Analysis of Left-Truncated Right-Censored Survival Data with Uncertainty of Onset Times

James Hugh McVittie

Department of Mathematics and Statistics,
McGill University, Montreal
July, 2017

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Science

Abstract

Survival analysis is a field of statistics in which the distribution of the time interval between two well defined events is studied. The distinguishing feature of survival analysis is that the time of the initial event and the time of the final event may be truncated and/or censored. By appropriately adjusting for such phenomena, statistical properties of the underlying distribution of the time interval can be obtained. In certain settings, the true time of the initial event may be completely unknown or censored. For example, in a study of a cohort of subjects with prevalent dementia, the initial date of dementia onset reported by a caregiver or family member can be assumed to be reported with error. The main purpose of this thesis is to review how uncertainty in the time of the initial event impacts the classical statistical inference techniques of survival analysis. We suggest new failure time density estimators under different measurement error models.

Résumé

L'analyse de survie est un domaine de la statistique qui s'intéresse à la distribution du temps de deux événements définis. Une des caractéristiques est que le temps du premier événement et le temps de l'événement final peuvent être tronqués et/ou censurés. En utilisant les méthodes pour ajuster pour ces phénomènes, les propriétés statistiques de la vraie distribution de l'intervalle de temps peuvent être obtenues. Dans certains contextes, le temps du premier événement peut être complètement inconnu ou censuré. Par exemple, dans une étude d'une cohorte prévalente de démence, le première temps du début de la démence qui est identifé par un soignant ou un membre de la famille peut être supposé obtenu avec erreur. L'objectif principal de ce mémoire est de réviser comment l'incertitude avec le temps du premier événement a un impact sur l'inférence statistique classique d'analyse de survie. Nous suggérons de nouveaux estimateurs sous l'hypothèse des modèles d'erreur de mesure.

Acknowledgements

I would like to thank my supervisors Professor David Stephens and Professor David Wolfson for introducing me to the field of survival analysis and for always providing meaningful direction for the research of this thesis. I would also like to thank Professor Russell Steele in his role as graduate program director for answering my many questions about the university both academic and administrative. Additionally, I would like to thank Alexandre Piché and Stephanie Long for reminding me on many occassions that it is just as important to relax over a glass of wine as it is to spend the night working through statistical proofs. Finally, I would like to especially thank my parents and my brother for their constant support and encouragement through every hurdle of this degree. The research of this degree was financially supported through the NSERC CGSM award.

Preface

Survival analysis is a field of statistics in which the distribution of the time interval between two well defined events is studied. The distinguishing feature of survival analysis is that the time of the initial event and the time of the final event may be truncated and/or censored. By appropriately adjusting for such phenomena, statistical properties of the underlying distribution of the time interval can be obtained. In certain settings, the true time of the initial event may be completely unknown or censored. For example, in a study of a cohort of subjects with prevalent dementia, the initial date of dementia onset reported by a caregiver or family member can be assumed to be reported with error. The main purpose of this thesis is to review how uncertainty in the time of the initial event impacts the classical statistical inference techniques of survival analysis. In Chapter 4, we extend the previously derived likelihood function and we suggest a deconvolution estimator to include uncertainty in the onset distribution under different measurement error models.

Contents

Chapter 1

Introduction

The Canadian Study of Health and Aging (CSHA-1) began in 1991 when approximately 10,000 Canadians over the age of 65 in community or institutional settings across nine provinces were screened for various forms of dementia. The objective of the CSHA-1 was "to determine the prevalence of dementia and its subtypes by sex and age group for five regions in Canada" [30]. In 1996, the CSHA entered its second phase (CSHA-2). A date and cause of death were recorded for individuals that died between 1991 and 1996. In addition, it was noted which participants who had screened positive in 1991 were still alive in 1996. Thus, the CSHA included two types of cohort: 1) An incident cohort consisting of those that were deemed disease free in CSHA-1 and then followed until CSHA-2 unless censored and 2) A prevalent cohort with follow-up consisting of those prevalent for disease in CSHA-1 who were followed until censoring or death in 1996.

A sub-study using the CSHA-1 and CSHA-2 data was conducted to estimate survival from onset of dementia and to determine predictors of survival with dementia [45]. For the sub-study, the importance of the ascertainment of the dates of onset, as opposed to diagnosis, of dementia (recorded from the CSHA databases) in producing time-to-

failure/censoring data for the survival analysis, is clear. One complication was the speculative nature of the reported onset times. For example, the dates of onset of Alzheimer's disease were determined among participants with prevalent disease through the recollections of their caregivers. While there is inherent uncertainty in the recollections of the caregivers, the very meaning of "date of onset" in the setting of dementia is ambiguous. The recalled date of onset can be used as the true date of onset without further interpretation or can be representative, with error, of the true date of onset of the clinical symptoms of the disease. It can also be interpreted as approximating the true date of physiological onset of the disease process.

This thesis is concerned with the survival analysis modelling of failure time data with measurement error in the reported time of the initial event, allowing for possible censoring of the failure event time. Chapter 2 introduces the notation of the thesis and elaborates on the details of incident and prevalent cohort studies. Chapter 3 discusses how uncertainty in the date of the initial event is incorporated into survival analysis models through various measurement error models and censoring types. Chapter 4 presents statistical inference techniques that adjust for the uncertainty in the initial event date and Chapter 5 presents simulation results of the methods discussed in Chapter 4. Chapter 6 concludes with some suggestions for further research.

Chapter 2

Notation and Terminology

2.1 Notation

Let T be a non-negative random variable representing the failure time and let $S(\cdot)$ denote the survivor function of T. That is,

$$S(t) = \begin{cases} 1 - F(t) = P(T > t), & \text{for } 0 \leq t < \infty \\ 1, & \text{for } t < 0 \end{cases} \tag{2.1}$$

where $F(t) = P(T \leq t)$ is the cumulative distribution function of T. Let $f(t)$ denote the *probability density function* or *probability mass function*, for the continuous or the discrete cases respectively, of the random variable T. For $-\infty < t < \infty$, let

$$\lambda(t) = \lim_{\Delta t \to 0^+} P(t \leq T < t + \Delta t | T \geq t)/\Delta t. \tag{2.2}$$

be the hazard function of T. Although the functions above are written in non-parametric form, they can also be defined parametrically, in which case they are denoted, respectively, by $S(t;\boldsymbol{\theta})$, $f(t;\boldsymbol{\theta})$ and $\lambda(t;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a p-dimensional vector of parameters [20] [23].

2.2 Censoring and Truncation

The feature that distinguishes survival analysis is that the failure times T_i, in the sample, may be incompletely observed. The two main ways in which this occurs is through censoring and/or truncation. Any statistical inference must take these features into account. Due to their role in later chapters, we define the various types of censoring and truncation that arise in survival analysis.

Definition 2.2.1. ***(Right/Left-Censoring)*** *T is* right-censored *by the non-negative random variable C if $C < T$. The observed quantity is then* $\min(T, C) = C$. *Equivalently, it is* left-censored *if $C > T$ and the observed quantity is* $\max(T, C) = C$.

For example, in a study with staggered entry, which terminates on some fixed date, all failure times that extend beyond this date are right-censored.

Definition 2.2.2. ***(Interval-Censoring)*** *T is* interval-censored *by the non-negative random variables C_1,C_2 with $C_1 < C_2$ if $C_1 < T < C_2$. The observed quantity is then the pair (C_1, C_2).*

For example, if the random variable T represents the onset time of HIV, and C_1 and C_2 represent the calendar times of the observed negative and positive HIV tests respectively, then T is interval-censored by (C_1, C_2) [10].

Definition 2.2.3. ***(Random Censoring)*** *T is* randomly censored *by C if the non-negative random variables T and C are independent.*

Random censoring is an assumption that, if appropriate, leads to simplified (although not necessarily simple) statistical analyses. Let the so-called censoring indicator, $\delta_i = 1$ if T_i

is censored by C_i and 0 otherwise. We denote the ith observed sample point by (X_i, δ_i). For an extensive discussion of censoring in survival analysis, see [27].

Definition 2.2.4. *(**Right/Left/Interval-Truncation**) T is* right-truncated *by the random variable D if T is observable only if $T < D$. T is* left-truncated *by D if T is observable only if $T > D$. T is* interval-truncated *by the random variables D_1 and D_2 $(D_1 < D_2)$, if T is observable only if $T < D_1$ or $T > D_2$.*

Note that if the random variable T is left/right-censored by the random variable C, then C is observed and T is unobserved. In contrast, if T is left/right-truncated by the random variable D, then both T and D are unobserved. Analogously, if T is interval-censored by the pair (C_1, C_2), then the pair is observed and T is unobserved and if T is interval-truncated by the pair (D_1, D_2), then both the pair and T are unobserved.

2.3 Incident Cohort Study

An *incident cohort study* can be used to estimate the incidence rate within a population or to estimate the survivor function from onset of a given disease or condition. In the latter case, two events must be defined: an initiating event, and a subsequent event, failure (onset and death, say). A disease-free cohort is followed for the occurrence of these two events. All subjects who do not experience onset during follow-up play no role in the estimation of survival from onset. The dates of those who experience onset (an initiating event) during follow-up are monitored for the occurrence of failure/censoring; these dates are also recorded. Figure 2.1 is a graphical representation of incident cohort data with various failure/censoring times.

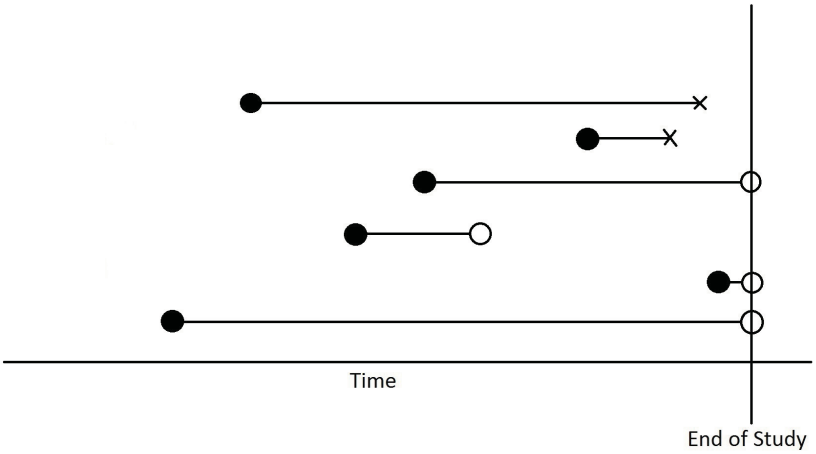

Figure 2.1 – *Graphical representation of incident cohort data. The filled circles, crosses and open circles denote, respectively, disease onset, failure and censoring.*

It is usually assumed that the incidence process of onset times is independent of the corresponding failure/censoring times [23]. Consequently, without loss of generality, the calendar dates of onset may be translated to a common time origin. The analysis is based on the time intervals from this origin to failure/censoring. Let T_i and C_i be the respective translated calendar dates of failure/censoring for individual i with an observed onset. Then the observed sample points in this study are $(X_i, \delta_i) = (\min(T_i, C_i), 1_{T_i \leq C_i}); i \in \{1, 2, ..., n\}$.

2.4 Prevalent Cohort Study

Imbedded in the CSHA was a prevalent cohort study with follow-up. The data from this sub-study were used to estimate survival of individuals with dementia from onset [45]. In such studies, a cohort of individuals with an existing (prevalent) condition is identified from a sample of subjects drawn from some population. These prevalent cases are followed from their respective recruitment dates to failure or censoring, which may

occur at the end of the study. This type of study is typically used for diseases with longer failure times as individuals entered into the study already have the condition, unlike in an incident cohort study where all subjects are initially disease-free. With longer failure times, the incident cohort study would have to follow subjects from their respective onset dates for a prolonged period of time to observe a sizeable amount of failures. The prevalent cohort study, which collects subjects already with an existing condition, does not have to be maintained for the same amount of time as the incident cohort study to observe failures, as subjects entered into the study may already have had the condition for an extended period of time.

For simplicity of exposition, it is almost always assumed that all subjects are recruited and screened on a single calendar date known as *prevalence day*, which we denote by R. This is a simplification of what occurs in practice as the dates of recruitment are usually staggered. If there is no cohort effect (i.e. survival is independent of the calendar date of recruitment), this assumption does not affect the survival analysis [2].

Each positively screened subject i is followed from R to the minimum of their calendar date of failure, denoted by T_i^{cal}, and their calendar date of censoring, denoted by C_i^{cal}. Note that all $T_i^{cal} < R$ are left-truncated by R and are unobserved [1]. Assuming some well defined notion of onset for the disease, the onset date for subject i in the prevalent cohort is obtained retrospectively. We denote this date as U_{0i}. The observed sample points in this study are $(X_i, \delta_i) = (Q_i - U_{0i}, \delta_i) = (\min(T_i^{cal}, C_i^{cal}) - U_{0i}, 1_{T_i \leq C_i})$ where $Q_i = \min(T_i^{cal}, C_i^{cal})$, $\forall i \in \{1, 2, ..., n\}, X_i > R - U_{0i}$. Figure 2.2 is a graphical representation of prevalent cohort data with various failure/censoring times.

Definition 2.4.1. *(**Forward/Backward Recurrence Times**) Let U_{0i} be the onset date, R be the prevalence date and Q_i be the calendar date of failure/censoring for subject*

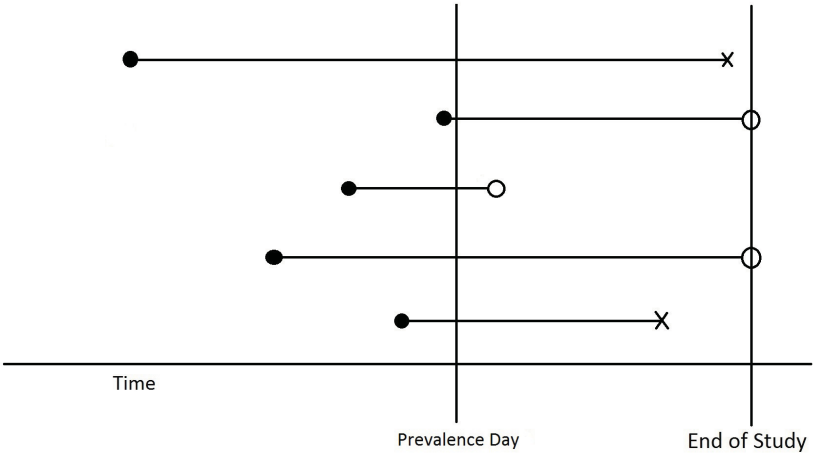

Figure 2.2 – *Graphical representation of prevalent cohort data. The filled circles, crosses and open circles denote, respectively, disease onset, failure and censoring.*

i. Then $R-U_{0i}$ *and* Q_i-R *are the* forward recurrence time *and the* backward recurrence time*, respectively corresponding to subject* i*.*

We give a detailed discussion of uncertainty in the ascertainment of the onset date and its effect on the forward/backward recurrence times in Chapter 3. Note that in a prevalent cohort study, only the subjects with prevalent disease on date R provide their onset dates. The onset dates of subjects with truncated failure times are missing and unobserved. We refer to the *full onset process* as the stochastic point process which generates the set of all observed and unobserved onset dates. A *stationary onset process* is the stochastic point process for which the set of all onset times arise from a stationary Poisson process. In this case, as is well known, conditioned on the number of onsets in a pre-specified interval, the onset times are independently and identically distributed uniform random variables on this interval.

