival object as returned by the `Surv` function.
L	vector of left truncation times
R	vector right truncation times
data	an optional data.frame vector, needed to interpret variables named in the `formula`
subset	an optional vector specifying a subset of observations to be used in the fitting process. All observations are included by default.
time.var	default = FALSE. If TRUE, specifies that time varying covariates are fit to the data.
subject	a vector of subject identification numbers. Only needed if time.var=TRUE.
B.SE.np	number of iterations for bootstrapped standard error (default = 200)
CI.boot	requests bootstrap confidence intervals (default==FALSE)
B.CI.boot	number of iterations for bootstrapped confidence intervals (default = 2000)
pvalue.boot	requests bootstrap confidence intervals (default==FALSE)
B.pvalue.boot	number of iterations for bootstrapped p-values (default = 200)
print.weights	requests the output of nonparametric selection probabilities (default==FALSE)
error	convergence criterion for nonparametric selection probabilities (default = 10e-6)
n.iter	maximum number of iterations for computation of nonparamteric selection probabilities (default = 10000)

Table 5.2: Summary of the arguments of the function `coxDT`.

left and right truncation times, L and R, are denoted by `L.time` and `R.time`, respectively. Below we request that the bootstrap technique be applied to estimate the standard error and 95% confidence limits of the distribution function for the time to development of AIDS from HIV infection, using 200 bootstrap resamples (`boot=TRUE,B.boot=200`). We also request the computation of the joint and marginal distributions of the truncation times (`joint=TRUE`). Finally, we request the plots for the estimated cumulative distribution function and survival function of the event time (`plot.cdf=TRUE`), as well as the marginal and joint distribution of the truncation times (`plot.joint=TRUE`).

```
> data(AIDS)
> fit1 <- cdfDT(AIDS$Induction.time,AIDS$L.time,AIDS$R.time,error=1e-6,
+         boot=TRUE,B.boot=200,joint=TRUE,plot.cdf=TRUE,plot.joint=TRUE)
number of iterations 21
time n.event cumulative.cdf survival
3       9         0.0136   0.9864
```

```
6        7          0.0238   0.9762
9       18          0.0495   0.9505
12      20          0.0771   0.9229
15      18          0.1022   0.8978
18      26          0.1393   0.8607
21      16          0.1636   0.8364
24      14          0.1861   0.8139
27      22          0.2232   0.7768
30      17          0.2558   0.7442
33      15          0.2867   0.7133
36      23          0.3384   0.6616
39      14          0.3750   0.6250
42       9          0.4016   0.5984
45       5          0.4187   0.5813
48      11          0.4640   0.5360
51      10          0.5110   0.4890
54       6          0.5424   0.4576
57       5          0.5737   0.4263
60       8          0.6343   0.3657
63       9          0.7131   0.2869
66       5          0.7664   0.2336
69       2          0.7949   0.2051
72       1          0.8136   0.1864
75       1          0.8339   0.1661
78       2          0.8937   0.1063
81       1          0.9296   0.0704
87       1          1.0000   0.0000
number of observations 295
```

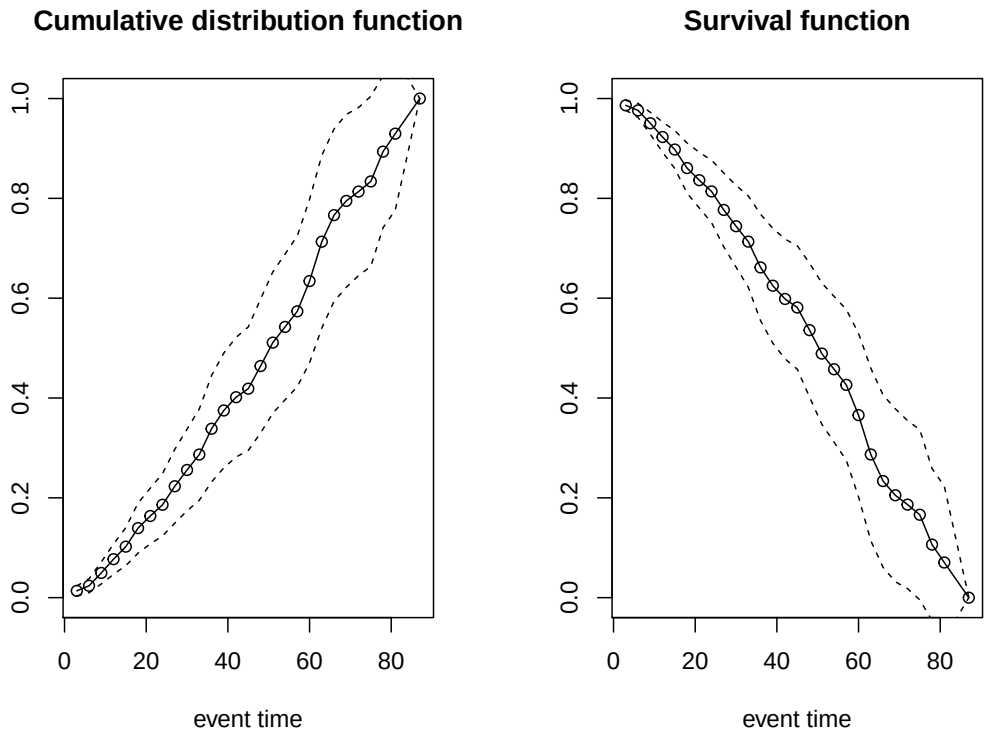
Running the function `cdfDT` automatically displays the number of unique event times, the number of observations for each unique event time, as well as estimates of the cumulative distribution function and survival function at each unique event time. The plots are automatically generated

and displayed in Figures 5.1, 5.2, and 5.3, respectively.

To display the remaining output, we need to call it. For example, the estimated selection probabilities for the first 5 subjects, $\widehat{\pi}_i = \hat{P}(L \leq T_i \leq R)$ for $i = 1, ..., 5$, can be called as follows:

```
> fit1$P.K[1:5]
[1] 0.16581658 0.03510388 0.16581658 0.16581658 0.01784253
```

Figure 5.1: NPMLE of the cumulative distribution function and survival function of the AIDS induction times.



Note: 95% upper and lower confidence limits displayed (since `boot=TRUE`).

5.3.3. Estimating the Cox regression coefficients using `coxDT`

The arguments for the function `coxDT` are similar to that of `coxph` in the `survival` package in R. Unlike the function `cdfDT`, here we can directly insert the variable names as long as we include the data set in the argument (e.g. `data=AIDS`). In this example, we specify 200 bootstrap resamples for estimation of the standard error for the regression coefficient (`B.SE.np=200`).

```
> data(AIDS)
```

```
> fit2 <- coxDT(Surv(Induction.time,status)~Adult,L.time,R.time,data=AIDS,
B.SE.np=200,print.weights=TRUE)
> fit2
$results.beta
Estimate     SE CI.lower CI.upper Wald statistic p-value
[1,]  -1.0545 0.5601   -2.152    0.043           3.54  0.0598

$CI
[1] "Normal approximation"
```

