Inference on Cure Rate Under Multivariate Random Censoring

Elnaz Ghadimi

A Thesis

In the Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Mathematics) at

Concordia University

Montréal, Québec, Canada

January 2018

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: Elnaz Ghadimi

Entitled: Inference on Cure Rate Under Multivariate Random Censoring

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Mathematics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

______________________________ Chair
Dr. Satyaveer Chauhan

______________________________ External Examiner
Dr. Masoud Asgharian

______________________________ Examiner
Dr. Prosper Dovonon

______________________________ Examiner
Dr. Frédéric Godin

______________________________ Examiner
Dr. Lisa Kakinami

______________________________ Thesis Supervisor
Dr. Arusharka Sen

Approved by ______________________________
Dr. Arusharka Sen, Graduate Program Director

March 28 2018 ______________________________
Dr. André Roy, Dean, Faculty of Arts and Science

Abstract

Inference on Cure Rate Under Multivariate Random Censoring

Elnaz Ghadimi, Ph.D.

Concordia University, 2018

In survival studies, it is often of interest to study cure rates. Sometimes the event of interest (such as death or the occurrence of a disease) may not be experienced by individuals under study, and the cure rate is the probability of the latter eventuality. Furthermore, survival times (i.e., the time to the event of interest) may be subject to random censoring due to dropping out or late entry of individuals during the study period. Interestingly, random censoring facilitates the estimation of the cure rate.

In this research, we study cure rates for multivariate survival times under multivariate random censoring. Specifically, three topics have been studied. In the first topic of this thesis, a new non-parametric multivariate cure rate estimator, based on a multivariate Kaplan-Meier estimator, is proposed. The asymptotic normality and an estimator of asymptotic variance of this estimator are obtained. In the second topic, a non-parametric cure rate estimator in the presence of covariates is constructed via kernel smoothing. The asymptotic normality of this estimator is obtained, and the optimal choice of the bandwidth via cross-validation is discussed. In the third topic of this thesis, we develop a test for the presence of immunes, i.e., we test if the cure rate is zero against the alternative that it is positive, under univariate random censoring. The limiting distribution of the test statistic under the null hypothesis is obtained using extreme-value theory. Theoretical results are supported by simulation studies.

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisor, Arusharka Sen, for his support, encouragement, valuable guidance and commitments during my PhD thesis. This thesis has benefited greatly from his insightful and endless support. His immense knowledge and inspiration helped me to overcome many difficulties during my research. I feel fortunate to have had the enlightening experience to work with him during my time at Concordia University.

I wish also to thank my committee members for their precious time, brilliant comments and insightful suggestions. Also thanks to the faculty members and staff who helped me whenever I approached them.

I would like to express my deepest gratitude and appreciation to my adored parents, my brother and my dear husband, Pooya, for their ultimate support, patience and encouragements. Their worthwhile support and love during my PhD thesis have encouraged me to do my best to achieve my goal. I do not know how I can express my appreciation to them for their ultimate support and grace with the words. I only hope that I have made their support worthwhile, and compensate some parts of their fortitude.

To my Mom, Dad, Brother and
my dear husband, Pooya
for their love, endless support and encouragement

Contribution of Authors

This thesis has been realized under the direction of Professor Arusharka Sen, professor at the Department of Mathematics and Statistics in Concordia University. It has been prepared in a “Manuscript-based” format. This research was financially supported by NSERC.

The thesis includes three manuscripts, co-authored by Professor Arusharka Sen. In all of the presented manuscripts, I have acted as the principal researcher and developed the mathematical models, identified the limiting distributions, conducted the simulations, analyzed the results, as well as having written of the first drafts of the articles. Professor Arusharka Sen has revised the thesis to obtain the final version prior to submission.

The first manuscipt entitled *’Multivariate cure rate estimation under random censoring’*, is co-authored with Professor Arusharka Sen.

The second manuscipt entitled *’Nonparametric multivariate cure rate estimation with covariates’*, is co-authored with Professor Arusharka Sen.

The third manuscipt entitled *’A nonparametric test for the presence of immunes in the univariate case’*, is co-authored with Professor Arusharka Sen.

Contents

List of Figures

List of Tables

Notation and abbreviations

EVD	Extreme Value Distribution
iid	Independent and Identically Distributed
iff	If and only if
KM	Kaplan-Meier
LRT	Likelihood Ratio Test
MLE	Maximum Likelihood Estimation
NPMLE	Non-Parametric Maximum-Likelihood Estimator

Chapter 1

Introduction

1.1 Background

The main aim of survival analysis is modeling and analyzing time-to-event data. Survival data consists of the occurrence of event of interest over time. The event of interest could be death, occurrence of a disease, re-arrest of a released prisoner, marriage, divorce, etc. The presence of immune individuals has an important role in the analysis of survival data. We define the cured or immune individuals as the ones who are not subject to the event of interest. Often our information on the survival times of individuals is not complete because there is censoring in the study. Censoring can happen due to loss of follow-up, drop out or termination of the study. There are different types of censoring such as right, left and interval censoring. Right censoring is the most common form of censoring. As an example of right censoring, we can refer to a patient who does not experience the event of interest during study or drops out of the study before the event could occur.