Chapter 3

Modelling Uncertainty in Failure Time Data

3.1 Measurement Error Models

In order to quantify the uncertainty between the true (unobserved) and the observed initial event times, we examine two forms of linear measurement error models. As mentioned in the introduction, for a dementia prevalent cohort, the true onset date (i.e. the true initial event time) may not be known or even defined. We assume in all further discussions, for mathematical and practical simplicity, that a true date of the initial event that gives rise to the collected failure time data exists and is clearly defined.

Let Z_1 be the observed random variable, Z_2 be the true unobserved random variable and let $\epsilon \sim N(0, \sigma^2)$ be an unobserved random variable representing some form of error. When the observed random variable Z_1 is expressed as the sum of Z_2 and ϵ:

$$Z_1 = Z_2 + \epsilon, \tag{3.1}$$

we denote this model as the *Classical Measurement Error Model*. When the unobserved

random variable Z_2 is expressed as the sum of Z_1 and ϵ:

$$Z_2 = Z_1 + \epsilon, \tag{3.2}$$

we denote this model as the *Berkson Measurement Error Model*. For both equations, the assumption that the errors are normally distributed about 0 can be relaxed to allow for skewed error distributions [5] [6]. In the case of an error term that is not symmetrically distributed about 0, the models 3.1 and 3.2 are not functionally equivalent. In practice, they represent two different forms of measurement error, which we denote respectively as *controlled* and *uncontrolled* [5] [6]. We examine the difference between these two measurement errors in the context of a chemical experiment as discussed by Berkson [4]. Suppose some intended amount of chemical A is added to a system and the resulting amount of chemical B by-product of the system reaction is observed. Under the assumption that there is error in the measurement of the added chemical A due to the inaccuracy of the scientific instruments, this form of error would be referred to as controlled. It is controlled to the extent that the experiementer controls the amount of chemical A that is intended to be added. In contrast, the amount of chemical B being observed has an uncontrolled measurement error because the true amount of chemical B resulting from the reaction is not known a priori [4]. Since there can be debate on which form of measurement error relates the true and observed dates of the initiating event in survival analysis more appropriately, we examine the implications of both in the context of incident and prevalent cohort studies.

3.2 Censoring Indicator Assumption with Measurement Error

For incident and prevalent cohort studies, some of the failure time data may be right-censored. The data collected from such studies are the pairs (X_i, δ_i) $i \in \{1, 2, ..., n\}$, where X_i is the observed failure/censoring time and δ_i is the right-censoring indicator. We assume that the measurement error terms in equations 3.1 or 3.2 are independent of the failure/censoring indicator $\delta_i, \forall i \in \{1, 2, ..., n\}$. We examine the validity of this assumption through the following example. Suppose a cohort of subjects are followed from their respective incidence/recruitment dates to their failure/censoring calendar dates. For each subject, the researcher either observes the death of the subject or the subject's survival. The retrospective inclusion of measurement error in the reported incidence/onset date does not affect the original observation of death or survival made by the researcher. The inclusion of measurement error only affects the time interval from the reported incidence/onset date to the failure/censoring calendar date. We ignore the possibility of the observation of death or survival being in error.

3.3 Effects of Measurement Error in Incident Cohort Studies

In the case of an incident cohort study, the failure time data are the differences between the calendar date of failure/censoring and the recorded calendar date of incidence. When we assume a measurement error model for the incidence calendar date, depending on the support of the error term, the error can translate the incidence date to the right or to the left. Thus, the error term shortens or lengthens the failure/censoring times depending

on whether it takes a negative or positive value, respectively. Let T_i^* and C_i^* represent, respectively, the true failure and censoring times for subject i with an observed onset. For incident cohort failure time data, the classical measurement error model has the form:

$$\min(T_i, C_i) = \min(T_i^*, C_i^*) + \epsilon_i, i \in \{1, 2, ..., n\} \tag{3.3}$$

and the random Berkson measurement error model has the form:

$$\min(T_i^*, C_i^*) = \min(T_i, C_i) + \epsilon_i, i \in \{1, 2, ..., n\} \tag{3.4}$$

We include the subscript i on the error term ϵ to allow the measurement error to vary according to each subject. Additionally, we will assume that all observed and unobserved failure/censoring random variables are independent of $\epsilon_i, \forall i \in \{1, 2, ..., n\}$ (i.e. the error does not vary according to the length of the observed/unobserved failure/censoring times). The two types of measurement error are represented graphically in figure 3.1.

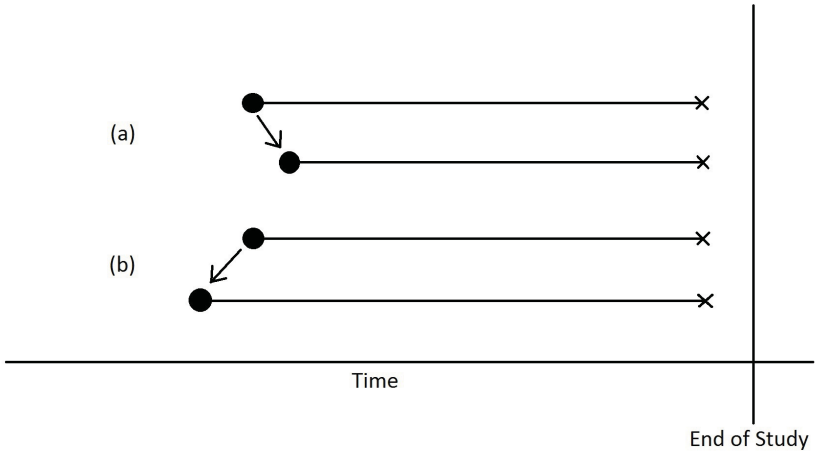

Figure 3.1 – *A graphical representation of incident cohort failure times where the incidence dates are affected by the presence of measurement error (denoted by arrows). In case (a), the measurement error translates the incident date to the right and in case (b), the measurement error translates the incident date to the left.*

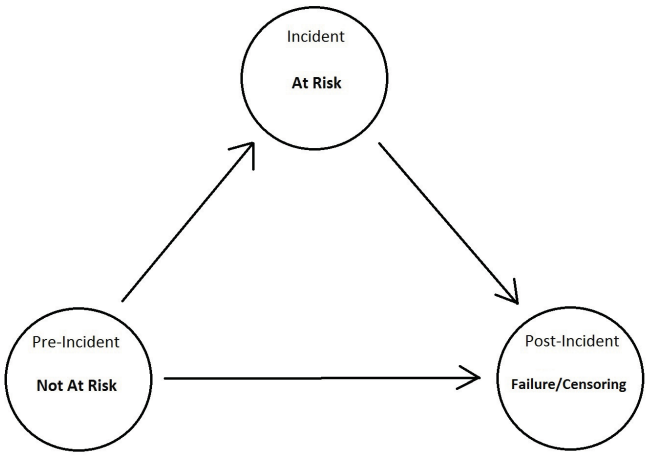

Figure 3.2 – *A three stage transition diagram for subjects in an incident cohort study. Transitions can occur between the pre-incident stage to the incident stage, from the incident stage to the failure/censoring stage and from the pre-incident stage to the failure/censoring stage.*

From figure 3.1, when assuming the classical measurement error model, the upper line, in cases (a) and (b), represents the true failure time lengths while the line below represents the observed failure time lengths. When assuming the Berkson measurement error model, the referencing to each line is exchanged. We examine the statistical implications of cases (a) and (b) with respect to both measurement error models.

Under the classical measurement error model, if ϵ_i is a non-positive random variable bounded below by $-\min_i(T_i^*, C_i^*), i \in \{1, 2, ..., n\}$, then the true incidence date lies to the left of the observed incidence date. This particular assumption on the support of ϵ_i is thus equivalent to the true incidence date being left-censored by the observed incidence calendar date. For the Berkson measurement error model, if ϵ_i is a non-positive random variable bounded below by $-\min_i(T_i, C_i), i \in \{1, 2, ..., n\}$, then the true unobserved incidence date occurs beween the observed incidence calendar date and the observed calendar date of failure/censoring (i.e. the true unobserved incidence date is interval-censored by the observed incidence date and the calendar date of failure/censoring). Similarly, if ϵ_i

is a non-negative random variable, $i \in \{1, 2, ..., n\}$, then under the classical measurement error model, the true incidence date occurs between the observed incidence date and the calendar date of failure/censoring (i.e. the true incidence date is interval-censored by the observed incidence date and the calendar date of failure/censoring). Under the Berkson measurement error model with the same error assumption, the true incidence date occurs before the observed incidence date (i.e. the true incidence date is left-censored by the observed incidence date). In figure 3.2, these error assumptions correspond to the scenario where a subject transitions from the pre-incident stage to the incident (at risk of failure) stage and then to the failure/censoring stage. These assumptions on the error thus imply that no observations are lost in the data collection step of the study.

When ϵ_i is negative, without further restrictions, there is positive probability that the error term will translate the incidence date past the observed failure/censoring calendar date.

Definition 3.3.1. *(**Type 1 and Type 2 Misclassification**) In the context of incident/prevalent cohort studies, with the inclusion of measurement error in the initial event time, a* type 1 misclassification error (MC-1) *of subject i occurs when the subject is incorrectly excluded from the study due to the effect of the measurement error. A* type 2 misclassification error (MC-2) *occurs when the subject is incorrectly included in the study.*

For the classical measurement error model, if the observed incidence date is to the right of the failure/censoring date, even though the true date falls before failure/censoring, a subject in the initial disease-free cohort would not be included in the cohort of incident cases. This would result in an MC-1 error. Under the Berkson measurement error model, if the true incidence date falls to the right of the observed failure date and the observed incidence date before, then the failure event would be incorrectly associated with "the

disease". Likewise, if the true incidence date falls to the right of the censoring date. Both cases in the Berkson measurement error model are examples of MC-2 errors. In figure 3.2, this scenario corresponds to a subject transitioning from the pre-incident stage directly to the post-incident failure/censoring stage (bypassing the second stage).

When the failure time data are both left-truncated and right-censored, the inclusion of measurement error may affect both the failure/censoring times as well as the observed truncation times. Due to this added effect, in section 3.4, we examine the scenario where measurement error is included in left-truncated right-censored failure time data in the context of a prevalent cohort study.

3.4 Effects of Measurement Error in Prevalent Cohort Studies

In a prevalent cohort study with follow-up, we model the uncertainty in the failure time data by an additive error term between the observed onset date and the true (unobserved) onset date. Let $V_{0i}, i \in \{1, 2, ..., n\}$ be the true onset dates for subjects in the prevalent cohort study.

Then the classical measurement error model has the form:

$$\min(T_i^{cal}, C_i^{cal}) - U_{0i} = \min(T_i^{cal}, C_i^{cal}) - (V_{0i} + \epsilon_i), i \in \{1, 2, ..., n\} \tag{3.5}$$

and the Berkson measurement error model has the form:

$$\min(T_i^{cal}, C_i^{cal}) - V_{0i} = \min(T_i^{cal}, C_i^{cal}) - (U_{0i} + \epsilon_i), i \in \{1, 2, ..., n\} \tag{3.6}$$

where we assume $U_{0i} \leq R$, and $\min(T_i^{cal}, C_i^{cal}) - U_{0i} > R, \forall i \in \{1, 2, ..., n\}$. Figure 3.3

represents the ways in which the error term can translate the onset time either positively or negatively. Under the classical measurement error model, the upper line in 3.3, in cases (a), (b1) and (b2), represents the true failure time lengths while the line below represents the observed failure time lengths. Under the Berkson measurement error model, the referencing is exchanged. As with the incident cohort study, we examine the statistical implications of various error assumptions on prevalent cohort failure time data.

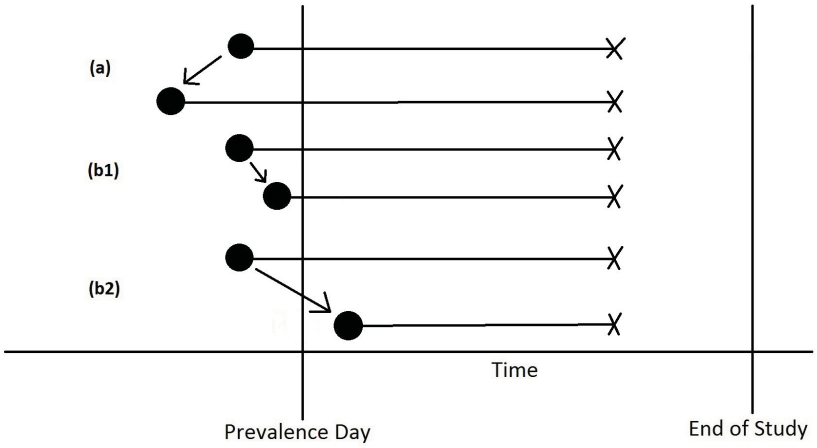

Figure 3.3 – *A graphical representation of prevalent cohort failure times where the onset dates are affected by the presence of measurement error (denoted by arrows). In case (a), the measurement error translates the onset date to the left. In case (b1), the measurement error translates the onset date to the right, prior to prevalence day. In case (b2), the measurement error translates the onset date to the right, past prevalence day.*

First, we examine the implications of case (a) where there is a negative translation and case (b1) where the translated onset date occurs prior to R. We assume in both cases, for simplicity, that there is no screening error on prevalence day, as the a priori assumption of the error terms implies that $V_{0i} < R$ for subjects included in the prevalent cohort. When ϵ_i is a non-positive random variable, $i \in \{1, 2, ..., n\}$, under the classical measurement error model, $V_{0i} > U_{0i}$ (i.e. the observed onset date occurs prior to the true

onset date). Since subject i is in the prevalent cohort and it is assumed that $V_{0i} < R$, then V_{0i} is interval-censored by (U_{0i}, R). Under the Berkson measurement error model, this error assumption implies that $V_{0i} < U_{0i}$ (i.e. the true onset date is left-censored by the observed onset date).