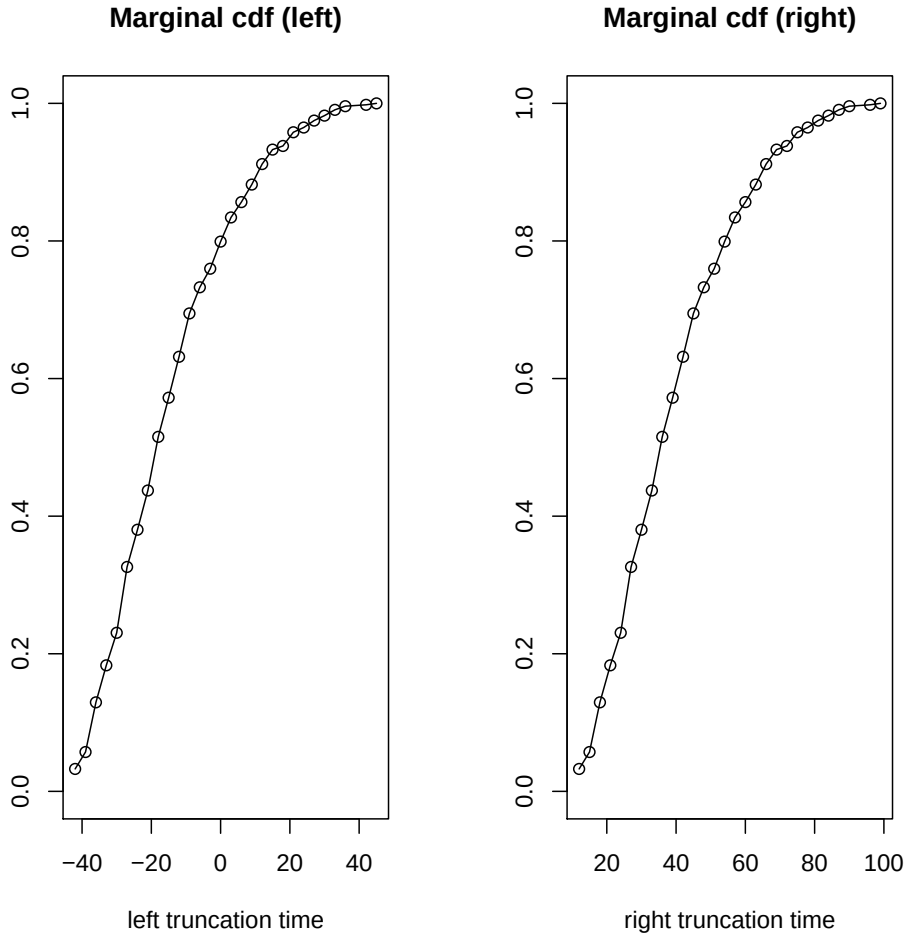


Figure 5.2: NPMLE of the marginal cumulative distribution function of left truncation time (left) and right truncation time (right).

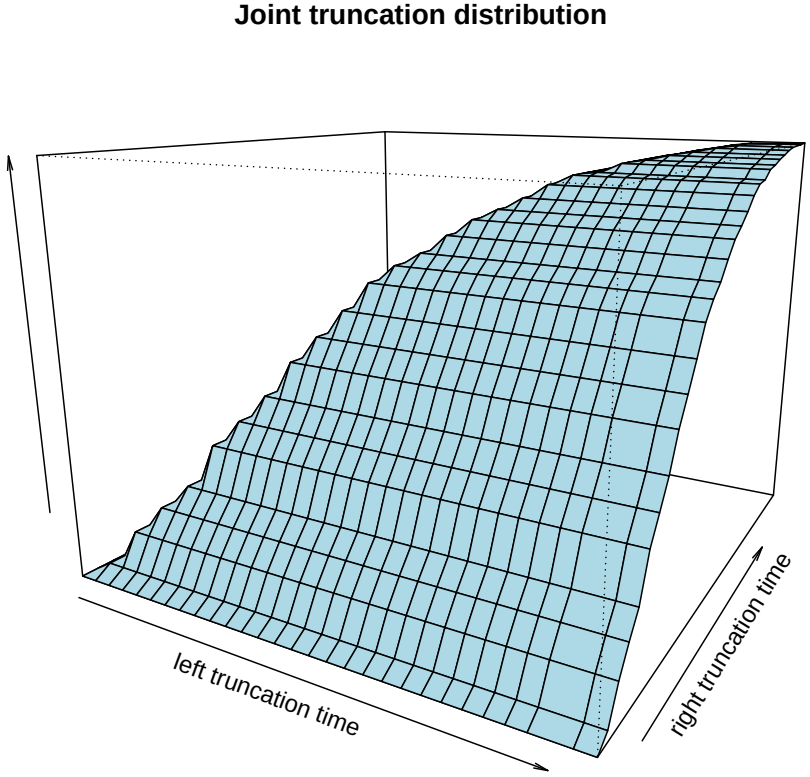


Figure 5.3: NPMLE of the joint cumulative distribution function of left and right truncation times.

```
$p.value
[1] "Normal approximation"


$weights
[1]  6.030760 28.486881  6.030760  6.030760 56.045875 . . .
```

The `Estimate` of the regression coefficient for adults is displayed in the output under `results.beta`. The value of -1.0545 indicates that adults are $\exp(-1.0545) = 0.35$ times less likely than children to develop AIDS after HIV infection. The output `CI = "Normal approximation"` and

`p.value = "Normal approximation"` indicate that the 95% confidence interval and Wald p-value were computed by assuming normality for the regression coefficient estimator $\widehat{\beta}_{\widehat{w}}$. The argument `print.weights=TRUE` outputs the weights for all subjects. Here we suppress the output to include the weights for the first 5 subjects only. Note that the weights displayed here are the inverse of the estimated selection probabilities which were output in the previous subsection. That is, the weights displayed here are simply $\widehat{w}_i = 1/\widehat{\pi}_i$ for $i = 1, ..., 5$.

The example provided here assumes that the covariates are time independent. The `coxDT` function can easily accommodate time-varying covariates in a similar manner to the `coxph` function. Details can be found in the help file for `coxDT`.

5.4. Conclusions

This paper discusses the implementation of software for the event time distribution function and Cox regression model when the survival times are subject to both left and right truncation. The event time distribution function is estimated in the **SurvTruncation** package in R using the algorithm introduced in Shen, 2010a. The Cox regression model is fit using the weighted estimating equation approach introduced in Rennert and Xie, 2017, which uses the inverse of the estimated selection probabilities from (Shen, 2010a) as weights.

The function `cdfDT` displays both numerical output and graphical displays of the estimates of the survival and truncation time distributions. Both `cdfDT` and the function to estimate the Cox regression model, `coxDT`, estimate the standard errors by using the simple bootstrap technique. The `CoxDT` function also allows for estimating the 95% confidence intervals and p-values of the regression coefficient estimators by the simple bootstrap technique. We note that both methods assume that the observed survival and truncation times are independent.

To our knowledge, the **SurvTruncation** package is the first to implement the Cox regression model under double truncation in a friendly way. The arguments for the `coxDT` function are similar to the `coxph` in the **Survival** package in R. The `coxDT` function also handles time-varying covariates in a similar fashion to `coxph`. The implementation of the Shen (2010) algorithm was previously done in the `shen` function in the **DTDA** package in R (Moreira, Una-Alvarez, and Crujeiras, 2010). While the output from `cdfDT` and `shen` are similar, the function `cdfDT` has a significantly

faster computation time. The **SurvTruncation** package can be downloaded at the following link: https://github.com/rennertl/SurvTruncation.

CHAPTER 6

DISCUSSION

The double truncation inherent in autopsy-confirmed studies of neurodegenerative diseases and methods to correct for it have so far received little attention in the literature. Due to the inaccuracy of clinical diagnoses and a lack of available biomarkers, many studies of neurodegenerative diseases rely on autopsy-confirmed diagnoses. We described how the selection bias arises due to the double truncation inherent in these studies, and showed that ignoring double truncation leads to biased estimators of the regression coefficients from the Cox regression model. In Chapter 2, we introduced a weighted estimating equation approach to adjust the Cox regression model under double truncation, by weighting the subjects in the score equation of the Cox partial likelihood by the inverse of the probability that they were observed (i.e. *not* truncated). The probability of being observed was estimated both parametrically and nonparametrically by methods introduced in Shen (2010; 2010) and Moreira and de Ũna-Álvarez (2010), and did not require any contribution from missing subjects. We proved the resulting regression coefficient estimators are consistent. The simulation studies confirmed that these estimators had little bias, while the naïve estimator which ignores truncation is biased. We proved the parametric weighted estimator is asymptotically normal, and a consistent estimator of its asymptotic variance was provided. Our simulations showed that the bootstrap estimate of the standard error for the nonparametric weighted estimator matched the observed sample standard deviation.

The consistency of the estimated selection probabilities used in this method rests on the assumption of independence between the survival and truncation times in the observable region. We showed in Chapter 3 that a violation of this assumption leads to biased estimators of regression coefficient estimators. We therefore proposed a novel method in Chapter 3 which relaxes the independence assumption between the observed survival and truncation times in the Cox model under left, right, or double truncation to an assumption of conditional independence between the observed survival and truncation times. We obtained consistent and asymptotically normal estimators of the regression coefficients and baseline hazard function by maximizing the conditional likelihood of the observed survival times using an EM algorithm. The simulation studies confirmed that the proposed estimators had little bias in small samples, while the naïve estimators from the

Cox models which ignore truncation or assume independence were biased. The existing methods which adjust for truncation but assume independence resulted in heavily biased estimators of the regression coefficients for risk factors of survival that were also correlated with the truncation times. Furthermore, the proposed estimators had a lower mean-squared error than the naïve estimators in most of the simulation settings.