Left censoring happens when the event of interest has occurred before the entrance of the person in the study. For example, if a person is asked at what age he used marijuana first and he does not remember the exact age, such situation is considered as left censoring. A more general type of censoring is called interval censoring. It occurs when the event occurs within an interval where the individual is not observed. As an example, we can refer to the periodic follow-up of a patient in the study. It is known that the event time falls in an interval between two follow-ups. The Kaplan-Meier estimator provides a nonparametric estimate of distribution function of a time to event variable in the presence of independent right censoring. As illustrate in [9], the Kaplan-Meier or product limit estimator of the survival function at time $t \geq 0$ is

$$\hat{\bar{F}}(t) = \prod_{j=1}^{n} \left(1 - \frac{d_j}{n_j}\right)^{I(t_j \leq t)},$$

where n_j and d_j are, respectively, the number of individuals at risk and the number of those who fail at time t_j and $t_1 \leq ... \leq t_n$ are observed failure times. Maller and Zhou [11] proposed the tail (i.e., $\hat{\bar{F}}(t_n)$) of the Kaplan-Meier estimator as an estimator of cure rate in the univariate case.

1.2 Objectives

The main objective of this research work is to construct and analyze non-parametric cure rate estimators for *multivariate* survival times in the presence of multivariate random censoring. Many examples of multivariate survival times have been studied in the literature. For instance, Chen et al. [5] consider the following type of bivariate survival times: (*time to disease relapse*, *time to death*) and (*time to first infection*, *time to second infection*). In these cases, the non-occurrence of one or both of the component events would be termed as cure.

First, we will consider the Maller-Zhou estimator and re-derive its asymptotics, but without using martingale theory as in Maller and Zhou [11]. The estimator will then be extended to the multivariate case using a recently proposed multivariate Kaplan-Meier estimator [18]. The asymptotic normality of the estimator will be established, and its asymptotic variance will be estimated.

In the second step, our aim is to include, via kernel smoothing, the effect of covariates in the multivariate cure rate model under random censoring. For this new model with covariates, asymptotic normality of the proposed estimator and estimator of its asymptotic variance will be derived. The optimal choice of the bandwidth will be found through cross-validation.

In the third step, the presence of immunes will be tested under univariate random censoring. The limiting distribution of the test-statistic will be obtained under the null hypothesis of zero cure rate using extreme value theory.

In all three steps, a simulation study is conducted to support the mathematical theory.

1.3 Organization of the Thesis

This section outlines the layout of this thesis. The thesis consists of five chapters. In Chapter 1 (present chapter), a brief description of the problem is given. Chapter 2 is dedicated to the new nonparametric multivariate cure rate estimator with random censoring. The asymptotic distribution of the proposed estimator and its variance are also identified in Chapter 2. In Chapter 3, the covariates are included in the model and a new nonparametric multivariate cure rate estimation with the effect of the covariates is proposed and studied. In Chapter 4, the presence of immunes in the univariate model is tested using the extreme value theory. In Chapter 5, a short summary of this thesis is provided.

Chapter 2

Multivariate Cure Rate Estimation under Random Censoring

2.1 Introduction

One of the popular and common studies in the survival analysis is analyzing survival data with cure rate. Survival analysis is used for modeling and analyzing time to event data and typically in survival data, an event is defined as death or failure. During a study, individuals can be cured (cured of a disease or no warranty claim for a car) or fail (death from a disease or claim for a warranty). Cured or immunes are the individuals who are not subject to the event of interest. The cure rate is the probability of being cured for an infected individual so in other words, $p = P(X = \infty)$ where X is time to the event of interest.

There are two main univariate cure models in the literature. The first one is a mixture model, $\bar{F}(t) = P(X \geq t) = p + (1-p)\bar{F}_o(t)$, where $\bar{F}_o(t)$ is the baseline survival function. The second one is bounded cumulative hazard (BCH) model, $P(X \geq t) = exp(-\int_0^t h(s)ds)$ such that $0 < p = exp(-\int_0^\infty h(s)ds) < 1$, where $h(.)$ is the hazard rate. There are not many studies in the literature focusing on the nonparametric cure rate estimation. One of the first studies in this field is Maller and Zhou [11] who estimate the cure rate nonparametrically using the tail of Kaplan-Meier estimator [10].