If subject i is in the prevalent cohort and $V_{0i} < R$, when ϵ_i is a non-negative random variable bounded above by $R - V_{0i}, i \in \{1, 2, ..., n\}$, then under the classical measurement error model, V_{0i} is left-censored by U_{0i}. When ϵ_i is a non-negative random variable bounded above by $R - U_{0i}, i \in \{1, 2, ..., n\}$, then under the Berkson measurement error model, V_{0i} is interval-censored by (U_{0i}, R). When it is only known that ϵ_i is bounded above by $R - U_{0i}, i \in \{1, 2, ..., n\}$ under the classical measurement error model or that ϵ_i is bounded above by $R - V_{0i}, i \in \{1, 2, ..., n\}$ under the Berkson measurement error model, it can only be concluded that V_{0i} is left-censored by R. Under these two error assumptions, it is not possible to further specify the location of V_{0i} as the error term is unobserved and it is not known whether the error translated the original onset date forwards or backwards.

For case (b2), in figure 3.3, when the error term translates the onset date past prevalence day, we must also consider the result of the screening test conducted on prevalence day. Suppose ϵ_i is a non-negative random variable and bounded below by $R - U_{0i}$, with $U_{0i} < R$ and $V_{0i} = U_{0i} + \epsilon_i > R$ for some $i \in \{1, 2, ..., n\}$. If the support of the error term is known a priori, then under the assumption of no possible screening error, subject i would be excluded from the prevalent cohort as they did not test positive on date R. Suppose ϵ_i is a non-negative random variable bounded below by $R - V_{0i}$, with $V_{0i} < R$ and $U_{0i} = V_{0i} + \epsilon_i > R$ for some $i \in \{1, 2, ..., n\}$. For subjects included in the prevalent cohort, we set their backward recurrence times to 0 and their failure/censoring

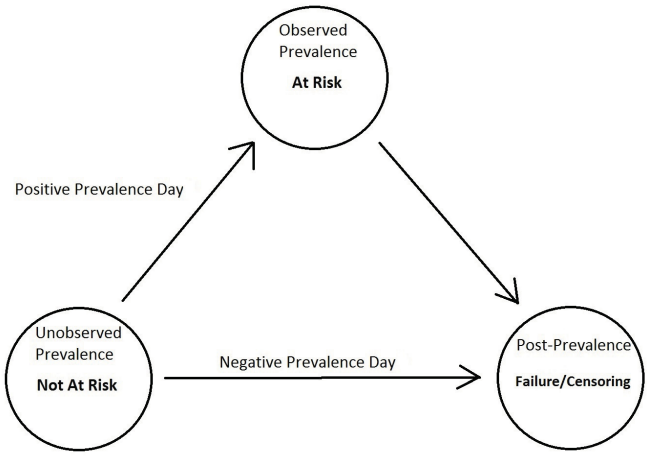

Figure 3.4 – *A three stage transition diagram for subjects in a prevalent cohort study. Transitions can occur between the unobserved prevalence stage to the observed prevalence stage, from the prevalence stage to the failure/censoring stage and from the unobserved prevalence stage to the failure/censoring stage.*

times to their respective forward recurrence times.

In the case where ϵ_i has unbounded positive and negative support, and it is not known a priori whether V_{0i} occurs before or after R, we must allow for the possibility of a screening error on prevalence day. If the test result on date R is positive, and $V_{0i} > R$, the recalled onset date and the positive test result would be incorrectly associated with "the disease". Subject i should not be included in the prevalent cohort (i.e. an MC-2 error). If the test result is negative, and $V_{0i} < R$, subject i would be incorrectly excluded from the prevalent cohort (i.e. an MC-1 error). We provide a transition diagram in figure 3.4 to graphically represent how a given subject can transition from the unobserved prevalence stage to the failure/censoring stage depending on the result of the prevalence day test. In figure 3.4, when the test result on prevalence day is positive, irrespective of whether the subject should be included in the prevalent cohort, the subject would transition from the unobserved prevalence stage to the observed prevalence stage and then to the

failure/censoring stage. When the test result is negative, the subject is excluded from the prevalent cohort and would transition directly from the unobserved prevalence stage to the failure/censoring stage. In Chapter 4, we present different ways in which uncertainty in the onset date may be included in a survival model and in Chapter 5, we examine the effect of measurement error on estimated survival under various parametric models in the prevalent cohort study setting.

Chapter 4

Survival Modelling with Uncertainty in the Reported Onset Time

We present three different methods that model the various forms of uncertainty discussed in Chapter 3. In Section 4.1, we examine the parametric approach of Zhong and Cook for left-truncated right-censored failure time data under a classical measurement error model [46]. In Section 4.2, we review the general theory behind convolution/deconvolution operations. We apply this theory in conjunction with kernel density estimation techniques to obtain a non-parametric density estimator for the density of the underlying failure time random variable in a classical measurement error model. In the final section, we present the numerical density estimation method of Sun which assumes the failure time data is doubly interval-censored and left-truncated [25] [38].

4.1 Parametric Maximum Likelihood Estimation

In this section, we review maximum likelihood methods in survival analysis for right-censored left-truncated failure time data. We carefully examine the approach developed by Zhong and Cook in [46], in the presence of measurement error for left-truncated right-censored failure time data.

Let $T_i, i \in \{1, 2, ..., n\}$, be independently and identically distributed failure times with parametric probability density function $f(t; \boldsymbol{\theta})$ and survivor function $S_T(t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional vector of parameters. Typical parametric families of failure time distributions in survival analysis include the exponential, Weibull, log-normal, generalized gamma, log-logistic or generalized F [23]. When all the T_i are observed and there is no censoring, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes the likelihood function:

$$L(\boldsymbol{\theta}|(t_1, ..., t_n)) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\theta}) \tag{4.1}$$

When a subset of the $T_i, i \in \{1, 2, ..., n\}$, are right-censored, this classic likelihood must be modified. Let $C_i, i \in \{1, 2, ..., n\}$ be i.i.d. censoring time random variables with parametric density function $g(t; \boldsymbol{\psi})$ and survivor function $S_C(t; \boldsymbol{\psi})$. We observe the pairs $(X_i = \min_i(T_i, C_i), \delta_i), i \in \{1, 2, ..., n\}$, where δ_i is the failure/censoring indicator function. To derive the likelihood function for $\boldsymbol{\theta}$, we consider the contributions to the full likelihood based on whether the failure time t was observed or censored [23]. For an observed failure time, $X = t, \delta = 1$, the contribution to the likelihood is, essentially, given by:

$$P(X \in [t, t + \Delta t), \delta = 1) = P(T \in [t, t + \Delta t), C \geq t)$$

which by random censoring, is approximately equal to

$$f(t;\boldsymbol{\theta})S_C(t;\boldsymbol{\psi})\Delta t$$

Similarly, for an observed censored failure time, $X = t, \delta = 0$, the contribution to the likelihood is essentially given by:

$$P(X \in [t, t+\Delta t), \delta = 0) = P(C \in [t, t+\Delta t), T > t)$$

which, by random censoring, is approximately equal to

$$g(t;\boldsymbol{\psi})S_T(t;\boldsymbol{\theta})\Delta t.$$

Now divide throughout by Δt and let $\Delta t \to 0$. Then, under the further assumption of non-informative censoring (i.e. $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ share no common parameters), the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes the likelihood function:

$$L(\boldsymbol{\theta}|(X,\delta)) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})^{\delta_i} S_T(x_i|\boldsymbol{\theta})^{1-\delta_i} \tag{4.2}$$

In this parametric setting, by the invariance of MLEs, $\hat{S}(t|\boldsymbol{\theta}) = S(t|\hat{\boldsymbol{\theta}})$, $\hat{f}(t|\boldsymbol{\theta}) = f(t|\hat{\boldsymbol{\theta}})$ and $\hat{\lambda}(t|\boldsymbol{\theta}) = \lambda(t|\hat{\boldsymbol{\theta}})$ [23].

In a prevalent cohort study setting, we adjust 4.2 for both left-truncated and right-censored failure time data. We follow the approach of Wang in [42], in a parametric setting. This requires virtually no modification of Wang's non-parametric derivation of the likelihood. Let X_i be the observed failure time, C_i be the potential censoring time and W_i be the left-truncation time for subject i, $i \in \{1, 2, ..., n\}$. Let $H(\cdot, \cdot; \boldsymbol{\psi})$ be the bivariate cumulative distribution function of (W_i, C_i) where $\boldsymbol{\psi}$ is the p-dimensional vector of parameters indexing the distribution of the random variable C_i. We assume $P(W_i < C_i) = 1$ (i.e. subjects that are not yet under follow-up cannot be lost to follow-up), X_i is left-truncated by W_i and X_i is independent of (W_i, C_i) $\forall i \in \{1, 2, .., n\}$. We

construct the full likelihood based on the observed failure/censoring and truncation times. For an observed failure time y and truncation time w, the contribution to the likelihood is, essentially, given by:

$$P(X \in [y, y+dy), W \in [w, w+dw), \delta = 1 | X > W; \boldsymbol{\theta}, \boldsymbol{\psi})$$

$$= P(X \in [y, y+dy), W \in [w, w+dw), C > y | X > W; \boldsymbol{\theta}, \boldsymbol{\psi})$$

$$\approx \frac{f(y; \boldsymbol{\theta}) \int_y^\infty dH(w, u; \boldsymbol{\psi})}{P(X > W; \boldsymbol{\theta})} dy dw I(\delta = 1) I(y > w)$$

For an observed censored failure time y and truncation time w, the contribution to the likelihood is, essentially, given by:

$$P(C \in [y, y+dy), W_i \in [w, w+dw), \delta = 0 | X > W; \boldsymbol{\theta}, \boldsymbol{\psi})$$

$$= P(C \in [y, y+dy), W_i \in [w, w+dw), X > y | X > W; \boldsymbol{\theta}, \boldsymbol{\psi})$$

$$\approx \frac{S(y; \boldsymbol{\theta}) dH(w, y; \boldsymbol{\psi})}{P(X > W; \boldsymbol{\theta})} dy dw I(\delta = 0) I(y > w)$$

Divide throughout by $dydw$ and let $dy \to 0$ and $dw \to 0$. The maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}})$ maximizes the likelihood function:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^n \frac{1}{P(X_i > W_i; \boldsymbol{\theta})} f(y_i; \boldsymbol{\theta})^{\delta_i} S(y_i; \boldsymbol{\theta})^{1-\delta_i} dH(w_i, y_i; \boldsymbol{\psi})^{1-\delta_i} \left(\int_{y_i}^\infty dH(w_i, y; \boldsymbol{\psi}) \right)^{\delta_i} \tag{4.3}$$

The likelihood in 4.3 can then be decomposed as the product of two functions, denoted by L_1 and L_2, by multiplying and dividing by the survivor function evaluated at the observed truncation time [42]:

$$L_1 = \prod_{i=1}^n \frac{f(y_i; \boldsymbol{\theta})^{\delta_i} S(y_i; \boldsymbol{\theta})^{1-\delta_i}}{S(w_i; \boldsymbol{\theta})} \tag{4.4}$$

$$L_2 = \prod_{i=1}^n \frac{1}{P(X_i > W_i; \boldsymbol{\theta})} S(w_i; \boldsymbol{\theta}) dH(w_i, y_i; \boldsymbol{\psi})^{1-\delta_i} \left(\int_{y_i}^\infty dH(w_i, u; \boldsymbol{\psi}) \right)^{\delta_i} \tag{4.5}$$

In [46], Zhong and Cook also stated a parametric likelihood function for left-truncated right-censored failure time data without the inclusion of measurement error. Without first writing down the full likelihood, they claimed that 4.4 is proportional to the conditional likelihood for $\boldsymbol{\theta}$, conditional on the observed onset dates. However, as pointed out by Wang, 4.4 cannot be interpreted as a conditional likleihood since 4.5 is dependent on the observed failure times y_i. In the nonparametric setting of [42], Wang showed that the MLE of L_1 with respect to $S(t)$ is equivalent to the MLE obtained from the full likelihood, even though the term $S(t)$ appears in L_2. In the "working data" case, in which all censoring times are observed even for subjects who have failed, 4.4 is a conditional likelihood. Making use of this observation, Wang justifies maximization of only L_1 with respect to $S(t)$ nonparametrically. The invocation of this so-called working dataset is not of course necessary for the validity of Wang's argument. Now in the parametric setting, without measurement error, the maximization of L_1 is not sufficient to obtain the MLE of $\boldsymbol{\theta}$. This can be seen, since the $\boldsymbol{\theta}$ parameters appear in both the prevalent cohort inclusion probability and survivor function terms of 4.5. Moreover, without the assumption of joint non-informativeness of the truncation and censoring times, the MLE of $\boldsymbol{\psi}$ may yield information on the estimation of $\boldsymbol{\theta}$. Zhong and Cook appear to use L_1 in the same context as Wang by supposing maximization of L_1, with respect to $\boldsymbol{\theta}$, provides a maximum likelihood estimator. L_1 in this setting is merely some function of $\boldsymbol{\theta}$. Certainly, the maximizer of L_1 with respect to $\boldsymbol{\theta}$ is not the maximizer of the full likelihood as is the case in the non-parametric setting.

We turn now to the measurement error model. The issues raised above remain for this model. For the sake of completeness, we present the derivation of what Zhong and Cook call the "correct conditional likelihood" for left-truncated right-censored failure time data

when there is uncertainty in the onset dates since the derivation, nevertheless, suggests a possible fruitful approach to the problem of modelling uncertainty in the date of onset. We then derive what we assert to be the full likelihood in the presence of measurement error in the onset dates.