We also conducted a simulation and case study to examine survival time distribution function estimators under double truncation. We showed that the SPMLE and NPMLE of the survival distribution function had little bias in small samples, while the naïve empirical CDF which ignores double truncation was heavily biased. We found that the empirical CDF had a much larger mean squared error relative to the SPMLE and NPMLE under moderate to severe truncation. Furthermore, the 95% confidence intervals of the empirical CDF were well below the nominal level, while those corresponding to the SPMLE and NPMLE were close to the nominal level under larger sample sizes.

When applied to our autopsy-confirmed data set, the survival probabilities based on the SPMLE and NPMLE were significantly greater for the AD group relative to the FTLD group at almost all time points. Furthermore, the difference in median survival time between AD and FTLD was over 5 years. Application of the empirical CDF to the AD and FTLD groups found that the survival probabilities were similar between the two groups, with median survival time less than one year apart. This is contrary to the previous literature hypothesizing that survival in AD is greater than that of FTLD (Rascovsky et al., 2005).

The main limitation with the SPMLE and NPMLE of the survival time distribution is that they require independence of the truncation and survival times. As shown in Chapter 3 and Chapter 4, these methods are very sensitive to this independence assumption. Therefore these estimators must be used with caution. An extension of these methods is needed to adjust for dependent truncation and survival times in the presence of double truncation.

The function to compute the Cox regression model using the weighted estimating equation approach in Chapter 2, and the function to estimate the survival and truncation time distributions under nonparametric assumptions, are implemented in our **SurvTruncation** package in R. This package is described in Chapter 5. To our knowledge, this package is the first to implement the Cox regression model under double truncation in a friendly way. The arguments for the `coxDT`

function are similar to the `coxph` in the **Survival** package in R. The `coxDT` function also handles time-varying covariates in a similar fashion to `coxph`. The implementation of the Shen (2010) algorithm was previously done in the `shen` function in the **DTDA** in R (Moreira, Una-Alvarez, and Crujeiras, 2010). While the output from `cdfDT` and `shen` are similar, the function `cdfDT` has a significantly faster computation time. This package can be downloaded at the following link: https://github.com/rennertl/SurvTruncation.

We applied our proposed methods from Chapters 2 and 3 to assess the effect of cognitive reserve on survival in an autopsy-confirmed sample of individuals with Alzheimer's disease (AD). AD is a major neurodegenerative disease which currently affects 5.3 million people in the United States according to the Alzheimer's Association. In 2017 alone, AD and other dementias will have cost the nation an estimated $259 billion. Therefore it is crucial to determine factors affecting survival. Using both education and highest occupational attainment as proxies for cognitive reserve, our data analyses concluded that cognitive reserve prolongs survival in subjects with Alzheimers disease.

Our proposed methods have useful implications for observational studies beyond autopsy-confirmed neurodegenerative diseases. Double truncation has been shown to be present in a variety of studies, such as those of clinically diagnosed Parkinson's disease (Mandel et al., 2017), childhood cancer (Moreira and Una-Alvarez, 2010), astronomy data (Efron and Petrosian, 1999), and studies based on registry data (Bilker and Wang, 1996; Shen and Liu, 2017). In fact, any data pulled from a disease registry will be subject to inherent right truncation, since data is only recorded for subjects who have the disease and are entered in the registry by the time the data is extracted (Bilker and Wang, 1996). In certain cases, the data will also be subject to left truncation (Bilker and Wang, 1996; Shen and Liu, 2017). In a similar fashion, studies which only include data from individuals whose event times fall within the time course of the study are subject to double truncation (Moreira and Una-Alvarez, 2010). Therefore careful consideration of the study design must be taken into account when fitting the Cox proportional hazards model. Furthermore, the assumption of independence should always be tested, given the high sensitivity of existing methods to this assumption. For example, a quick application of a Kendall's conditional Tau test (Martin and Betensky, 2005) revealed this independence assumption is violated in the AD data set analyzed in Chapter 3 and the AIDS data used in Shen and Liu (2017). When time varying covariates are not of interest, we recommend estimating the regression coefficients using the EM method in Chapter

3, since the resulting estimators have little bias, and in most situations, have a lower mean-squared error compared to existing estimators under left, right, or double truncation, and under a wide range of dependence structures. When time-varying covariates are of interest, the weighted estimating equation approach from Chapter 2 is more suitable, as long as the independence assumption is not violated. Future methods are needed to develop methods to handle time varying covariates under double truncation when the assumption of independence is violated.

APPENDIX A

REGULARITY ASSUMPTIONS OF PROPOSED E-M ESTIMATOR

Regularity Assumptions

1. The true hazard function $\lambda_0(\cdot)$ is continuously differentiable, $\Lambda_0(0) = 0$, and $\Lambda_0(\tau) < \infty$.

2. The true parameter vector $\boldsymbol{\beta}_0$ lies in a compact set $\mathbb{B}$. The set $\mathbb{A}$ contains all nondecreasing functions Λ satisfying regularity assumption 1.

3. $E\|Z\|$ and $E\|\exp(\boldsymbol{\beta}'\boldsymbol{Z})\|$ are bounded, where $\|\boldsymbol{z}\| \equiv \sqrt{z_1^2 + ... + z_p^2}$.

4. The information matrix $-\partial^2 E[l_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}))]/\partial\boldsymbol{\beta}^2|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is positive definite. Here $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \boldsymbol{\lambda}_{em}$ is used to emphasize the dependence on $\boldsymbol{\beta}$.

5. If $P(\boldsymbol{b}'\boldsymbol{Z} = c) = 1$ for some constant c, then $\boldsymbol{b} = 0$.

Assumptions 1 and 2 are required for stochastic approximation. Assumptions 3 and 4 are needed to establish the asymptotic properties of the regression parameter estimates from the Cox model (Andersen et al., 1997). Assumption 5 implies no covariate colinearity and thus ensures that the model is identifiable.

APPENDIX B

PROOF OF THEOREM 3.1

Since each function of $\boldsymbol{\lambda}$ in $l_n(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is concave or strictly concave, and the summation of concave functions is concave, the log-likelihood function $l_n(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is strictly concave in $\boldsymbol{\lambda}$. Therefore we can find a unique maximizer $\widehat{\boldsymbol{\lambda}}(\cdot, \boldsymbol{\beta})$ of $l_n(\boldsymbol{\beta}, \boldsymbol{\lambda})$ for each $\boldsymbol{\beta}$ in a compact set $\mathbb{B}$. The existence of the NPMLE for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ follows by compactness of $\mathbb{B}$ for the likelihood $l_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\lambda}}(\cdot, \boldsymbol{\beta}))$, which is continuous in $\boldsymbol{\beta}$. Uniqueness is guaranteed by Assumption 4 in Appendix A for large samples.

Here we show that if $\widehat{\boldsymbol{\theta}}_n$ converges, it must converge to $\boldsymbol{\theta}_0$. As $\widehat{\boldsymbol{\theta}}_n$ maximizes the log-likelihood given in (3.2), $l_n(\boldsymbol{\theta})$, the empirical Kullback-Leibler distance $l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)$ must be nonnegative. Suppose $\widehat{\boldsymbol{\theta}}_n$ converges to some $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \Lambda^*)$. Then by the strong law of large numbers (SLLN), $l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)$ must converge to the negative Kullback-Leibler distance between $P_{\boldsymbol{\theta}^*}$ and $P_{\boldsymbol{\theta}_0}$. Here $P_{\boldsymbol{\theta}}$ is the probability measure under the parameter $\boldsymbol{\theta}$. Since the Kullback-Leibler distance and $l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)$ are nonnegative, the Kullback-Leibler distance between $P_{\boldsymbol{\theta}^*}$ and $P_{\boldsymbol{\theta}_0}$ must be zero. Therefore $P_{\boldsymbol{\theta}^*} = P_{\boldsymbol{\theta}_0}$ almost surely, and it then follows from model identifiability that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. Therefore if $\widehat{\boldsymbol{\theta}}_n$ converges, it must converge to $\boldsymbol{\theta}_0$.

The technical details to show that $\widehat{\boldsymbol{\theta}}_n$ indeed converges are similar to those in (Murphy, 1994). The idea is to find a further convergent subsequence for any subsequence of $\widehat{\boldsymbol{\theta}}_n$, and then apply Helly's selection theorem. Here we provide only a sketch of the proof. The first step is to show that $\widehat{\boldsymbol{\theta}}_n$ stays bounded. By regularity assumption 3, $\widehat{\boldsymbol{\beta}}_n$ is in a compact set and is therefore bounded. To show $\widehat{\Lambda}_n$ is bounded, we make use of the fact that the empirical Kullback-Leibler distance $l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\bar{\boldsymbol{\theta}})$ is always non-negative for each $\bar{\boldsymbol{\theta}}$ in the parameter set. Using the approach of Murphy (1994), it can be shown that if $\widehat{\Lambda}_n$ does indeed diverge to $\pm\infty$, then it is possible to construct some sequence $\bar{\boldsymbol{\theta}}_n$ such that $l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\bar{\boldsymbol{\theta}})$ eventually becomes negative infinity, which contradicts the nonnegativity of the empirical Kullback-Leibler distance.

Since $\widehat{\boldsymbol{\theta}}_n$ stays bounded, we can apply Helly's selection principal to find a further convergent subsequence $\widehat{\boldsymbol{\theta}}_{n_k} = (\widehat{\boldsymbol{\beta}}_{n_k}, \widehat{\Lambda}_{n_k})$ for any subsequence of $\widehat{\boldsymbol{\theta}}_n$ indexed by $\{1, ..., n\}$. By the classical Kullback-Leibler information approach, and the SLLN, $\widehat{\boldsymbol{\theta}}_{n_k}$ must converge to $\boldsymbol{\theta}_0$. It then follows from Helly's selection theorem that the entire sequence $(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n(t))$ must converge to $(\boldsymbol{\beta}_0, \Lambda_0(t))$ for ev-

ery $t \in [0, \tau]$, where $\tau = t_d$ is the maximum of the observed event times. Since $\Lambda_0(\cdot)$ is assumed to be monotone and continuous, the convergence of $\widehat{\Lambda}_n(t)$ is uniformly in $t \in [0, \tau]$. Because the proof is carried out for a fixed ω in the underlying probability space Ω, where the SLLN is applied countably many times, the convergence here is also almost surely a true convergence.

APPENDIX C

PROOF OF THEOREM 3.2

Here we outline the proof for the weak convergence of $\widehat{\boldsymbol{\theta}}_n$, which follows the proof for weak convergence in (Qin et al., 2011). The proof consists of the application of empirical process theory and the Z-theorem for infinite dimensional estimating equations (Vaart and Wellner, 2000).

Denote the score equation for $\boldsymbol{\beta}$ by $\boldsymbol{U}_{1n}(\boldsymbol{\theta}) = \partial l_n(\boldsymbol{\theta})/\partial\boldsymbol{\beta}$. To obtain the score equation $\Lambda(\cdot)$, we define the submodel $d\Lambda_\epsilon = (1+\epsilon h)d\Lambda$, where h is a bounded and integrable function. Setting $h(\cdot) = I(\cdot \leq t)$, the score equation for Λ is given by $\boldsymbol{U}_{2n}(t,\boldsymbol{\theta}) = \frac{\partial l_n(\boldsymbol{\beta},\Lambda_\epsilon)}{\partial\epsilon}|_{\epsilon=0}$.

We denote the vector of the score functions by $\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}) = [\boldsymbol{U}_{1n}(\boldsymbol{\theta}), \boldsymbol{U}_{2n}(t,\boldsymbol{\theta})]$. The expectation E_0 under the true value $\boldsymbol{\theta}_0$ is given by $\boldsymbol{U}_0(\cdot,\boldsymbol{\theta}) = [\boldsymbol{U}_{10}(\boldsymbol{\theta}), \boldsymbol{U}_{20}(t,\boldsymbol{\theta})]$, where $\boldsymbol{U}_{10}(\boldsymbol{\theta}) = E_0[\boldsymbol{U}_{1n}(\boldsymbol{\theta})]$ and $\boldsymbol{U}_{20}(t,\boldsymbol{\theta}) = E_0[\boldsymbol{U}_{2n}(\cdot,\boldsymbol{\theta})]$.

By the definition of the MLE, $\boldsymbol{U}_n(\cdot,\widehat{\boldsymbol{\theta}}_n) = 0$. Since $\boldsymbol{U}_0(\cdot,\boldsymbol{\theta}_0) = 0$, we can show that $|\sqrt{n}\{\boldsymbol{U}_0(\cdot,\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{U}_n(\cdot,\widehat{\boldsymbol{\theta}}_n)\} - \sqrt{n}\{\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}_0) - \boldsymbol{U}_0(\cdot,\boldsymbol{\theta}_0)\}| = o_p(1)$. The estimating equation evaluated at $\boldsymbol{\theta}_0$, $\sqrt{n}\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}_0) = \sqrt{n}\{\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}_0) - \boldsymbol{U}_0(\cdot,\boldsymbol{\theta}_0)\}$, is a sum of iid terms. We can therefore use empirical process theory to show that $\sqrt{n}\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}_0)$ converges weakly to $\mathbb{W} = (\mathbb{W}_1, \mathbb{W}_2)$, where $\mathbb{W}_1$ is a Gaussian random vector and $\mathbb{W}_2$ is a tight Gaussian process. The covariance matrix for $\mathbb{W}_1$ is given by $\boldsymbol{\Sigma}_{11} = E_0\{\boldsymbol{U}_{1n}(\boldsymbol{\theta}_0)^{\otimes 2}\}$, and the covariance between $\mathbb{W}_2(s)$ and $\mathbb{W}_2(t)$ is given by $\Sigma_{22}(s,t) = E_0\{\boldsymbol{U}_{2n}(s,\boldsymbol{\theta}_0)\boldsymbol{U}_{2n}(t,\boldsymbol{\theta}_0)'\}$.

Applying the Z-theorem for the infinite dimensional estimating equations, Theorem 3.3.1 in Van der Vaart and Wellner (2000), we have that under the regularity conditions in A.1, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges weakly to a tight mean-zero Gaussian process $-\dot{U}^{-1}_{\boldsymbol{\theta}_0}(\mathbb{W})$.

Here $\dot{U}_{\boldsymbol{\theta}_0}$ is the Fréchet derivative of the map $\boldsymbol{U}_0(\cdot,\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$. Using arguments similar to Appendix A.5 in (Qin et al., 2011), we can show $\boldsymbol{U}_0(\cdot,\boldsymbol{\theta})$ is Fréchet differentiable and the Fréchet derivative, $\dot{U}_{\boldsymbol{\theta}_0}$, is continuously invertible. By definition of the Fréchet derivative, we have that $\dot{U}_{\boldsymbol{\theta}_0}\{\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\} = -\sqrt{n}\{\boldsymbol{U}_n(\cdot,\boldsymbol{\theta}_0) - \boldsymbol{U}_0(\cdot,\boldsymbol{\theta}_0)\} + o_p(1)$. This completes the proof.