Let the observed failure time $T_i = V_{1i} - U_{0i}, i \in \{1, 2, ..., n\}$ where V_{1i} is the calendar date of failure and U_{0i} is the reported date of onset. Let R be the fixed prevalence date. Let V_{0i} be the true random onset date with marginal density $f_0(\cdot)$ and assume that $A < U_{0i} < R$ and $A < V_{0i} < R$ for $i \in \{1, 2, ..., n\}$ for some constant lower bound A. As in Chapter 3, we assume the classical measurement error model for the observed onset date U_{0i} and the true onset date V_{0i}:

$$U_{0i} = V_{0i} + \epsilon_i \tag{4.6}$$

where $\epsilon_i \sim^{i.i.d} N(0, \sigma^2)$ is the random measurement error. Since $A < U_{0i} < R$, this implies that

$$A - V_{0i} < \epsilon_i < R - V_{0i}$$

and thus the random error term ϵ_i has a normal distribution truncated to the right by $R - V_{0i}$ and to the left by $A - V_{0i}$. By conditioning on $V_{0i} = v_{0i}$, it follows that $U_{0i} \stackrel{d}{=} \epsilon_i + v_{0i}$. By applying the standard density transformation formula for a single random variable, the density function of the observed onset time, conditional on $V_{0i} = v_{0i}$ and $A \leq U_{0i} \leq R$, is derived as follows:

$$g(u_{0i}|V_{0i} \in [v_{0i}, v_{0i} + dv_{0i}), A \leq U_{0i} \leq R; \sigma^2) du_{0i} dv_{0i}$$

$$\approx \frac{\mathbb{P}(U_{0i} - V_{0i} \in [u_{0i} - v_{0i}, u_{0i} - v_{0i} + du_{0i} - dv_{0i}); \sigma^2)}{\mathbb{P}(V_{0i} \in [v_{0i}, v_{0i} + dv_{0i}), A \leq U_{0i} \leq R; \sigma^2)}$$

$$= \frac{\mathbb{P}(\epsilon_i \in [u_{0i} - v_{0i}, u_{0i} - v_{0i} + du_{0i} - dv_{0i}); \sigma^2)}{\mathbb{P}(V_0 \in [v_{0i}, v_{0i} + dv_{0i}), A \leq U_{0i} \leq R; \sigma^2)}$$

$$\approx \frac{f_\epsilon(u_{0i} - v_{0i}; \sigma^2)}{\mathbb{P}(U_{0i} - V_{0i} \leq R - v_{0i}; \sigma^2) - \mathbb{P}(U_{0i} - V_{0i} \leq A - v_{0i}; \sigma^2)} du_{0i} dv_{0i}$$

Dividing through both sides by $du_{0i}dv_{0i}$ and letting them tend to 0, the density of the observed onset date, conditional on the unobserved true onset date, is given by:

$$g(u_{0i}|v_{0i}, A \leq U_{0i} \leq R; \sigma^2) = \frac{f_\epsilon(u_{0i} - v_{0i}; \sigma^2)}{F_\epsilon(R - v_{0i}; \sigma^2) - F_\epsilon(A - v_{0i}; \sigma^2)} \tag{4.7}$$

where f_ϵ and F_ϵ are the pdf and cdf of the normal distribution with mean 0 and variance σ^2.

To obtain the maximum likelihood estimator for $\boldsymbol{\theta}$, Zhong and Cook derived the so-called "correct conditional likelihood" function based on the calendar date of failure, conditional on the reported onset date U_{0i}. For subject i observed to fail at v_{1i}, the contribution to the conditional likelihood is given by:

$$\mathbb{P}(V_{1i} \in [v_{1i}, v_{1i} + dv_{1i}), \delta_i = 1 | V_{1i} > R, U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), A \leq V_{0i} \leq R; \boldsymbol{\theta}, \sigma^2)$$

$$= \frac{\mathbb{P}(V_{1i} \in [v_{1i}, v_{1i} + dv_{1i}), U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), A \leq V_{0i} \leq R; \boldsymbol{\theta}, \sigma^2)}{\mathbb{P}(V_{1i} > R, U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), A \leq V_{0i} \leq R; \boldsymbol{\theta}, \sigma^2)} I(\delta_i = 1) I(v_{1i} > R)$$

$$= \frac{\int_A^R \mathbb{P}(V_{1i} - V_{0i} \in [v_{1i} - v_0, v_{1i} - v_0 + dv_{1i}), U_{0i} \in [u_{0i}, u_{0i} + du_{0i}); \boldsymbol{\theta}, \sigma^2 | V_{0i} = v_0) f_0(v_0) dv_0}{\int_A^R \mathbb{P}(V_{1i} - V_{0i} > R - v_0, U_{0i} \in [u_{0i}, u_{0i} + du_{0i}) | V_{0i} = v_0; \boldsymbol{\theta}, \sigma^2) f_0(v_0) dv_0} \times$$

$$I(\delta_i = 1) I(v_{1i} > R)$$

which reduces to

$$= \frac{\int_A^R f_T(v_{1i} - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0} dv_{1i} du_{0i} I(\delta_i = 1) I(v_{1i} > R) \tag{4.8}$$

Similarly, if subject i is censored at v_{1i}, the contribution to the conditional likelihood follows from the above derivation by replacing the expression $V_{1i} \in [v_{1i}, v_{1i} + dv_{1i})$ with $V_{1i} > v_{1i}$. Thus, we obtain:

$$\frac{\int_A^R S_T(v_{1i} - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0} dv_{1i} du_{0i} I(\delta_i = 0) I(v_{1i} > R) \tag{4.9}$$

Dividing through 4.8 and 4.9 by dv_{1i} and du_{0i} and letting them tend to 0, the likelihood for $\boldsymbol{\theta}$ based only on the observed failure times, conditional on the observed onset times, is the following:

$$L_C(\boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \left(\frac{\int_A^R f_T(v_{1i} - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0} \right)^{\delta_i} \times \left(\frac{\int_A^R S_T(v_{1i} - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) g(u_{0i}|v_0; \sigma^2) f_0(v_0) dv_0} \right)^{1-\delta_i} \tag{4.10}$$

Based on the above arguments, as in the case without measurement error, the function defined in 4.10 cannot be interpreted as a conditional likelihood for $\boldsymbol{\theta}$. Since the "correct" unconditional full likelihood of Zhong and Cook is a function of 4.10 and also does not account for the censoring/truncation time random variables, we disregard its derivation. Therefore, we derive the unconditional full likelihood function based on the observed failure, censoring and truncation times with the inclusion of measurement error in the reported onset date.

Let $H(\cdot, \cdot; \boldsymbol{\psi}, \sigma^2)$ denote the bivariate distribution function of (U_{0i}, C_i) for all $i \in \{1, 2, ..., n\}$ indexed by the parameters $\boldsymbol{\psi}$ and σ^2. Let $f_C(\cdot; \boldsymbol{\psi})$ and $S_C(\cdot; \boldsymbol{\psi})$ denote the respective density and survivor functions of C_i, $i \in \{1, 2, ..., n\}$. We assume the failure time $T_i = V_{1i} - V_{0i}$ is independent of (U_{0i}, C_i) for all $i \in \{1, 2, ..., n\}$. As in the derivation of 4.3, we consider the probabilistic contributions to the likelihood function based on the observed failure/censoring and truncation times. For subject i observed to fail at the calendar date v_{1i}, the contribution to the likelihood is given by:

$$\mathbb{P}(V_{1i} \in [v_{1i}, v_{1i} + dv_{1i}), U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), C_i > V_{1i} - V_{0i}, \delta_i = 1 | V_{1i} > R; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2)$$

$$= \frac{\mathbb{P}(V_{1i} \in [v_{1i}, v_{1i} + dv_{1i}), U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), C_i > V_{1i} - V_{0i}, V_{1i} > R; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2)}{\mathbb{P}(V_{1i} > R; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2)} I(\delta_i = 1) I(v_{1i} > R)$$

$$= \frac{\int_A^R \mathbb{P}(V_{1i} \in [v_{1i}, v_{1i} + dv_{1i}), U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), C_i > V_{1i} - v_0 | V_{0i} = v_0; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2) f_0(v_0) dv_0}{\int_A^R \mathbb{P}(V_{1i} > R | V_{0i} = v_0; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2) f_0(v_0) dv_0} \times$$

$$I(\delta_i = 1) I(v_{1i} > R)$$

which simplifies to

$$= \frac{\int_A^R f_T(v_{1i} - v_0; \boldsymbol{\theta}) \int_{v_{1i}-v_0}^{\infty} dH(u_{0i}, w; \boldsymbol{\psi}, \sigma^2) dw f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) f_0(v_0) dv_0} dv_{1i} du_{0i} I(\delta_i = 1) I(v_{1i} > R)$$

Similarly, for subject i censored at v_{1i}, the contribution to the likelihood is given by:

$$\mathbb{P}(V_{1i} > v_{1i}, U_{0i} \in [u_{0i}, u_{0i} + du_{0i}), C_i \in [v_{1i} - V_{0i}, v_{1i} - V_{0i} + dv_{1i}), \delta_i = 0 | V_{1i} > R; \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2)$$

which reduces to

$$= \frac{\int_A^R S_T(v_{1i} - v_0; \boldsymbol{\theta}) dH(u_{0i}, v_1 - v_0; \boldsymbol{\psi}, \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) f_0(v_0) dv_0} dv_{1i} du_{0i} I(\delta_i = 0) I(v_{1i} > R)$$

Dividing through by dv_{1i} and du_{0i} and letting them tend to 0, the likelihood for $(\boldsymbol{\theta}, \boldsymbol{\psi})$, based on the observed failure/censoring and truncation times, is the following:

$$L_F(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} \left(\frac{\int_A^R f_T(v_{1i} - v_0; \boldsymbol{\theta}) \int_{v_{1i}-v_0}^{\infty} dH(u_{0i}, w; \boldsymbol{\psi}, \sigma^2) dw f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) f_0(v_0) dv_0} \right)^{\delta_i} \times \left(\frac{\int_A^R S_T(v_{1i} - v_0; \boldsymbol{\theta}) dH(u_{0i}, v_{1i} - v_0; \boldsymbol{\psi}, \sigma^2) f_0(v_0) dv_0}{\int_A^R S_T(R - v_0; \boldsymbol{\theta}) f_0(v_0) dv_0} \right)^{1-\delta_i} \tag{4.11}$$

Unlike the case without measurement error, there is no obvious factorization of 4.11 into a product $L_1 L_2$. Hence, based on this observation, there is no $L_C(\boldsymbol{\theta}, \sigma^2)$ function as proposed by Zhong and Cook. From a preliminary examination of 4.11, it is clear that analytical maximization is infeasible and numerical procedures are required to obtain an estimator for the $\boldsymbol{\theta}$ parameters. An EM algorithm may provide another approach to

estimate the true underlying parameters by augmenting the likelihood with the unobserved exact onset dates as missing data. However, it is not clear how to evaluate the expectation in this algorithm. Given the complexity of the parametric estimation problem, obtaining a non-parametric estimator for the survivor function S_T from 4.11 appears even more challenging. Due to the difficulty of this optimization, we examine a simpler non-parametric modelling approach in the context of a deconvolution problem in section 4.2.

4.2 Non-Parametric Density Estimation

Given the classical measurement error model of 3.1, if the error term ϵ is assumed to be independent of the true unobserved random variable Z_2, then the following relationship holds between their respective density functions: $f_{Z_1} = f_{Z_2} * f_\epsilon$ [31]. In turn, to estimate f_{Z_2} based on the observed data Z_1, and an assumed known error density f_ϵ, this suggests a deconvolution estimation problem [12]. In this section, we examine the standard deconvolution procedure as presented by Delaigle in [12] and describe how the procedure can be adapted for left-truncated right-censored failure time data based on the equations 3.3 and 3.5.

Let $T_i, i \in \{1, 2, ..., n\}$ be independently and identically distributed failure time random variables with non-parametric density function $f_T(\cdot)$. When there is no censoring and all T_i are fully observed, the density $f_T(\cdot)$ can be estimated using a kernel density estimator. The kernel density estimator is an average of smoothing functions (i.e. kernels)

evaluated at predefined points [35]. The kernel density estimator for f_T is given by:

$$\hat{f}_T(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - T_i}{h}\right) \tag{4.12}$$

where h is a constant tuning parameter. When the random variable T_i is observed with error, we consider the following classical measurement error model:

$$Y_i = T_i + \epsilon_i \tag{4.13}$$

where the Y_is are the observed failure times, $T_i \sim^{i.i.d} f_T(\cdot)$ are the true underlying failure times and the ϵ_is are i.i.d with known parametric density $f_\epsilon(\cdot; \boldsymbol{\theta})$. As shown in [12] and [31], the authors derived a non-parametric estimator of $f_T(\cdot)$ using the properties of the convolution operator $*$ and the Fourier transform $\mathcal{F}$. Given two independent random variables X_1 and X_2 with respective densities f_1, f_2, the density of $X_1 + X_2$ is defined through the convolution operator between the two density functions [31]:

$$(f_1 * f_2)(x) = \int_{-\infty}^{\infty} f_1(w) f_2(x - w) dw \tag{4.14}$$

The Fourier transform of a random variable X is the characteristic function of the random variable [7]. Therefore, as is well known,

$$\mathcal{F}(f_1 * f_2)(s) = \mathcal{F}(f_1)(s) \times \mathcal{F}(f_2)(s)$$

Let $\hat{f}_Y(x)$ be the kernel density estimator based on the i.i.d observed failure times Y_i, of 4.13. Through the use of the operations described above, we derive a non-parametric estimator for the underlying ("true") p.d.f. f_T, based on the measurement error model 4.13. Thus,

$$\hat{f}_Y(x) = f_T(x) * f_\epsilon(x; \boldsymbol{\theta})$$

implies

$$\mathcal{F}(\hat{f}_Y)(s) = \mathcal{F}(f_T)(s) \times \mathcal{F}(f_\epsilon)(s; \boldsymbol{\theta}).$$

That is,

$$\mathcal{F}(f_T)(s) = \frac{\mathcal{F}(\hat{f}_Y)(s)}{\mathcal{F}(f_\epsilon)(s;\boldsymbol{\theta})}$$

leading to the estimator

$$\hat{f}_T(x) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\hat{f}_Y)(s)}{\mathcal{F}(f_\epsilon)(s;\boldsymbol{\theta})}\right). \tag{4.15}$$

For the estimator in 4.15 to be well defined, we assume that $\mathcal{F}(f_\epsilon)(s;\boldsymbol{\theta}) = \mathbf{0}$ only on a set of probability 0, for all $\boldsymbol{\theta}$ in the parameter space.

When a subset of all the observed failure times Y_i is right-censored, we can adjust the method described above to obtain a non-parametric density estimator for f_T. Given that the Y_is are right-censored, we consider the following measurement error model:

$$Y_i = \min(T_i, C_i) = \min(T_i^*, C_i^*) + \epsilon_i = \min(T_i^* + \epsilon_i, C_i^* + \epsilon_i) \tag{4.16}$$

where T_i, C_i are the observed failure/censoring times and T_i^*, C_i^* are the true failure/censoring times, respectively, since as discussed in Chapter 3, the censoring indicator δ_i is assumed to be unaffected by measurement error.

From 4.16, when a subset of the observed failure times are right-censored, as in the non-censored case, we first estimate the density of the contaminated failure time random variable using the observed contaminated right-censored failure time data and then apply the deconvolution operation to remove the contamination.