APPENDIX D

R FUNCTION FOR COX REGRESSION COEFFICIENT ESTIMATOR UNDER DOUBLE TRUNCATION USING EM ALGORITHM

```
# user-defined functions that will be called in the algorithm
fun.geq=function(a,b) I(a>=b)*1;
fun.leq=function(a,b) I(a<=b)*1;
fun.eq=function(a,b) I(a==b)*1;
fun.length.which=function(a,b) length(which(a==b));

# function to compute the baseline hazard function
lambda.0=function(beta,y,z,t) return(1/apply(sapply(y,t,FUN="fun.geq")
%*%as.matrix(exp(z%*%beta),nrow=length(y)),1,sum))

# function to compute the weight vector w
fun.W=function(beta,lambda,y.unique,y,z,u,v) {
n=length(y);
temp.1=sapply(y.unique,y,FUN="fun.eq") # n x d matrix I(T_i = t_j)
# computing n x d matrix I(t_j < U_i) + I(t_j > V_i)
temp.2=sapply(y.unique,u,FUN="fun.leq")+sapply(y.unique,v,FUN="fun.geq")
## checked this- it works
temp.3=exp(z%*%beta)%*%t(lambda);
temp.4=exp(-exp(z%*%beta)%*%t(cumsum(lambda)));
## checked this- it works
temp.5=matrix(rep(exp(-exp(z%*%beta)*sapply(y.unique,u,FUN="fun.leq")
%*%lambda),length(y.unique)),nrow=n,ncol=length(y.unique));
temp.6=matrix(rep(exp(-exp(z%*%beta)*sapply(y.unique,v,FUN="fun.leq"
%*%lambda),length(y.unique)),nrow=n,ncol=length(y.unique));
W=temp.1+temp.2*temp.3*temp.4/(temp.5-temp.6)
return(W)}
```

```
# fun.EM: function to implement the EM algorithm
# formula = same as formula from coxph function in Survival package
# u and v are left and right truncation times, respectively
# Difference between successive estimates of beta.EM must be less than
# prespecified error and before max number of iterations n.iter achieved
fun.EM=function(formula,u,v,error,n.iter) {

# extracting the variable names
mf = model.frame(formula=formula);
z=model.matrix(attr(mf,"terms"),data=mf)[,-1]
y=model.response(mf);

data=data.frame(y,u,v,z); n=dim(data)[1];
newdata=data[order(y),]; # Ordering data set by survival time
y=as.numeric(newdata$y)[1:n]; u=newdata$u; v=newdata$v;
z=as.matrix(newdata[,4:dim(data)[2]]); y.unique=unique(y)

# number of unique observations, and covariates
d=length(y.unique); num.cov=dim(newdata)[2]-3;

# Beginning EM Algorithm
beta.EM=matrix(0,nrow=n.iter,ncol=num.cov)
lambda.EM=matrix(0,nrow=n.iter,ncol=d)

# Step 1 (initial values)
beta.EM[1,]=coxph(formula)$coefficients
lambda.EM[1,]=lambda.0(beta.EM[1,],y,z,y.unique)

# Creating weights
W=fun.W(beta.EM[1,],lambda.EM[1,],y.unique,y,z,u,v);
```

```
w=as.vector(t(W));
w.plus.j=apply(W,2,"sum")
w.i.plus=apply(W,1,"sum")

# Creating new data to apply coxph function with weight vector of length n*d
y.new=rep(y.unique,length(y))
status.obs.new=rep(1,length(y.new))
z.temp=matrix(0,nrow=n*d,ncol=num.cov)
for(i in 1:num.cov) z.temp[,i]=rep(z[,i],each=d)
znam <- paste0("z.new", 1:num.cov)
colnames(z.temp)=znam;

new.data=data.frame(y.new,status.obs.new,z.temp)
new.formula=as.formula(paste("Surv(y.new,status.obs.new) ~ ",
paste(znam, collapse= "+")))

# Step 2 (Maximizing expected complete data likelihood)
beta.EM[2,]=coxph(new.formula,data=new.data,weights=w,
subset=which(w>0))$coefficients

temp=t(W)%*%exp(z%*%beta.EM[2,]) # column j of W times exp(z*beta)
lambda.EM[2,]=sapply(1:d, function(j) w.plus.j[j]/sum(temp[j:d]))

# Step k (Iterating through step 2 until convergence)
k=2;
while(max(abs(beta.EM[k,]-beta.EM[k-1,]))>error)
{
if(k>=n.iter) break;
k=k+1;
W=fun.W(beta.EM[k-1,],lambda.EM[k-1,],y.unique,y,z,u,v)
w=as.vector(t(W));
```

```
w.plus.j=apply(W,2,"sum")
w.i.plus=apply(W,1,"sum")
beta.EM[k,]=coxph(new.formula,data=new.data,weights=w,
subset=which(w>0))$coefficients
temp=t(W)%*%exp(z%*%beta.EM[k,]) # column j of W times exp(z*beta)
lambda.EM[k,]=sapply(1:d, function(j) w.plus.j[j]/sum(temp[j:d]))
if(k>n.iter) break;
#print(k)
}
beta.hat.EM=beta.EM[k,]
lambda.hat.EM=lambda.EM[k,]

# Indicator of whether the maximum number of iterations was reached
max.iter_reached=0; if(k>=n.iter) max.iter_reached=1;

if(k<n.iter) return(list(beta.hat=beta.hat.EM,lambda.hat=lambda.hat.EM,
n.iterations=k,max.iter_reached=max.iter_reached))
if(k>=n.iter) return(list(max.iter_reached=max.iter_reached))
}
```

APPENDIX E

R FUNCTION FOR NONPARAMETRIC WEIGHTED COX REGRESSION COEFFICIENT ESTIMATOR UNDER DOUBLE TRUNCATION

```
coxDT = function(formula,L,R,data=list(),subset,time.var=FALSE,subject=NULL,
B.SE.np=200,CI.boot=FALSE,B.CI.boot=2000,pvalue.boot=FALSE,
B.pvalue.boot=200,print.weights=FALSE,error=10^-6,n.iter=10000)
{
set.seed(1312018)
data=data[subset,]
# extracting outcomes and covariates
mf = model.frame(formula=formula,data=data)
X=model.matrix(attr(mf,"terms"),data=mf)[,-1]
p=1; n=length(X);
# number of predictors and observations
if(length(dim(X))>0) {p=dim(X)[2]; n=dim(X)[1]}

Y=as.numeric(model.response(mf))[1:n];



# extracting truncation times
L=deparse(substitute(L)); R=deparse(substitute(R));
formula.temp=paste(L,R,sep="~")
mf.temp=model.frame(formula=formula.temp,data=data)
obs.data=sapply(rownames(mf),rownames(mf.temp),FUN=function(x,y) which(x==y))
L=mf.temp[obs.data,1]; R=mf.temp[obs.data,2];

# estimating individual nonparametric probabilities of being observed
P.obs.y.np=cdfDT(Y,L,R,error,n.iter,display=FALSE)$P.K
# weights
```

```
weights.np=1/P.obs.y.np
# estimating nonparametric probability of observing random subject
P.obs.NP=n*(sum(1/P.obs.y.np))^(-1)

# creating new data set which incorporate weights
data.new=data.frame(data,weights.np)
# computing estimates of nonparametric weighted regression coefficient estimator
beta.np=coxph(formula,data=data.new,weights=weights.np)$coefficients

# computing bootstrapped standard errors
# first, we import the vector of subject id's
# for bootstrapping data with time-varying coefficients
if(time.var==TRUE)
{
subject=deparse(substitute(subject))
formula.temp2=paste(subject,subject,sep="~");
subjects=model.response(model.frame(formula.temp2,data=data));
n.subject=length(unique(subjects));
}

B=B.SE.np
if(CI.boot==TRUE) B=max(B.SE.np,B.CI.boot)

beta.boot.np=matrix(0,nrow=B,ncol=p)
for(b in 1:B) {
repeat{ # creating repeat loop in case Shen algorithm fails
if(time.var==FALSE) {temp.sample=sort(sample(n,replace=TRUE))};
if(time.var==TRUE) {
temp.obs=sort(sample(n.subject,replace=TRUE))
temp.sample=unlist(sapply(temp.obs,subjects,FUN=function(x,y) which(x==y)))
}
```

```
Y.temp=Y[temp.sample]; L.temp=L[temp.sample]; R.temp=R[temp.sample];
if(p==1) X.temp=X[temp.sample];
if(p>=2) X.temp=X[temp.sample,];

P.obs.NP.temp=cdfDT(Y.temp,L.temp,R.temp,error,n.iter,display=FALSE)$P.K

if(length(which(P.obs.NP.temp<.01))==0) {break}
} # ending repeat loop

# non-parametric weights (Shen) for cox regression
weights.np.temp=1/P.obs.NP.temp

# updating data set to include bootstrapped observations
data.temp=data.frame(data[temp.sample,],weights.np.temp)

# computing estimates of nonparametric weighted estimator
beta.boot.np[b,]=coxph(formula,data=data.temp,
weights=weights.np.temp)$coefficients
}
# standard error
se.beta.np=apply(beta.boot.np,2,sd);

# If bootstrap not requested, return normal confidence intervals
if(CI.boot==FALSE) {
CI.lower=beta.np-1.96*se.beta.np; CI.upper=beta.np+1.96*se.beta.np;
CI.beta.np=round(cbind(CI.lower,CI.upper),3)
}
# If bootstrap not requested, return p-values based on normality assumption
if(pvalue.boot==FALSE)  {
Test.statistic=(beta.np/se.beta.np)^2;
p.value=round(2*(1-pnorm(abs(beta.np/se.beta.np))),4)
```

```
}

# computing 95% confidence intervals
if(CI.boot==TRUE)
{
CI.beta.np=matrix(0,nrow=p,ncol=2)
for(k in 1:p) CI.beta.np[k,]=round(c(quantile(beta.boot.np[,k],seq(0,1,0.025))[2],
quantile(beta.boot.np[,k],seq(0,1,0.025))[40]),3)
}




if(pvalue.boot==TRUE)
{
B1=B2=B.pvalue.boot
beta.boot.np1=matrix(0,nrow=B1,ncol=p); beta.boot.np2=matrix(0,nrow=B2,ncol=p);
beta.boot.np.sd1=matrix(0,nrow=B1,ncol=p)
for(b1 in 1:B1) {
repeat{ # creating repeat loop in case Shen algorithm fails
if(time.var==FALSE) {temp.sample1=sort(sample(n,replace=TRUE))};
if(time.var==TRUE) {
temp.obs1=sort(sample(n.subject,replace=TRUE))
temp.sample1=unlist(sapply(temp.obs1,subjects,FUN=function(x,y) which(x==y)))
}
Y.temp1=Y[temp.sample1]; L.temp1=L[temp.sample1]; R.temp1=R[temp.sample1];
if(p==1) X.temp1=X[temp.sample1,];
if(p>=2) X.temp1=X[temp.sample1,];

P.obs.NP.temp1=cdfDT(Y.temp1,L.temp1,R.temp1,error,n.iter,display=FALSE)$P.K

if(length(which(P.obs.NP.temp1<.01))==0) {break}
```

```
} # ending repeat loop

weights.np.temp1=1/P.obs.NP.temp1

# updating data set to include bootstrapped observations
data.temp1=data.frame(data[temp.sample1,],weights.np.temp1)
# computing estimates of nonparametric weighted estimator
beta.boot.np1[b1,]=coxph(formula,data=data.temp1,
weights=weights.np.temp1)$coefficients


# The loop below is to compute the standard error of each bootstrap estimate
for(b2 in 1:B2) {
repeat{ # creating repeat loop in case Shen algorithm fails
if(time.var==FALSE) {temp.sample2=sort(sample(temp.sample1,replace=TRUE))};
if(time.var==TRUE) {
temp.obs2=sort(sample(temp.obs1,replace=TRUE))
temp.sample2=unlist(sapply(temp.obs2,subjects,FUN=function(x,y) which(x==y)))
}
Y.temp2=Y[temp.sample2]; L.temp2=L[temp.sample2]; R.temp2=R[temp.sample2];
if(p==1) X.temp2=X[temp.sample2,];
if(p>=2) X.temp2=X[temp.sample2,];

P.obs.NP.temp2=cdfDT(Y.temp2,L.temp2,R.temp2,error,n.iter,display=FALSE)$P.K

if(length(which(P.obs.NP.temp2<.01))==0) {break}
} # ending repeat loop

# non-parametric weights (Shen) for cox regression
weights.np.temp2=1/P.obs.NP.temp2
```

```
# updating data set to include bootstrapped observations
data.temp2=data.frame(data[temp.sample2,],weights.np.temp2)
# computing estimates of nonparametric weighted estimator
beta.boot.np2[b2,]=coxph(formula,data=data.temp2,
weights=weights.np.temp2)$coefficients
}
beta.boot.np.sd1[b1,]=apply(beta.boot.np2,2,"sd")
}

# computing test statistics, bootstrapped test statistics, and p-values
test_statistic.beta.np=numeric(p)
test_statistic.beta.np.boot=matrix(0,nrow=B1,ncol=p)
p.value=numeric(p)
for(k in 1:p) {
test_statistic.beta.np[k]=(beta.np[k]/se.beta.np[k])^2
test_statistic.beta.np.boot[,k]=
((beta.boot.np1[,k]-beta.np[k])/beta.boot.np.sd1[,k])^2
p.value[k]=round(length(which(test_statistic.beta.np.boot[,k]>
test_statistic.beta.np[k]))/B1,4)
}
Test.statistic=test_statistic.beta.np
}

beta.np=round(beta.np,4);
se.beta.np=round(se.beta.np,4);
Test.statistic=round(Test.statistic,2)

results.beta=cbind(beta.np,se.beta.np,CI.beta.np,Test.statistic,p.value)
rownames(results.beta)=colnames(X); colnames(results.beta)=
c("Estimate","SE","CI.lower","CI.upper","Wald statistic","p-value")
```

```
weights="print option not requested";
if(print.weights==TRUE) weights=weights.np;


if((CI.boot==TRUE)&(pvalue.boot==FALSE)) return(list(results.beta=results.beta,
CI="Bootstrap",p.value="Normal approximation",weights=weights));
if((CI.boot==FALSE)&(pvalue.boot==TRUE)) return(list(results.beta=results.beta,
CI="Normal approximation",p.value="Bootstrap",weights=weights));
if((CI.boot==TRUE)&(pvalue.boot==TRUE))  return(list(results.beta=results.beta,
CI="Bootstrap",p.value="Bootstrap",weights=weights));
if((CI.boot==FALSE)&(pvalue.boot==FALSE))  return(list(results.beta=results.beta,
CI="Normal approximation",p.value="Normal approximation",weights=weights));
}
```