Observe that the estimator 4.12 can be interpreted as a simple weighted average of kernels with the weight of $\frac{1}{n}$ for each observed failure time T_i. When a subset of all T_i are right-censored, we need to adjust the weights in 4.12 to account for the fact that not all event times are failure times. Given the observed data $Y_i = \min_i(T_i^* + \epsilon_i, C_i^* + \epsilon_i)$, $i \in \{1, 2, ..., n\}$, a natural approach is to use weights that are dependent on the "jump

sizes" of the *Kaplan-Meier* estimator of the survivor function:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{r_i}\right) \tag{4.17}$$

where $t_1 < t_2 < ... < t_m$ are the observed (with error) distinct failure times, r_i is the number of individuals at risk up to time t_i and d_i are the number of individuals who fail at time t_i. Thus, for ordered right-censored failure time data $Y_i = \min_i(T_i, C_i), i \in \{1, 2, ..., n\}$, we define the weights of the kernel density estimator as follows:

$$s_j = \begin{cases} \hat{S}(Y_j) - \hat{S}(Y_{j+1}), & \text{if } j = 1, ..., n-1 \\ \hat{S}(Y_n), & \text{if } j = n. \end{cases} \tag{4.18}$$

Since the Kaplan-Meier estimator is a non-increasing piecewise constant function, the weights in 4.18 are non-zero if and only if at least one failure is observed to occur between the sequential observations of Y_j for $j \in \{1, 2, ..., n-1\}$. This choice of weights ensures the kernel density estimator is only fit to the failure time data with a non-zero weight based on whether a failure was actually observed. Note that if the final observation, Y_n, is censored, the Kaplan-Meier estimator evaluated at that point, $\hat{S}(Y_n)$, is not defined. To ensure the weighted kernel density estimator for the contaminated failure time random variable $T_i^* + \epsilon_i$ is well-defined, we assume for simplicity in all further discussion that the final observation of Y_n is a failure time. Using the weights of 4.18, we define the weighted kernel density estimator for the density of $T_i^* + \epsilon_i$ by:

$$\hat{f}_{T^*+\epsilon}(x) = \frac{1}{h} \sum_{j=1}^{n} s_j K\left(\frac{x - Y_j}{h}\right). \tag{4.19}$$

where $Y_j, j \in \{1, 2, ..., n\}$ are the ordered observed failure/censoring times [3]. For a discussion on the choice of kernel function and the optimal choice of bandwidth, see [3] [34]. We do not consider these topics in this thesis as the main focus is the development of the estimation methodology. Using this estimator for the density of $T_i^* + \epsilon_i$, we substitute

this expression into equation 4.15 in place of $\hat{f}_Y$ to obtain a non-parametric density estimator for f_{T^*}.

In a prevalent cohort study setting, using the model given in 3.5, where the observed failure times are left-truncated and right-censored, we further adjust the kernel density estimator given in 4.19 by replacing the weights provided by the Kaplan-Meier estimator with those of the *Lynden-Bell* estimator:

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{\gamma_{1i}}{\gamma_{2i}}\right) \tag{4.20}$$

where $t_{(1)} < t_{(2)} < ... < t_{(m)}$ are the observed failure times, γ_{1i} is the number of individuals observed to fail at time $t_{(i)}$ and $\gamma_{2i} = \#\{j : w_j \leq t_{(i)} \leq y_j\}$ for $i \in \{1, 2, ..., n\}$ where w_j is the truncation time and y_j is the observed failure/censoring time (i.e. γ_{2i} is the number of subjects under follow-up at risk of failing up to time y_j) [28]. That is, in 4.18, we exchange the Kaplan-Meier estimator with the Lynden-Bell estimator. As in the right-censored case, we first estimate the density of the contaminated failure time random variable $T_i + \epsilon_i$ using instead the Lynden-Bell-weighted kernel density estimator and then apply the deconvolution procedure as described in 4.15. Since the underlying failure time random variable T_i is non-negative, the kernel density estimator must have non-negative support. In order to accomodate this constraint, we may apply a simple boundary correction at 0 as given in [22]. Kernel boundary corrections use linear or quadratic approximations of the given kernel, K, near the boundary while maintaining the integration to 1 property of the resulting density estimator. An example of a linearly corrected boundary kernel is the following:

$$K_L(x) = \frac{(a_2(p) - a_1(p))K(x)}{a_0(p)a_2(p) - a_1^2(p)} \tag{4.21}$$

where $K(\cdot)$ is the uncorrected kernel and $a_l^j(p) = \int_{-1}^{p} x^l K^{(j)}(x)dx$. For a listing of various boundary corrected kernels and their respective properties, see [22].

Now, as discussed in Chapter 3 for incident and prevalent cohort studies, assumptions on the support of the error distribution relating the observed/unobserved onset date are equivalent to assumptions on the type of censoring of the observed onset date. For example, the truncated Normal error distribution assumed by Zhong and Cook implies in a prevalent cohort setting that all observed incidence dates are left-censored by prevalence day. In Section 4.3, in the setting of a prevalent cohort study, we examine a non-parametric method used to estimate the cdf of the underlying failure time random variable when the initial event times and terminating event times may both be interval-censored. This method suggests an alternative approach to survival analysis with uncertain initiation dates.

4.3 A Non-parametric Discrete Weight Likelihood Method

In [40], Turnbull developed a non-parametric estimation algorithm for the survivor function of i.i.d. failure time random variables subject to arbitrary truncation and censoring. That is, he assumed the failure time T_i is truncated by the interval B_i (i.e. T_i is unobserved if $T_i \notin B_i$) and is interval-censored by the interval $A_i \subset B_i$ for all $i \in \{1, 2, ..., n\}$. He showed that by discretizing the failure time random variables, it is possible to estimate the survivor function through a so-called "self-consistency" algorithm. As in [40], we define an unknown parameter/function θ as *self-consistent* if $\theta = \pi(\theta)$ for some function π and we define a *self-consistent* estimate of θ as any solution to the equation $\theta = \pi(\theta)$. A self-consistent algorithm attempts to find the fixed point by recursively refining the estimate of θ through an iterative updating procedure. Specifically, if θ^k is an estimate of θ at the k^{th} iteration, we obtain the updated estimate,

denoted by θ^{k+1}, through π such that $\theta^{k+1} = \pi(\theta^k)$.

Now, for a general estimation problem, the statistical importance of a fixed point solution for a given function π is unclear. Moreover, the selection of the function π, such that the resulting fixed point estimator has meaningful statistical properties, is not obvious. In [40], Turnbull remarked that his self-consistency algorithm can be regarded as a special example of the well known EM algorithm. Turnbull used the "complete dataset" principle of the EM algorithm and defined two indicator functions in which one was observable and the other unobservable such that both their expectations were observable. Based on the observed expectations, Turnbull derived an update function π. He subsequently showed that solving the equation $\theta = \pi(\theta)$ was equivalent to finding the maximum likelihood estimator for abitrarily censored and truncated failure time random variables. Thus, the use of Turnbull's self-consistency estimation algorithm is justified as it provides an easily implementable method of obtaining the maximum likelihood estimator.

The title of Turnbull's article ("The Empirical distribution Function with Arbitrarily Grouped, Censored and Truncated Data") from [40] is perhaps misleading. The proposed self-consistency algorithm only applies to failure time data in which the terminating event is interval-censored and interval-truncated. In [10], Lagakos and De Gruttola applied Turnbull's self-consistency algorithm to estimate the survivor function for failure time data that are doubly interval-censored (i.e. the onset and failure dates are both interval-censored). In [38], Sun adapted Turnbull's self-consistency algorithm for failure time data that are both doubly interval-censored and left-truncated. Lagakos and De Gruttola, and Sun, derived update functions π, for their respective types of failure time data, and used Turnbull's self-consistency algorithm to obtain estimates for the unknown survivor

functions. As in [40], for Lagakos and De Gruttola, and Sun, the choice of π was most likely motivated by the derivation of the respective likelihood score equations.

In Chapter 3, in a prevalent cohort study setting, we showed that specific error distribution assumptions on the onset dates in 3.5 and 3.6 are equivalent to the failure time data being left-truncated and doubly interval-censored. Thus, it follows that Sun's adaptation of Turnbull's self-consistency algorithm is directly applicable to this specific type of failure time data. In this section, we review how Sun adapted Turnbull's estimation procedure by first presenting the algorithm in full and then validating its use through the derivation of the likelihood score equations.

Let V_{0i} be the calendar date of disease onset, V_{1i} be the calendar date of the failure event, $S_i = V_{1i} - V_{0i}$ be the survival time of interest, and let $F(s)$ and $f(s)$ and $H_i^*(x)$ and $h_i^*(x)$ be the cumulative distribution functions and corresponding density functions of S_i and V_{0i} respectively for $i \in \{1, 2, ..., n\}$. We assume that S_i is independent of V_{0i} (i.e. survival is independent of the date of onset, that is, there is no "cohort effect" on survival) and that V_{0i}, V_{1i} and S_i take discrete values. For left-truncated doubly interval-censored failure time data, we assume the end event date V_{1i} is truncated on $B_i = [B_i^1, B_i^2]$ (i.e. if $V_{1i} \notin B_i$ then the subject is unobserved) where for non-truncated failure times, we observe V_{0i} in the interval $[E_i, R_i]$ and V_{1i} in the interval $A_i = [L_i, U_i] \subset B_i$. Under this setup, we derive the conditional likelihood (i.e. conditional on the inclusion in the prevalent cohort) based on the probabilistic contributions of observing V_{0i} in $[E_i, R_i]$ and V_{1i} in A_i. Let $h_i(x) = \frac{h_i^*(x)}{\sum_{x \in [E_i, R_i]} h_i^*(x)}$ denote the conditional probability density function of the onset times, conditional on $V_{0i} \in [E_i, R_i]$. For subject i observed to fail in A_i with observed onset in $[E_i, R_i]$, the contribution to the conditional likelihood is given by:

$$\mathbb{P}(V_{1i} \in [L_i, U_i], V_{0i} \in [E_i, R_i] | V_{1i} \in [B_i^1, B_i^2])$$

$$= \frac{\mathbb{P}(V_{1i} \in [L_i, U_i], V_{0i} \in [E_i, R_i], V_{1i} \in [B_i^1, B_i^2])}{\mathbb{P}(V_{1i} \in [B_i^1, B_i^2])}$$

$$= \frac{\sum_{x \in [E_i, R_i]} \mathbb{P}(V_{1i} \in [L_i, U_i] | V_{1i} = x) h_i(x)}{\sum_{x \in [E_i, R_i]} \mathbb{P}(V_{1i} \in [B_i^1, B_i^2] | V_{1i} = x) h_i(x)}$$

$$= \frac{\sum_{x \in [E_i, R_i]} \mathbb{P}(V_{1i} \in [L_i - x, U_i - x]) h_i(x)}{\sum_{x \in [E_i, R_i]} \mathbb{P}(V_{1i} \in [B_i^1 - x, B_i^2 - x]) h_i(x)}$$

$$= \frac{\sum_{x \in [E_i, R_i]} (F(U_i - x) - F((L_i - x)-)) h_i(x)}{\sum_{x \in [E_i, R_i]} (F(B_i^2 - x) - F((B_i^1 - x)-)) h_i(x)}$$

Therefore, because of between-subject independence, the conditional likelihood is given by:

$$L = \prod_{i=1}^{n} \frac{\sum_{x \in [E_i, R_i]} (F(U_i - x) - F((L_i - x)-)) h_i(x)}{\sum_{x \in [E_i, R_i]} (F(B_i^2 - x) - F((B_i^1 - x)-)) h_i(x)} \tag{4.22}$$

To apply Turnbull's self-consistency algorithm, we assume the $h_i(x)$ are known a priori and without loss of generality assume that the failure time data are centered such that $E_i = E$ for some constant E, $\forall i \in \{1, 2, ..., n\}$. Under these assumptions, the observed data have the form: $\{[E, R_i], A_i = [L_i, U_i], B_i = [B_i^1, B_i^2]\}$ for $i \in \{1, 2, ..., n\}$. To simplify 4.22, let $u_0 = 0 < u_1 < u_2 < ... < u_m$ denote the distinct finite ordered values of $\{0, L_i - R_i, U_i - E, ...\}$, $\forall i \in \{1, 2, ..., n\}$. These ordered values represent the collection of all the distinct smallest/largest possible finite failure time intervals for a given dataset.

Figure 4.1 is a graphical representation of the possible values for a particular subject. Using this discretization, it follows that 4.22 only depends on F evaluated at the u_j values and not on values between them [40]. Since the observation of the discretized values depends on whether they are observed in particular intervals, let α_{ij} be the indicator function of the event $u_j \in [L_i - R_i, U_i - E]$ and β_{ij} be the indicator function of the event $u_j \in [B_i^1 - R_i, B_i^2 - E]$ with $f_j = F(u_j) - F(u_{j-1})$. Using these functions, the likelihood

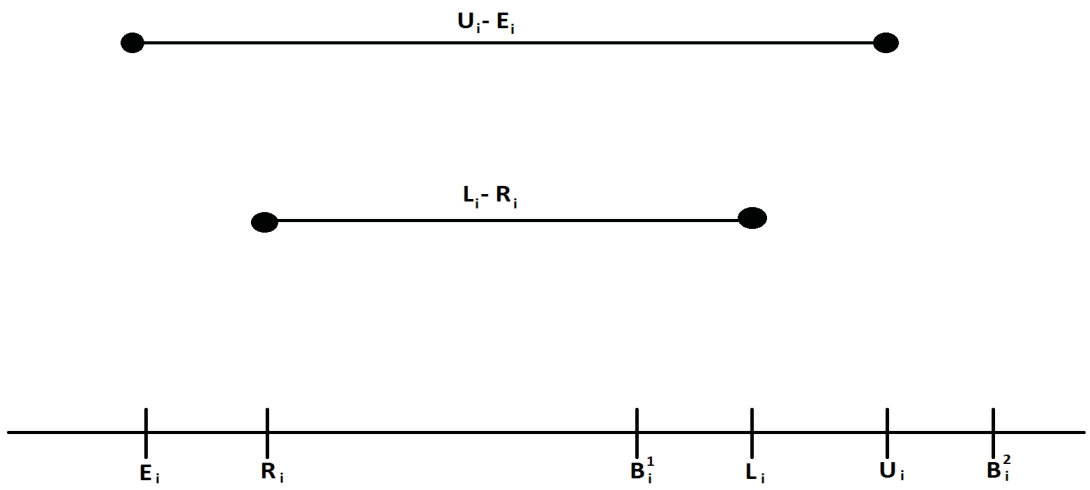

Figure 4.1 – *Representation of values ($U_i - E_i$ and $L_i - R_i$) for subject i in Sun's self-consistency algorithm. Sun's algorithm uses the collection of these values over all subjects as the discrete support of the probability mass function of the underlying failure time random variable.*

in 4.22 simplifies to

$$L = \prod_{i=1}^{n} \frac{\sum_{j=1}^{m} \alpha_{ij}\alpha_{ij}^{*} f_j}{\sum_{j=1}^{m} \beta_{ij}\beta_{ij}^{*} f_j} \tag{4.23}$$

where

$$\alpha_{ij}^{*} = \begin{cases} \sum_{x:x+u_j \in A_i} h_i(x), & \text{if } u_j \in [L_i - R_i, U_i - E] \\ 1, & \text{otherwise.} \end{cases} \tag{4.24}$$

and

$$\beta_{ij}^{*} = \begin{cases} \sum_{x:x+u_j \in B_i} h_i(x), & \text{if } u_j \in [B_i^1 - R_i, B_i^2 - E] \\ 1, & \text{otherwise.} \end{cases} \tag{4.25}$$