APPENDIX F

R FUNCTION FOR NONPARAMETRIC ESTIMATION OF SURVIVAL DISTRIBUTION FUNCTION AND SELECTION PROBABILITIES UNDER DOUBLE TRUNCATION

```
cdfDT=function(y,l,r,error=1e-6,n.iter=10000,boot=FALSE,B.boot=200,joint=FALSE,
plot.cdf=FALSE,plot.joint=FALSE,display=TRUE)
{
if(joint==FALSE) plot.joint=FALSE;

# removing rows from data frame with missing data
temp.data=data.frame(y,l,r);
missing.data=unique(which((is.na(temp.data[,1])==TRUE)|
(is.na(temp.data[,2])==TRUE)|(is.na(temp.data[,3])==TRUE)))
if(length(missing.data)==0) {
y=temp.data[,1]; u=temp.data[,2]; v=temp.data[,3]};
if(length(missing.data)>=1) {temp.data2=temp.data[-missing.data,];
y=temp.data2[,1]; u=temp.data2[,2]; v=temp.data2[,3];}
n=length(y);

fun.U=function(y,u) I(y>=u)*1;
fun.V=function(y,v) I(y<=v)*1;

fun.DT=function(y,u,v)
{
n=length(y);
temp.U=sapply(y,u,FUN="fun.U")
temp.V=sapply(y,v,FUN="fun.V")

J=temp.U*temp.V
```

```
K=matrix(0,nrow=n.iter+1,ncol=n); F=matrix(0,nrow=n.iter+1,ncol=n);
f=matrix(0,nrow=n.iter+1,ncol=n); k=matrix(0,nrow=n.iter+1,ncol=n);

# Step 0
F.0=apply(J,2,"sum")/n

# Step 1
k[1,]=(sum(1/F.0)^-1)/F.0
K[1,]=apply(k[1,]*J,2,"sum")
f[1,]=(sum(1/K[1,])^-1)/K[1,]
F[1,]=apply(f[1,]*t(J),2,"sum")

# Step 2
k[2,]=(sum(1/F[1,])^-1)/F[1,]
K[2,]=apply(k[2,]*J,2,"sum")
f[2,]=(sum(1/K[2,])^-1)/K[2,]
F[2,]=apply(f[2,]*t(J),2,"sum")

# Step s - iterating through step 2
s=2;
while(sum(abs(f[s,]-f[s-1,]))>error)
{
s=s+1;

# Step s.1
k[s,]=(sum(1/F[s-1,])^-1)/F[s-1,]
K[s,]=apply(k[s,]*J,2,"sum")

# Step s.2
f[s,]=(sum(1/K[s,])^-1)/K[s,]
F[s,]=apply(f[s,]*t(J),2,"sum")
```

```
if(s>n.iter) break;
}
# P.K = P(L<T_i<R); P.F = P(L_i<T<R_i)
P.K=K[s,]; P.F=F[s,]

n.unique.y=length(unique(y))
distF=numeric(n.unique.y)

# computing CDF estimate at unique (ordered) survival times
for(i in 1:n.unique.y) {
distF[i]=round(sum(1/P.K)^-1*sum(I(y<=sort(unique(y))[i])/P.K),4)}

# f = density of observed survival times
# k = joint density of observed truncation times
f=round(f[s,],4); k=round(k[s,],4);

max.iter_reached=0; if(s>=n.iter) max.iter_reached=1;
return(list(f=f,k=k,P.K=P.K,P.F=P.F,distF=distF,n.iterations=s,
max.iter_reached=max.iter_reached))
}

out=fun.DT(y,u,v);
P.K=out$P.K; P.F=out$P.F; distF=out$distF; f=out$f;
k=out$k; n.iterations=out$n.iterations;
max.iter_reached=out$max.iter_reached;
#########################################################################
# NPMLE of truncation distribution
if(joint==TRUE)
{
# NPMLE of joint truncation distribution
```

```
unique.u=sort(unique(u)); unique.v=sort(unique(v));
Joint.UV=matrix(0,nrow=length(unique.u),ncol=length(unique.v));

for(a in 1:length(unique.u)) {
for(b in 1:length(unique.v)) {
Joint.UV[a,b]=(sum(1/(P.F)))^-1*
sum(I(u<=unique.u[a])*I(v<=unique.v[b])/(P.F))
}
}
# NPMLE of marginal truncation distributions
Q.U=Joint.UV[,length(unique.v)]; R.V=Joint.UV[length(unique.u),]
for(a in 1:length(unique.u)) {
for(b in 1:length(unique.v)) {
Joint.UV[a,b]=(sum(1/P.F))^-1*
sum(I(u<=unique.u[a])*I(v<=unique.v[b])/P.F)
}
}
Q.U=round(Joint.UV[,length(unique.v)],4);
R.V=round(Joint.UV[length(unique.u),],4)
Joint.UV=round(Joint.UV,4)
}

#########################################################################

# computing standard errors and confidence interavls for survival CDF
if(boot==TRUE)
{

temp.data=data.frame(y,u,v);
temp.data=temp.data[order(temp.data$y),]
y.sort=temp.data$y; u.sort=temp.data$u; v.sort=temp.data$v;
```

```
y.unique=sort(unique(y));
n.unique.y=length(unique(y));

F.boot=matrix(-1,nrow=B.boot,ncol=n.unique.y)
for(b in 1:B.boot) {
repeat{ # creating repeat loop in case program does not converge
temp.sample=sort(sample(n,replace=TRUE))
# Creating new data set based off of bootstrapped samples
y.boot=y.sort[temp.sample]; u.boot=u.sort[temp.sample];
v.boot=v.sort[temp.sample];
y.boot.unique=(unique(y.boot))

#########################################################################
#########################################################################
# Survival distribution is at observed survival times, need to
# impute survival function for survival times not selected
# by the bootstrap procedure
x1=which(is.element(y.unique,y.boot.unique)==FALSE);
x2=which(is.element(y.unique,y.boot.unique)==TRUE);
x3=x1[which((x1>min(x2))&(x1<max(x2)))];
x3min=x1[which(x1<min(x2))]; x3max=x1[which(x1>max(x2))];

out.boot=fun.DT(y.boot,u.boot,v.boot)

if(out.boot$max.iter_reached==0) {break}
} # ending repeat loop

F.boot[b,x2]=out.boot$distF

if(length(x1)>0)
{
```

```
if(length(x3min)>0) F.boot[b,x3min]=0;
if(length(x3max)>0) F.boot[b,x3max]=1;
if(length(x3)>0)
{
while(min(F.boot[b,x3])<0)
{
F.boot[b,x3]=F.boot[b,x3-1]
}
}
}
########################################################################
########################################################################
}
# computing standard error of bootstrapped samples
sigma=apply(F.boot,2,"sd")

# computing confidence intervals based on normality assumption
CI.lower.F=distF-1.96*sigma; CI.upper.F=distF+1.96*sigma;
}

# getting density at unique survival times
f=f[which(duplicated(y)==FALSE)];
f=f[order(unique(y))];

# printing plots
if(display==TRUE)
{
if(max.iter_reached==0)
{
cat("number of iterations", n.iterations, "\n")
summary <- cbind(event.time = sort(unique(y)), n.event = table(sort(y)),
```

```
F = distF, Survival = 1-distF)
colnames(summary) <- c("time", "n.event",  "cumulative.df", "survival")
rownames(summary) <- rep("", times = length(unique(y)))
print(summary, digits = 4, justify = "left")
cat("number of observations", n, "\n")
}
if(max.iter_reached==1) print("Maximum number of iterations reached.
Program did not converge")
}

if(plot.cdf==TRUE)
{
dev.new()
par(mfrow=c(1,2))
plot(distF~sort(unique(y)),ylim=c(0,1),xlab="event time",ylab="",
main="Cumulative distribution function")
lines(distF~sort(unique(y)))
if(boot==TRUE)
{
lines(CI.lower.F~sort(unique(y)),lty=2)
lines(CI.upper.F~sort(unique(y)),lty=2)
}
plot((1-distF)~sort(unique(y)),ylim=c(0,1),xlab="event time",ylab="",
main="Survival function")
lines((1-distF)~sort(unique(y)))
if(boot==TRUE)
{
lines((1-CI.lower.F)~sort(unique(y)),lty=2)
lines((1-CI.upper.F)~sort(unique(y)),lty=2)
}
}
```

```
if(plot.joint==TRUE)
{
dev.new()
par(mfrow=c(1,2))
plot(Q.U~sort(unique(u)),ylim=c(0,1),xlab="left truncation time",ylab="",
main="Marginal cdf (left)")
lines(Q.U~sort(unique(u)))
plot(R.V~sort(unique(v)),ylim=c(0,1),xlab="right truncation time",ylab="",
main="Marginal cdf (right)")
lines(R.V~sort(unique(v)))

dev.new()
persp(sort(unique(u)),sort(unique(v)),Joint.UV,
theta=30,expand=0.75,col="lightblue",
main="Joint truncation distribution",
xlab="left truncation time",ylab="right truncation time",zlab="")
}

if(boot==TRUE)
{

if(joint==TRUE) return(invisible(list(time=round(sort(unique(y)),4),
n.event = table(sort(y)), F = distF, Survival = 1-distF,sigma.F=sigma,
CI.lower.F=CI.lower.F,CI.upper.F=CI.upper.F,P.K=P.K,
Joint.LR=Joint.UV,Marginal.L=Q.U,Marginal.R=R.V,n.iterations
=n.iterations,max.iter_reached=max.iter_reached)));
if(joint==FALSE) return(invisible(list(time=round(sort(unique(y)),4),
n.event = table(sort(y)), F = distF, Survival = 1-distF, F=distF,
sigma.F=sigma,CI.lower.F=CI.lower.F,CI.upper.F=CI.upper.F,P.K=P.K,
n.iterations=n.iterations,max.iter_reached=max.iter_reached)));
```

```
}

if(boot==FALSE)
{
if(joint==TRUE) return(invisible(list(time=round(sort(unique(y)),4),
n.event = table(sort(y)), F = distF, Survival =1-distF,P.K=P.K,
Joint.LR=Joint.UV,Marginal.L=Q.U,Marginal.R=R.V,
n.iterations=n.iterations,max.iter_reached=max.iter_reached)));
if(joint==FALSE) return(invisible(list(time=round(sort(unique(y)),4),
n.event = table(sort(y)), F = distF, Survival = 1-distF,P.K=P.K,
n.iterations=n.iterations,max.iter_reached=max.iter_reached)));
}
}
```

BIBLIOGRAPHY

Andel, R, Silverstein, M, and Kareholt, I (2014). The role of midlife occupational complexity and leisure activity in late-life cognition. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 70.2, 314–321.

Andersen, P, Borgan, O, Gill, R, and Keiding, N (1997). *Statistical models based on counting processes*. English. OCLC: 968632889. New York: Springer. ISBN: 978-1-4612-4348-9.

Beach, TG, Monsell, SE, Phillips, LE, and Kukull, W (2012). Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer's Disease Centers, 20052010. *J Neuropathol Exp Neurol* 71.4, 266–273. DOI: 10.1097/NEN.0b013e31824b211b.

Bilker, WB and Wang, MC (1996). A semiparametric extension of the Mann-Whitney test for randomly truncated data. eng. *Biometrics* 52.1, 10–20. ISSN: 0006-341X.

Binder, D (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* 79, 139–147.

Correa Ribeiro, PC, Lopes, CS, and Loureno, RA (2013). Complexity of lifetime occupation and cognitive performance in old age. *Occupational medicine* 63.8, 556–562.

Cox, D (1972). Regression Models and Life-Tables. *JRSSB* 34.2, 187–220.

Cox, D (1975). Partial likelihood. en. *Biometrika* 62.2, 269–276. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/62.2.269.

Efron, B and Petrosian, V (1999). Nonparametric Methods for Doubly Truncated Data. *Journal of the American Statistical Association* 94.447, 824–834. ISSN: 0162-1459. DOI: 10.1080/01621459.1999.10474187.

Efron, B and Tibshirani, RJ (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Enroth, L, Raitanen, J, Hervonen, A, Nosraty, L, and Jylh, M (2014). Is socioeconomic status a predictor of mortality in nonagenarians? The vitality 90+ study. *Age and ageing* 44.1, 123–129.

Grossman, M and Irwin, DJ (2016). The mental status examination in patients with suspected dementia. *CONTINUUM: Lifelong Learning in Neurology* 22.2 Dementia, 385.

Ientile, L, De Pasquale, R, Monacelli, F, Odetti, P, Traverso, N, Cammarata, S, Tabaton, M, and Dijk, B (2013). Survival rate in patients affected by dementia followed by memory clinics (UVA) in Italy. *Journal of Alzheimer's Disease* 36.2, 303–309.

Kalbfleisch, JD and Lawless, JF (1991). Regression models for right truncated data with application to AIDS incubation times and reporting lags. *Statistica Sinica* 1.1, 19–32.

Klein, JP and Moeschberger, ML (2003). *Survival analysis: techniques for censored and truncated data*. 2nd ed. Statistics for biology and health. New York: Springer. ISBN: 978-0-387-95399-1.

Lai, TL and Ying, Z (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* 19.2, 531–556.

Lin, D (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* 87.1, 37–47.

Mandel, M, Una-Alvarez, J de, Simon, DK, and Betensky, RA (2017). Inverse probability weighted Cox regression for doubly truncated data: Cox Regression for Doubly Truncated Data. en. *Biometrics*. ISSN: 0006341X. DOI: 10.1111/biom.12771.

Martin, EC and Betensky, RA (2005). Testing Quasi-Independence of Failure and Truncation Times via Conditional Kendall's Tau. en. *Journal of the American Statistical Association* 100.470, 484–492. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214504000001538.

Massimo, L, Zee, J, Xie, SX, McMillan, CT, Rascovsky, K, Irwin, DJ, Kolanowski, A, and Grossman, M (2015). Occupational attainment influences survival in autopsy-confirmed frontotemporal degeneration. *Neurology* 84.20, 2070–2075. DOI: 10.1212/WNL.0000000000001595.

Massimo, L, Xie, SX, Rennert, L, Fick, DM, Halpin, A, Placek, K, Williams, A, Rascovsky, K, Irwin, DJ, Grossman, M, and McMillan, CT (2018). Occupational attainment influences longitudinal decline in behavioral variant frontotemporal degeneration. en. *Brain Imaging and Behavior*, 1–9. DOI: 10.1007/s11682-018-9852-x.

Meng, X and DArcy, C (2012). Education and Dementia in the Context of the Cognitive Reserve Hypothesis: A Systematic Review with Meta-Analyses and Qualitative Analyses. en. *PLoS ONE* 7.6. Ed. by J Laks, e38268. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0038268.

Moreira, C, Ivarez, J de na, and Meira-Machado, L (2016). Nonparametric regression with doubly truncated data. *Computational Statistics and Data Analysis* 93.January 1987, 294–307. ISSN: 01679473. DOI: 10.1016/j.csda.2014.03.017.

Moreira, C, Ivarez, J de na, and Van Keilegom, I (2014). Goodness-of-fit tests for a semiparametric model under random double truncation. *Computational Statistics* 29.1, 1365–1379.

Moreira, C and Una-Alvarez, J de (2010). A semiparametric estimator of survival for doubly truncated data. en. *Statistics in Medicine* 29.30, 3147–3159. ISSN: 02776715. DOI: 10.1002/sim.3938.

Moreira, C, Una-Alvarez, J de, and Crujeiras, R (2010). **DTDA** : An *R* Package to Analyze Randomly Truncated Data. en. *Journal of Statistical Software* 37.7. ISSN: 1548-7660. DOI: 10.18637/jss.v037.i07.

Murphy, SA (1994). Consistency in a Proportional Hazards Model Incorporating a Random Effect. *The Annals of Statistics* 22.2, 712–731. ISSN: 0090-5364.

Pan, Q and Schaubel, DE (2008). Proportional hazards models based on biased samples and estimated selection probabilities. *Canadian Journal of Statistics* 36.1, 111–127. ISSN: 03195724. DOI: 10.1002/cjs.5550360111.

Paradise, M, Cooper, C, and Livingston, G (2009). Systematic review of the effect of education on survival in Alzheimer's disease. en. *International Psychogeriatrics* 21.01, 25. ISSN: 1041-6102, 1741-203X. DOI: 10.1017/S1041610208008053.

Qin, J, Ning, J, Liu, H, and Shen, Y (2011). Maximum Likelihood Estimations and EM Algorithms With Length-Biased Data. en. *Journal of the American Statistical Association* 106.496, 1434–1449. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/jasa.2011.tm10156.

Rascovsky, K, Salmon, DP, Lipton, AM, Leverenz, JB, DeCarli, C, Jagust, WJ, Clark, CM, Mendez, MF, Tang-Wai, DF, Graff-Radford, NR, and Galasko, D (2005). Rate of progression differs in frontotemporal dementia and Alzheimer disease. eng. *Neurology* 65.3, 397–403. ISSN: 1526-632X. DOI: 10.1212/01.wnl.0000171343.43314.6e.

Rennert, L and Xie, SX (2017). Cox regression model with doubly truncated data. eng. *Biometrics*. ISSN: 1541-0420. DOI: 10.1111/biom.12809.

Sanchez, JL, Torrellas, C, Martn, J, and Barrera, I (2011). Study of sociodemographic variables linked to lifestyle and their possible influence on cognitive reserve. *Journal of Clinical and Experimental Neuropsychology* 33.8, 874–891. ISSN: 1380-3395. DOI: 10.1080/13803395.2011.567976.

Seaman, SR and White, IR (2013). Review of inverse probability weighting for dealing with missing data. *Stat Methods in Medical Research* 22.3, 278–295. ISSN: 1477-0334. DOI: 10.1177/0962280210395740.

Shen, PS (2010a). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics* 62.5, 835–853. ISSN: 15729052. DOI: 10.1007/s10463-008-0192-2.

Shen, PS (2010b). Semiparametric analysis of doubly truncated data. *Communications in Statistics-Theory and Methods* 39.1, 3178–3190. ISSN: 0361-0926. DOI: 10.1007/s10463-008-0192-2.

Shen, PS (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics* 28.2, 581–596. ISSN: 09434062. DOI: 10.1007/s00180-012-0318-0.

Shen, P-s and Liu, Y (2017). Pseudo maximum likelihood estimation for the Cox model with doubly truncated data. en. *Statistical Papers*. ISSN: 0932-5026, 1613-9798. DOI: 10.1007/s00362-016-0870-8.

Stern, Y, Albert, S, Tang, MX, and Tsai, WY (1999). Rate of memory decline in AD is related to education and occupation: cognitive reserve? eng. *Neurology* 53.9, 1942–1947. ISSN: 0028-3878.

Stern, Y (2012). Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology* 11.11, 1006–1012. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(12)70191-6.

Stern, Y, Tang, MX, Denaro, J, and Mayeux, R (1995). Increased risk of mortality in alzheimer's disease patients with more advanced educational and occupational attainment. *Annals of Neurology* 37.5, 590–595.

Turnbull, BW (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* 38.1, 290–295.

Vaart, AWvd and Wellner, JA (2000). *Weak convergence and empirical processes: with applications to statistics*. English. OCLC: 45749647. New York: Springer. ISBN: 978-0-387-94640-5 978-1-4757-2547-6.

Valenzuela, MJ and Sachdev, P (2007). Assessment of complex mental activity across the lifespan: development of the Lifetime of Experiences Questionnaire (LEQ). *Psychological Medicine* 37.1, 1015–1025.

Wang, M-C (1996). Hazards regression analysis for length-biased data. *Biometrika* 83.2, 343–354.

Woodroofe, M (1985). Estimating a distribution function with truncated data. *Annals of Statistics* 13.1, 163–177.