In [13], Efron derived a self-consistent estimator for the survivor function for right-censored failure time data. He used an average of the observed number of failure times

and the expected number of failure times from the observed right-censored failure times up to predefined time points to estimate $S(t)$. In [40], Turnbull followed a similar approach as Efron by using the expected number of observed failure times (denoted by μ_{ij}) and the expected number of unobserved failure times (denoted by ν_{ij}) excluded by left-truncation to derive an update function π. We define the two expectations below:

$$\mu_{ij}(f) = E_f(I(u_j \in [L_i - R_i, U_i - E]))$$

and

$$\nu_{ij}(f) = E_f(\#\{\text{unobserved } u_j \in [L_i - R_i, U_i - E]\}).$$

Note that due to left-truncation in a prevalent cohort study, the event defined in $\nu_{ij}(f)$, (i.e. $\#\{\text{unobserved } u_j \in [L_i - R_i, U_i - E]\}$) is unobserved. However, the expectation of the event has a closed form which can be represented using the observed indicator functions α_{ij} and β_{ij} and the discretized quantities given in 4.24 and 4.25. As discussed above, this particular expectation can be interpreted as an "Expectation" step in the EM algorithm. The expectations $\mu_{ij}(f)$ and $\nu_{ij}(f)$ simplify to the following:

$$\mu_{ij}(f) = \frac{\alpha_{ij}\alpha^*_{ij}f_j}{\sum_{j=1}^m \alpha_{ij}\alpha^*_{ij}f_j} \text{ and } \nu_{ij}(f) = \frac{(1-\beta_{ij}\beta^*_{ij})f_j}{\sum_{j=1}^m \beta_{ij}\beta^*_{ij}f_j} \tag{4.26}$$

Using the functions in 4.26, we present the self-consistency algorithm from [38]:

1. Start with a given $f = (f_1, f_2, ..., f_m)$ such that $0 \leq f_j \leq 1$ and $\sum_{j=1}^m f_j = 1$
2. Calculate $\mu_{ij}(f) = \frac{\alpha_{ij}\alpha^*_{ij}f_j}{\sum_{j=1}^m \alpha_{ij}\alpha^*_{ij}f_j}$ and $\nu_{ij}(f) = \frac{(1-\beta_{ij}\beta^*_{ij})f_j}{\sum_{j=1}^m \beta_{ij}\beta^*_{ij}f_j}$
3. Calculate $\pi_j(f) = \frac{1}{M(f)} \sum_{i=1}^n (\mu_{ij}(f) + \nu_{ij}(f))$ where $M(f) = \sum_{i=1}^n \sum_{j=1}^m (\mu_{ij}(f) + \nu_{ij}(f))$
4. Update f by the following: $f_j^{k+1} = \pi_j(f^k)$ where f^k is the value of f at the k^{th} iteration of the self-consistency algorithm

5. Iterate through steps 2-4 until convergence: $\sum_{j=1}^{m} |f_j^{k+1} - f_j^k| < \epsilon$ for some convergence criterion value $\epsilon > 0$.

6. After convergence $\hat{F}(s) = \sum_{j:u_j \leq s} \hat{f}_j$

As explained by Turnbull in [40] and Sun in [38], the self-consistency algorithm is equivalent to maximum likelihood estimation. The reasoning is as follows. Using 4.23, the log-likelihood is given by:

$$\log(L) = \sum_{i=1}^{n} \left(\log \left(\sum_{j=1}^{m} \alpha_{ij} \alpha_{ij}^* f_j \right) - \log \left(\sum_{j=1}^{m} \beta_{ij} \beta_{ij}^* f_j \right) \right).$$

By differentiating with respect to the component f_j, we obtain:

$$\frac{\partial \log(L)}{\partial f_j} = d_j(f) = \sum_{i=1}^{n} \left(\frac{\alpha_{ij} \alpha_{ij}^*}{\sum_{j=1}^{m} \alpha_{ij} \alpha_{ij}^* f_j} - \frac{\beta_{ij} \beta_{ij}^*}{\sum_{j=1}^{m} \beta_{ij} \beta_{ij}^* f_j} \right)$$

Observe that we can express $\pi_j(f)$ from step 3 of the algorithm above using the function $d_j(f)$ as follows:

$$\begin{aligned} \pi_j(f) &= \frac{f_j}{M(f)} \left(d_j(f) + \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{m} \beta_{ij} \beta_{ij}^* f_j} \right) \\ &= f_j \left(\frac{d_j(f)}{M(f)} + 1 \right) \end{aligned}$$

Thus, f is a maximum likelihood estimate if and only if $d_j(f) = 0$ or $d_j(f) \leq 0$ with $f_j = 0$ for each $j \in \{1, 2, ..., m\}$. The former equality implies that $\pi_j(f) = f_j$ which proves the equivalence between the self-consistency algorithm and maximum likelihood estimation. In Chapter 5, we use simulated prevalent cohort failure time data with the inclusion of measurement error to examine the performance of the non-parametric procedures discussed in this chapter.

Chapter 5

Simulations

In this chapter, we use simulated prevalent cohort failure time data to assess the impact of measurement error on the standard survival analysis modelling techniques and to examine the modelling performance of the non-parametric estimation methods discussed in Chapter 4. In section 5.1, we describe the simulation procedure used to generate left-truncated right-censored failure time data with the inclusion of measurement error. In section 5.2, we examine how the standard Lynden-Bell estimator is affected when it is fit to left-truncated right-censored failure time data with the inclusion of measurement error in the reported onset date. We evaluate the modelling performance of the deconvolution density estimation technique and the numerical estimation procedure of Sun in sections 5.3 and 5.4 respectively.

5.1 Prevalent Cohort Data Simulation Procedure

To generate simulated prevalent cohort failure time data with the inclusion of measurement error, we randomly sampled observations from an onset distribution, a fail-

ure time distribution, a censoring time distribution and an error distribution. For simplicity, in all simulations in sections 5.2 to 5.4, we assumed a fixed prevalence date of $R = 100$. To generate a sample of size n of left-truncated right-censored failure times, we first sampled a single onset date from either a $Unif(95, 100)$ distribution, or from an $\chi \sim Exponential(0.2)$ distribution such that $R - \chi$ is the sampled onset date. The former choice of onset distribution corresponds to a stationary onset process and the latter simulates the effect that subjects are more likely to obtain onset of "the disease" closer to the calendar date R. We provide a diagram of the exponential onset distribution in Figure 5.1.

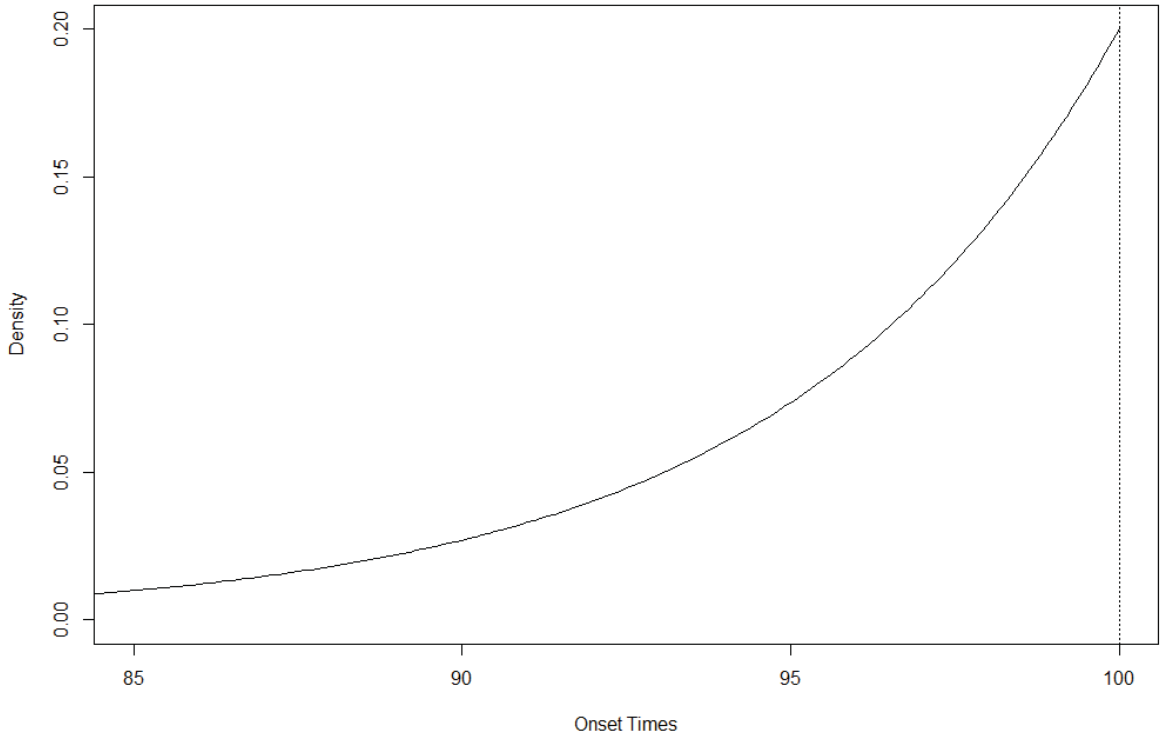

Figure 5.1 – *Exponential(0.2) onset density to the left of prevalence day ($R = 100$)*

For each sampled onset date, we sampled a failure time from a $Weibull(shp, scl)$ distribution and added its value to the sampled onset date. If the sum was greater than

R, the sampled onset date and failure time were kept in the sample, otherwise, they were both discarded. The action of retaining or discarding a given sample point depending on whether the sum surpassed the date R, simulates the effect of left-truncation. We continued this procedure of sampling an onset date and adding a failure time until we obtained a sample of n observations in our prevalent cohort.

Given the sample of n onset dates and corresponding failure times, we administratively censored the failure times by a constant value of c to allow for approximately 30% censoring. That is, if the calendar date of failure was greater than c, the sample point was denoted as censored with corresponding censoring time equal to $c - v_{0i}$ where v_{0i} is the sampled onset date. After censoring, the simulated sample comprised the observed onset dates, failure/censoring times and a censoring indicator vector. To simulate the effect of the measurement error, we added either a $Normal(0, \sigma^2)$ error to the sampled onset date or we subtracted an $Exponential(\psi)$ error from the sampled onset date. In the case of the Normal error, if the perturbed onset was greater than R, we set the observed onset date equal to R as we assume each subject in the simulated prevalent cohort was screened positive on date R. Unlike in the simulations of Zhong and Cook, the error distribution is not a truncated Normal distribution but rather a continuous error distribution with a discrete non-zero probability component at time R. The exponential error simulated the scenario in which subjects report a date which occurs prior to their true date of onset. In sections 5.2 to 5.4, we used different combinations of onset, failure, and error distribution while varying the given parameters to examine the effect of measurement error on estimation procedures and the performance of the non-parametric estimation techniques.

5.2 Impact of Measurement Error on Estimation in a Prevalent Cohort Study

In the parametric model setting of [46], Zhong and Cook used simulated prevalent cohort failure time data to examine how the inclusion of measurement error affected the estimates of the unknown parameters. We followed a similar simulation approach by empirically examining how the fit of the non-parametric Lynden-Bell estimator changed depending on whether it was fit to data with or without the inclusion of measurement error. In Figures 5.2 and 5.3, we plotted the Lynden-Bell estimators for 250 simulation runs for prevalent cohort study sample sizes of 1000 observations with/without error to examine how the inclusion of measurement error affected the variance and bias of the estimators.

In Figure 5.2, using a $Normal(0, 5)$ error distribution, the variance of the Lynden-Bell estimator appeared to be essentially unchanged with the inclusion of measurement error, however, the estimator appeared biased. Recall that the Normal error distribution for the onset dates is truncated by the prevalence date R which implies the expected value of the error term is non-zero. The Lynden-Bell estimator underestimated survival for shorter failure times as it overadjusted for survival for longer failure times. The adjustment mechanism of the Lynden-Bell estimator did not correct for the inclusion of measurement error. Thus, on average, the Lynden-Bell estimator either over/under fitted the true underlying failure time survivor function. In Figure 5.3, using an $Exponential(0.5)$ error distribution, the Lynden-Bell estimator fitted to the failure time data with measurement error appeared both biased and with larger variance as compared to the estimator without

the inclusion of the measurement error. Since the expectation of the exponential error is non-negative, the Lynden-Bell estimator over estimated the probability of survival. The sudden drop in some of the Lynden-Bell estimates in Figure 5.3 is attributed to the lengthening of the truncation times by the Exponential error distribution. Specifically, for left-truncated right-censored failure time data, if the backward recurrence times are very long relative to the forward recurrence times, the Lynden-Bell estimate may drop near 0 for shorter failure times. Pan and Chappell proposed a solution to account for the survival estimate drop in [32]. The correction, however, does not account for the effect of measurement error in the onset dates.

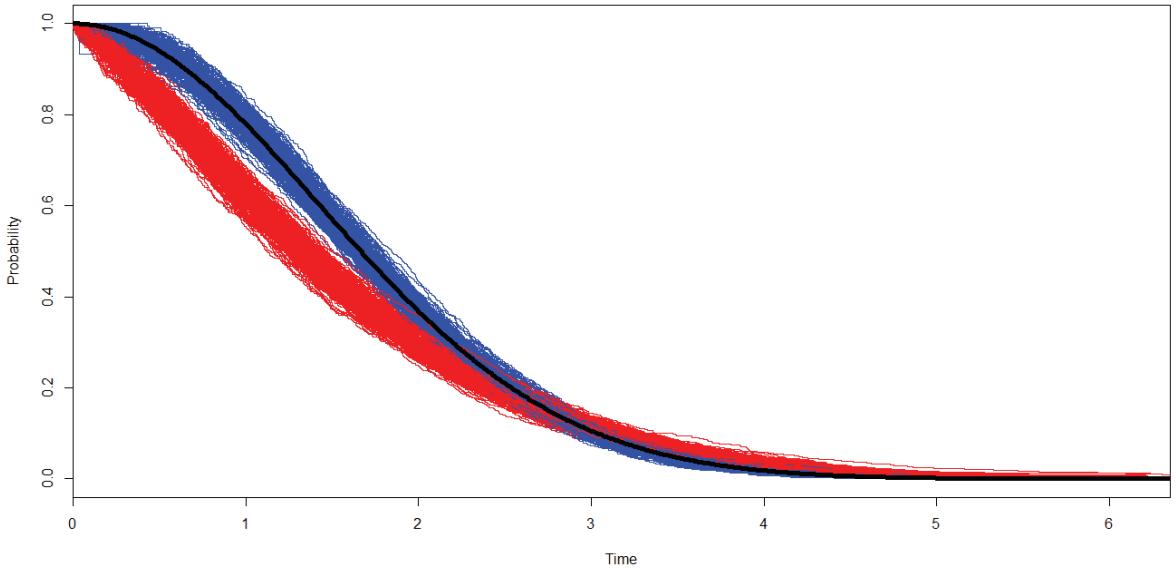

Figure 5.2 – *Graphical comparison of the Lynden-Bell estimator fit to failure time data with/without the inclusion of measurement error (represented by the red curves and blue curves, respectively) to the true failure time survivor function (black curve). The simulated prevalent cohort data follow a uniform onset distribution with* $Weibull(2.0, 2.0)$ *failure times with the addition of a* $Normal(0, 5)$ *error distribution to the observed onset dates. Number of simulation runs = 250. Sample size = 1000.*

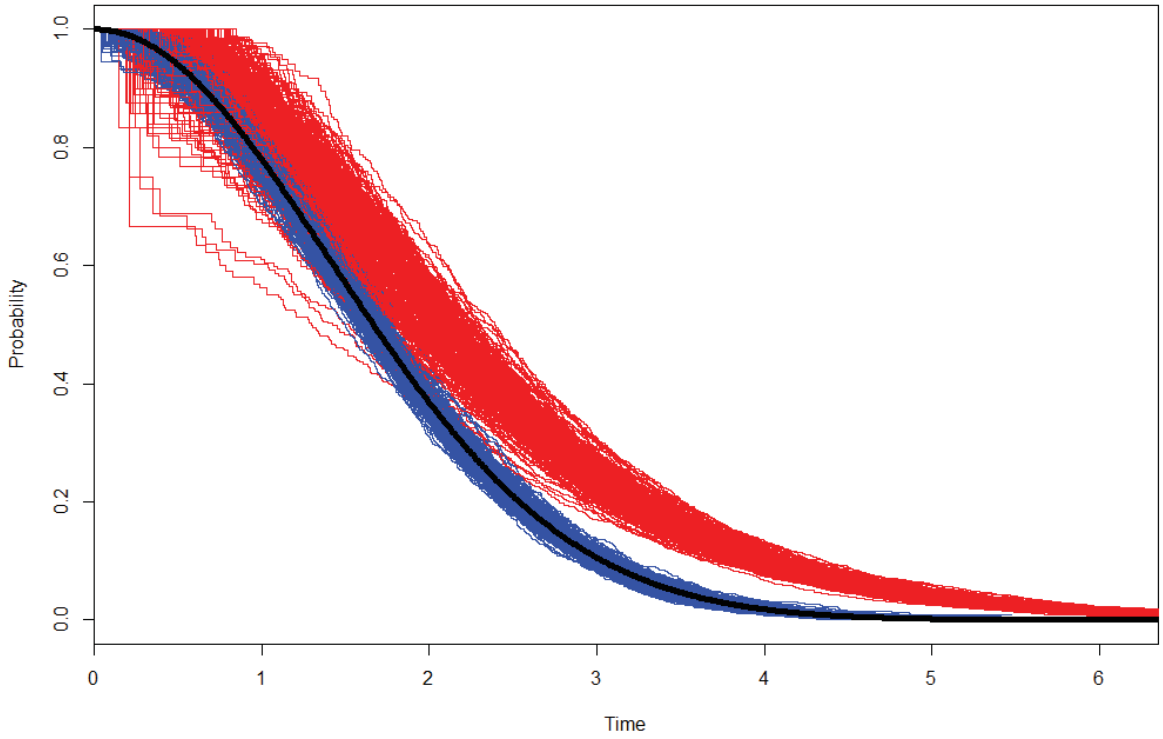

Figure 5.3 – *Graphical comparison of the Lynden-Bell estimator fit to failure time data with/without the inclusion of measurement error (represented by the red curves and blue curves, respectively) to the true failure time survivor function (black curve). The simulated prevalent cohort data follow a uniform onset distribution with* $Weibull(2.0, 2.0)$ *failure times with the addition of an* $Exponential(0.5)$ *error distribution to the observed onset dates. Number of simulation runs = 250. Sample size = 1000.*

To numerically examine the effect of the measurement error on survival estimation, we compared the average median values calculated from the Lynden-Bell estimator for various combinations of onset, error and failure time distributions. We listed the average median values, calculated over 1000 simulation runs with sample sizes of 500 observations, in Table 5.1. For the distribution combinations using a Normal error distribution and either Weibull(1.0,1.0) or Weibull(1.0,2.0) failure time distributions, we did not ob-

serve large differences in the average estimated medians as compared to the true median values. This lack of change in the estimated medians may have been attributable to the underlying shape of the failure time distributions, as shorter failure times were more likely to be sampled. Thus, as described above, for failure/censoring times with short backward recurrence times, there was an approximate 0.5 probability that the Normal error distribution only perturbed the onset dates to R yielding very little difference between the true and observed failure/censoring times. However, this phenomenon did not occur for the Weibull(2.0, 2.0) failure time distribution as the underlying distribution is mound shaped about 1 with a short tail near 0. In these cases, the Normal error distributions perturbed the onset dates closer to (or exactly to) the date R thus shortening the observed failure/censoring times yielding a significant underestimation of the median. For the exponential error distributions, we found that the average median value with the inclusion of measurement error tended to overestimate the true median value. As discussed above, since the exponential error distribution, on average, lengthened the true underlying failure times, the calculated medians with the inclusion of measurement error were larger than those calculated without error.

5.3 Non-Parametric Deconvolution Estimation

To examine the modeling performance of the deconvolved kernel-density estimator from 4.15, we compared it to the true underlying failure time density using the code of Delaigle from [11]. We also fitted the weighted kernel-density estimator, from 4.19, to the observed censored/failure times (with the inclusion of measurement error) and compared its fit to the true underlying failure time density. We denote this estimator

		Weibull Failure Time Distribution Parameters					
		(1.0, 1.0)		(1.0,2.0)		(2.0,2.0)	
	True Median	0.693		1.386		1.665	
Onset Distribution	Error Distribution	With Error	Without Error	With Error	Without Error	Wtih Error	Without Error
Uniform	Normal(0,1)	0.691	0.692	1.389	1.397	1.324	1.666
	Normal(0, 3)	0.692		1.392		1.076	
	Exponential(0.5)	0.784		1.496		2.029	
	Exponential(1.0)	0.756		1.451		1.983	
Negative Exponential	Normal(0, 1)	0.692	0.688	1.381	1.375	1.356	1.661
	Normal(0.3)	0.693		1.387		1.130	
	Exponential(0.5)	0.756		1.478		2.029	
	Exponential(1.0)	0.728		1.450		1.982	

Table 5.1 – *Average median values calculated from Lynden-Bell estimators fit with/without the inclusion of measurement error. Number of simulation runs = 1000. Sample size = 500.*

as the weighted "naive" kernel-density estimator as it does not account for the inclusion of the random measurement error. The bandwidth was selected using the R "density" function. To measure the fit of the estimators, we defined a mesh of sample points and calculated the average absolute differences between the estimators compared to the true underlying density evaluated at these points. We reported the average absolute error values for prevalent cohort study sample sizes of 500 observations over 1000 simulation runs for the naive and deconvolved density estimators in Table 5.2.

From the values listed in Table 5.2, we found that the naive kernel-density estimator provided a better fit to the underlying true density for the Weibull failure time distributions with shape parameter equal to 1. When the shape and scale parameters of the Weibull failure time distribution were both chosen to equal 2, the deconvolved kernel-density estimator outperformed the naive kernel-density estimator. These differences in performance may be attributable to the behaviour of the underlying failure time density at the boundary. As the estimated densities appeared to capture the overall shape of the underlying density function, the non-optimal selection of the bandwidth parameter has not greatly affected the fit of the estimator. That is, when the shape parameter is equal to 1, the underlying density function has non-negligible mass near 0 (i.e. there is no negligible tail for the density of the failure times arbitrarily close to 0) whereas when the shape parameter is 2, the majority of the density function is defined away from the boundary. Specifically, when the density of the observed failure/censoring times (with error) is "mound shaped", the deconvolution procedure cannot correct for the error such that the resulting estimate is not mound shaped. Due to this drawback, the deconvolved kernel-density estimator only outperformed the naive kernel-density estimator when the underlying failure time density was Weibull(2.0,2.0).

In addition, to implement the deconvolution procedure, we used discrete Fourier transforms and discrete inverse Fourier transforms to compute the density of the underlying failure time random variables from the observed sample data through the code in [11]. The discrete transforms are only approximations of the continuous Fourier transforms and so the resulting estimates were only approximations to the true underlying failure time densities. This approximation may account for the observed differences in the performance of the deconvolved estimator. In figure 5.4, for a sample size of 200000, we plotted the deconvolved kernel-density estimator against the naive kernel-density estimator and the underlying true failure time distribution. While the deconvolved estimator appeared to capture the same shape of the underlying failure time density, it was somewhat skewed to the right. This skew may be caused by the implementation of the deconvolution procedure using the discretized versions of the continuous transformations.

5.4 Left-Truncated Doubly Interval-Censored Non-Parametric Survival Estimation

As described in Section 4.3, there is an equivalence between the support of the error distribution for the observed dates and the type of censoring of the true unobserved onset dates. In this section, we examined the performance of the self-consistency estimator of Sun from [38] for left-truncated doubly-interval censored failure time data. Based on personal correspondence with Victor De Gruttola and Jianguo Sun, there is no open-source R code or R package to implement the self-consistency algorithm for (left-truncated) doubly-censored failure time data. The R package "survival" only allows for data that

		Weibull Failure Time (Shape, Scale) Parameters		
Onset Distribution	Exponential Error Parameter	(1.0, 1.0)	(1.0, 2.0)	(2.0, 2.0)
Uniform	1.0	(0.0424, 0.0281)	(0.0211, 0.0145)	(0.0292, 0.0322)
	0.5	(0.0352, 0.0213)	(0.0209, 0.0138)	(0.0163, 0.0229)
	0.1	(0.0117, 0.0102)	(0.0087, 0.0072)	(0.0078, 0.0104)
Left Exponential	1.0	(0.0437, 0.0278)	(0.0250, 0.0174)	(0.0303, 0.0327)
	0.5	(0.0353, 0.0212)	(0.0234, 0.0151)	(0.0169, 0.0235)
	0.1	(0.0116, 0.0109)	(0.00869, 0.00712)	(0.00815, 0.0106)

Table 5.2 – *Average L1 Error of naive kernel-densitty estimator and deconvolved kernel-density estimator compared to the true underlying failure time density. The error distribution acting on the true onset dates is Exponential(λ). Sample Size = 500, Number of Simulation Runs =1500. Pairs are in terms of (deconvolved kernel-density L1 error, naive kernel-density L1 error)*

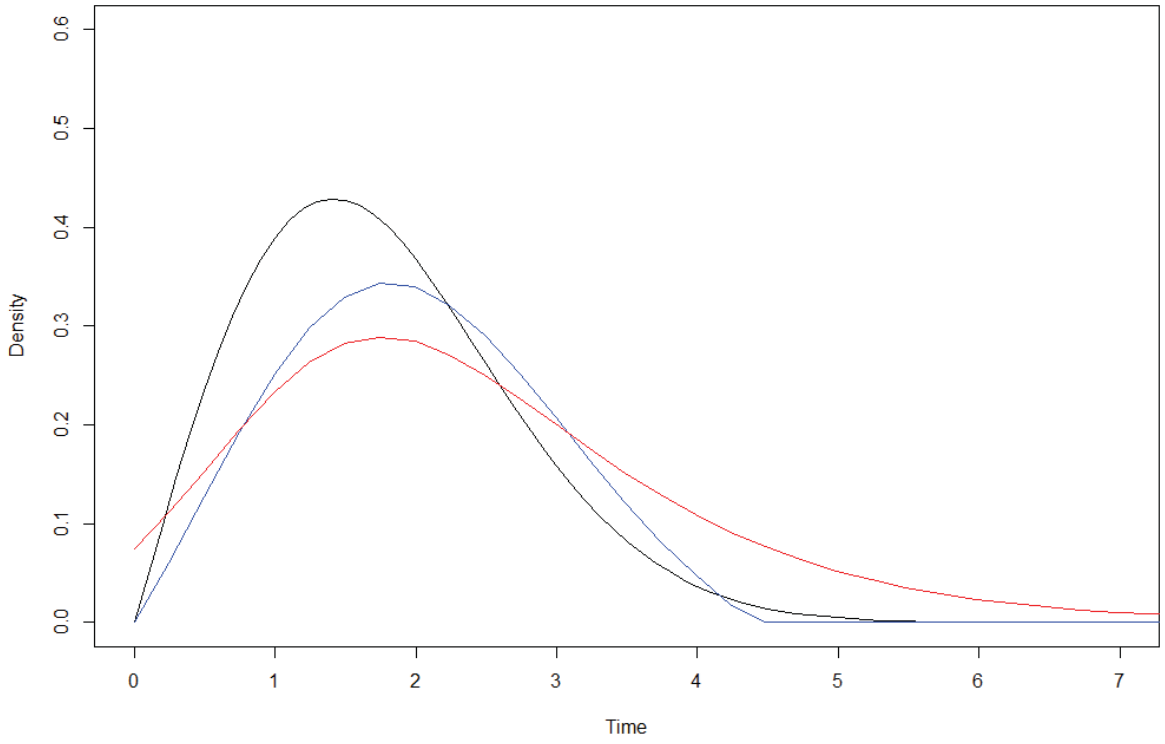

Figure 5.4 – *Graphical comparison of the weighted kernel-density estimator fit to failure time data with the inclusion of measurement error (red curve), the deconvolution kernel-density estimator (blue curve) to the true failure time survivor function (black curve). The simulated prevalent cohort data follow a uniform onset distribution with* $Weibull(2.0, 2.0)$ *failure times with the addition of an* $Exponential(0.1)$ *error distribution to the observed onset dates. Sample size = 200000.*

are truncated with one form of censoring (right/left) and the package "dblcens" only allows for alternatively censored failure time data (i.e. the failure times are either left-censored or right-censored but not doubly-censored). Neither Sun nor Lagakas and De Gruttolla carried out simulations to validate their proposed methods.

Using simulated prevalent cohort failure time data with the inclusion of measurement error, we compared the estimator of the survivor function derived using Sun's self-

consistency algorithm to the survivor function of the true underlying failure time random variables. Using a single sample of size 200 with an exponential error term, we empirically examined the difference between the Lynden-Bell estimator fit to the failure time data with the inclusion of (exponential) measurement error and the estimator using Sun's algorithm. Figure 5.5 displays this comparison. As expected, the ("regular") Lynden-Bell estimator overestimated the survivor function since the reported failure times were much longer (on average) than the true underlying failure times. We see that Sun's algorithm overestimated survival for shorter failure times (i.e. time lengths between 0 and 2) and was relatively accurate for longer failure times (i.e. time lengths larger than 2). Due to the exponential error, the censoring interval of each onset date is defined between the observed onset date and prevalence date R. From the defined discretized time intervals of Sun's algorithm, approximately half of the discretized points are equal to the forward recurrence times. The overestimation of survival, from Sun's algorithm, at shorter failure times was attributed to the large number of small discretized intervals based solely on the forward recurrence times.

In Table 5.3, we compared the fit of Sun's estimator of the survivor function and the Lynden-Bell estimator fit to the failure time data with error, to the true underlying survivor function. Using a mesh of points, we calculated the average absolute difference between the estimators and the true survivor function. In Sun's algorithm, because the number of parameters is approximately equal to the size of the sample, we used prevalent cohort study samples of size 15 over 500 simulation runs. It is not surprising therefore, that Sun's estimator did not perform better than the Lynden-Bell estimator when fit to the failure time data with measurement error. Based on the fit of Sun's estimator in Figure 5.4, for a single sample size of 200, the self-consistent estimator may provide a

better fit, on average, than the Lynden-Bell estimator for larger sample sizes.

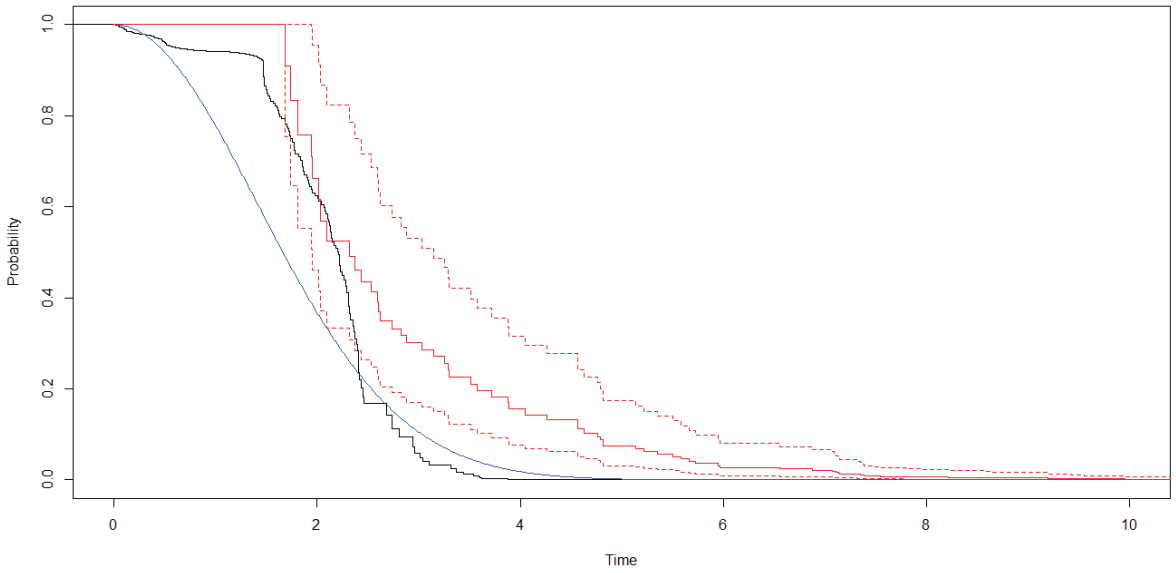

Figure 5.5 – *Graphical comparison of the Lynden-Bell estimator (with point-wise confidence intervals - dotted curves) fit to failure time data with the inclusion of measurement error (represented by the solid red curve) and Sun's nonparametric survivor function estimator (represented by the black curve) to the true failure time survivor function (blue curve). The simulated prevalent cohort data follow a uniform onset distribution with* $Weibull(2.0, 2.0)$ *failure times with the addition of an* $Exponential(0.1)$ *error distribution to the observed onset dates. Sample size = 200.*

Error Distribution	L1 error of Sun's Estimator	L1 error of Lynden-Bell Estimator
Exponential(1)	0.222	0.175
Exponential(0.5)	0.212	0.164
Exponential(0.1)	0.177	0.109

Table 5.3 – *Average L1 Error of the Lynden-Bell estimator and Sun's survivor function estimator compared to true underlying survivor function. The simulated prevalent cohort data follow a uniform onset distribution with* $Weibull(2.0, 2.0)$ *failure times, and with the addition of an* $Exponential(\psi)$ *error. Sample size = 15. Number of Simulation Runs = 500.*

Chapter 6

Discussion and Conclusions

The goal of this thesis was to present the problem of measurement error in failure time data and to review/propose various modelling techniques.

In Chapter 1, we provided the motivation for including measurement error in failure time data by examining the practical data collection problems that arose in the Canadian Study of Health and Aging. In Chapter 2, we reviewed the standard estimation techniques used in survival analysis for failure time data without the inclusion of measurement error. In Chapter 3, we examined the theoretical impact of measurement error in the context of an incident cohort study and a prevalent cohort study with follow-up. We showed that depending on the assumptions of the support of the error distribution, subjects can be incorrectly misclassified and excluded. We also showed, under particular error support assumptions, an equivalence between observing the true onset date with error and the true onset date being censored.

In Chapter 4, we examined three distinct modelling methods to incorporate measurement error in failure time data. In section 4.1, we reviewed the parametric estimation

method of Zhong and Cook. We argued that the proposed parametric likelihoods of Zhong and Cook appear to be incorrect and derived the correct parametric likelihood functions for the failure time parameters based on the observed failure/censoring times and the observed truncation times. In section 4.2, we proposed a new non-parametric density estimator for the true underlying failure time density with the inclusion of measurement error. In the framework of a classical measurement error model, we applied a deconvolution procedure on a weighted kernel-density estimator to obtain a non-parametric kernel-density estimator for the underlying failure time density. In section 4.3, using the equivalence between measurement error and censoring from Chapter 3, we reviewed the numerical procedure of Sun for left-truncated doubly interval-censored failure time data. Our goal was to investigate the usefulness of viewing the error-in-onset problem through an interval-censoring lens.

In chapter 5, we used simulated data to examine the effect of measurement error on the standard non-parametric Lynden-Bell survivor function estimator and to examine the modelling performance of the non-parametric methods discussed in chapter 4. We found that under certain parametric error distribution assumptions, the Lynden-Bell estimator, when fit to the failure time data with the inclusion of measurement error, has larger variance and is biased as compared to the Lynden-Bell estimator fit to the failure time data without measurement error. We found that the deconvolution procedure did not outperform the naive kernel-density estimator in all parametric settings due to estimation of the underlying density near the boundary as well as the implementation of the deconvolution procedure being based on the discrete approximation of the continuous Fourier transform. For the numerical method of Sun, we were only able to test its performance on simulated datasets with small sample sizes as the computational time

required for convergence of the self-consistency algorithm is dependent on the size of the dataset.

While the methods discussed in this thesis each provide possibly reasonable ways to model failure time data with error in the initiating event, they have their drawbacks. Zhong and Cook's approach needs to be revisited using the correct likelihood based on the observed/censored failure times and truncation times, as discussed in section 4.1. Additionally, more work is required on improving the implementation of the deconvolution kernel density estimator. While the method is theoretically sound, computational adjustments to Delaigle's deconvolution code from [11] are needed. Similarly, due to the computational costs of Sun's numerical estimation procedure, improvements on the implementation of the self-consistency algorithm are required. All discussion in this thesis is based on modelling procedures without the inclusion of observed covariates. It is an open question whether the models described in this thesis can be adapted to include covariates or covariates also observed with some form of measurement error.

Bibliography

[1] M. Asgharian, C. Wolfson, and D.B. Wolfson. Analysis of Biased Survival Data: The Canadian Study of Health and Aging and beyond. *Statistics in Action, A Canadian Outlook*, CRC, 2014.

[2] M. Asgharian, and D.B. Wolfson. Asymptotic Behavior of the Unconditional NPMLE of the Length-Biased Survivor Function from Right Censored Prevalent Cohort Data. *The Annals of Statistics*, 33 (5): 2109-2131, 2005.

[3] A. Berg, and D. N. Politis. Density Estimation of Censored Data with Infinite-Order Kernels. *arXiv:0704.3281v1 [math.ST]*, 2007.

[4] J. Berkson. Are There Two Regressions?. *Journal of the American Statistical Association*, 45 (250): 164-180, 1950.

[5] J.P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications.* Chapman & Hall/CRC, 2010.

[6] R.J. Carroll, D. Ruppert, L.A. Stefanski and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Chapman & Hall/CRC, 2006.

[7] G. Casella, R.L. Berger. *Statistical Inference, Second Edition.* Duxbury Advanced Series, 2002.

[8] D. Collett. *Modelling Survival Data in Medical Research, Second Edition.* Chapman & Hall/CRC, 2003.

[9] D.R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society - Series B*, 34 (2): 187-220, 1972.

[10] V. De Gruttola, and S.W. Lagakos. Analysis of Doubly-Censored Survival Data, with Applications to AIDS. *Biometrics*, 45 (1): 1-11, 1989.

[11] A. Delaigle. *Deconvolution R code* `http://www.ms.unimelb.edu.au/%7Eaurored/links.html#Code`

[12] A. Delaigle, and A. Meister. Density estimation with heteroscedastic error. *Bernoulli*, 14 (2): 562-579, 2008.

[13] B. Efron. The two sample problem with censored data. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 4: 831-853, 1967.

[14] B. Efron. Censored Data and the Bootstrap. *Journal of the American Statistical Association*, 76 (374): 312-319, 1981.

[15] R. Etzioni, and Y. Shen. Estimating Asymptomatic Duration in Cancer: The AIDS connection. *Statistics in Medicine*, 16 (6): 627-644, 1997.

[16] H. Frydman, T. Gerds, R. Gron and N. Keiding. Nonparametric estimation in an "illness-death" model when all transition times are interval censored. *Biometrical Journal*, 55 (6): 823-843, 2013.

[17] P. Groeneboom, G. Jongbloed. Density estimation in the uniform deconvolution model. *Statistica Neerlandica*, 57 (1): 136-157, 2003.

[18] D.W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data Second Edition*. Wiley Series in Probability and Statistics, John Wiley & Sons, 2008.

[19] N.P. Jewell, H.M. Malani, and E. Vittinghoff. Nonparametric Estimation for a form of Doubly Censored Data, with Application to Two Problems in AIDS. *Journal of the American Statistical Association*, 89 (425): 7-18, 2012.

[20] J.G. Ibrahim, M-H. Chen, D. Sinha. *Bayesian Survival Analysis*. Springer Series in Statistics, 2001.

[21] P. Joly, D. Commenges, C. Helmer and L. Letenneur. A penalized likelihood approach for an illness-death model with interval censored data: application to age-specific incidence of dementia. *Biostatistics*, 3 (3): 433-443, 2002.

[22] C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3 (3): 135-146, 1993.

[23] J.D. Kalbfleisch, R.L. Prentice. *The Statistical Analysis of Failure Time Data, Second Edition*. Wiley Series in Probability and Statistics, 2002.

[24] N.L. Komarova, and C.J. Thalhauser. High Degree of Heterogeneity in Alzheimer's Disease Progression Patterns. *PLOS Computational Biology* 7 (11): e1002251. doi:10.1371/journal.pcbi.1002251, 2011.

[25] S.W. Lagakos, and N. Reid. Estimating convolutions from partially censored data. *Biometrika*, 68 (1): 113-117, 1981.

[26] L. Lam. The Analysis of Doubly Censored Survival Data: An Application to Data Collected from the Amsterdam Cohort Studies on HIV Infection and AIDS. uuid:04474b33-542c-4d34-9821-cb672924a1f0, 1997.

[27] K.M.. Leung, R.M. Elashoff, and A.A. Afifi. Censoring Issues in Survival Analysis. *Annual Review of Public Health*, 18: 83-104, 1997.

[28] D. Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3CR Quasars. *Monthly Notices of the Royal Astronomical Society*, 155 (1): 95-118, 1971.

[29] E.L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53 (282): 457-481, 1958.

[30] I. McDowell. Canadian Study of Health and Aging: study methods and prevalence of dementia. *Canadian Medical Association Journal*, 150 (6): 899-913, 1994.

[31] A. Meister. Deconvolution Problems in Nonparametric Statistics. *Lecture Notes in Statistics* 193, Springer-Verlag, 2009.

[32] W. Pan, and R. Chappell. A Nonparametric Estimator of Survival Functions for Arbitrarily Truncated and Censored data. *Lifetime Data Analysis*, 4 (2): 187-202, 1998.

[33] F. Rouah, and C. Wolfson. A Recommended Method for Obtaining the Age at Onset of Dementia From the CSHA Database. *International Psychogeriatrics*, 3, Supp 1: 57-70, 2001.

[34] C. Sánchez-Sellero, W. Gonzalez-Manteiga, and R. Cao. Bandwidth Selection in Density Estimation with Truncated and Censored Data. *Annals of the Institute of Statistical Mathematics*, 51 (1): 51-70, 1999.

[35] S.J. Sheather. Density Estimation. *Statistical Science*, 19 (4): 588-597, 2004.

[36] B.W. Silverman. Density Estimation for Statistics and Data Analysis. *Monographs on Statistics and Applied Probability*, Chapman & Hall, 1986.

[37] N. Sourial. Impact of Uncertainty in the Onset Time of Disease on the Estimation of the Survival Function. *Master of Science Thesis, McGill University*, 2003.

[38] J. Sun. Empirical Estimation of a Distribution Function with Truncated and Doubly Interval-Censored Data and Its Application to AIDS Studies. *Biometrics*, 51 (3): 1096-1104, 1995.

[39] C.J. Thalhauser, and N.L. Komarova. Alzheimer's disease: rapid and slow progression. *Journal of the Royal Society Interface*, 9 (66): 119-126, 2012.

[40] B.W. Turnbull. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society - Series B (Methodological)*, 38 (3): 290-295, 1976.

[41] W.Y. Tsai, N. P. Jewell and M-I. Wang. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*. 74 (4): 883-886, 1987.

[42] M.C. Wang. Nonparametric Estimation from Cross-Sectional Survival Data. *Journal of the American Statistical Association*, 86 (413): 130-143, 1991.

[43] X-F. Wang, and B. Wang. Deconvolution Estimation in Measurement Error Models: The R Package decon. *Journal of Statistical Software*, 39 (1): i10, 2011.

[44] X.F. Wang, and D. Ye. Conditional density estimation in measurement error problems. *Journal of Multivariate Analysis*, 133: 38-50, 2015.

[45] C. Wolfson, D. Wolfson, M. Asgharian, C.E. M'Lan, T. Ostbye, K. Rockwood, D.B. Hogan. A Reevaluation of the Duration of Survival after the Onset of Dementia. *The New England Journal of Medicine*, 344: 1111-116, 2001.

[46] Y.Zhong, and R.J. Cook. Measurement Error for Age of Onset in Prevalent Cohort Studies. *Applied Mathematics*, 5 (11): 1672-1683, 2014.