Chen et al. [5] proposed a Bayesian approach for a multivariate cure rates. They used data from a melanoma clinical trial done by ECOG (Eastern Cooperative Oncology Group) where $X = (X_1, X_2)$, X_1 is time to relapse after treatment and X_2 is time from relapse to death. A frailty term with a positive stable distribution is introduced in their model to characterize the association between the failure times. They emphasized that sufficient follow-up of patients is crucial to their approach. They also considered the effect of covariates in the model. However, their model is based on a parametric approach and their model cannot be used for nonparametric approach. Xu and Peng [24] obtained a nonparametric cure rate estimator incorporating the impact of covariates. Their proposed estimator is consistent and asymptotically normal, but it only handles the univariate cure rate.

Sen and Stute [18] proposed a multivariate Kaplan-Meier estimator, which was the unique solution of an eigenfunction equation via a mass-shifting method. In this paper, we use the tail (excess mass) of the Sen-Stute estimator as an estimator of cure rates under multivariate random censoring, taking a cue from Maller and Zhou [13]. The computation of our estimator is straightforward, unlike that of Chen et al. [5]. Further, even though we use the sufficient follow-up condition to establish asymptotic normality of our estimator, our simulation model violates this condition. Yet the performance of our estimator is not affected. This shows that our estimator is quite robust.

This manuscript is organized as follows. In Section 2.2, the asymptotic normality of the Maller-Zhou estimator is established and its variance is estimated without using martingale theory, unlike in Maller and Zhou [13]. In Section 2.3, the asymptotic normality of the multivariate cure rate estimator is established and its variance estimator is obtained. In Section 2.4, the results from Sections 2.2 and 2.3 are illustrated using simulations. In Section 2.5, the model is applied on real data. In Section 2.6, a short summary of the research work is given.

2.2 Univariate Cure Rate Estimator

In this section, the univariate cure rate estimator is considered. Let X_i, $1 \leq i \leq n$, be independent and identically distributed (iid), non-negative random variables, each having a distribution function F(.) such that $F(x) = P\{X \leq x\} = (1-p)F_o(x)$. The survival function gives the probability of surviving beyond a specific time. It is defined as $\bar{F}(x) = P\{X \geq x\} = p + (1-p)\bar{F}_o(x)$. Due to the limited study time, the X_i's may not be observed. So in this case, there will be censoring variables $Y_i, 1 \leq i \leq n$, an independent set of iid random variables with distribution function $G(.)$, so that only $(\delta_i, Z_i), 1 \leq i \leq n$, can be observed where $\delta_i = \mathbf{1}\{X_i \leq Y_i\}$ and $Z_i = \min(X_i, Y_i)$ ($\mathbf{1}(A)$ indicates the indicator function of event A).

Define,

$$H_1\left(t\right) = E\left(\delta \; \mathbf{1}\left(Z \leq t\right)\right) = \left(1-p\right) \int_0^t \bar{G}\left(s\right) F_o\left(ds\right), \tag{1}$$

$$\bar{H}\left(t\right) = P(Z \geq t) = \big(\left(1-p\right) \bar{F}_o\left(t\right) + p\big) \; \bar{G}\left(t\right). \tag{2}$$

$H_1(.)$ and $\bar{H}(.)$ are estimated by the empirical versions of Eqs. (1) and (2) as

$$H_{n1}(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \; \mathbf{1}(Z_i \leq t), \qquad \bar{H}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i \geq t). \tag{3}$$

The cure rate p can be estimated by [6]

$$\hat{p} = \prod_{i=1}^{n} (1 - \frac{\delta_i}{n\bar{H}_n\left(Z_i\right)}), \tag{4}$$

the Kaplan-Meier estimator at the largest observation. Below we establish consistency and asymptotic normality of $\hat{p}$, but avoiding martingale arguments of Maller and Zhou [13]. Although for the sake of convenience, the absolute continuity of F_o and G are assumed.

Theorem 1 Suppose F_o and G are absolutely continuous and satisfy the sufficient follow-up assumption ($\tau_{F_o} \leq \tau_G$ where τ_{F_o} and τ_G are the right extremes of F_o and G, respectively, i.e, $\tau_{F_o} = sup(t \geq 0 : F_0(t) < 1)$), then $\sqrt{n}(\hat{p} - p)$ is asymptotically Normal with mean zero and limit variance

$$\sigma^2 = p^2(1-p)\int_0^\infty \frac{F_o(dx)}{(p + (1-p)\bar{F}_o(x))^2\bar{G}(x)},$$

A consistent estimator of σ^2 is

$$\hat{\sigma}^2 = n\hat{p}^2 \sum_{i=1}^{n} \frac{\delta_{[i:n]}}{\left(n-i+1\right)^2},$$

where $\delta_{[i:n]}$ is (the concomitant) attached to the i-th order statistics $Z_{i:n}$, $1 \leq i \leq n$.

Proof: First we take logarithm of both sides of Eq.(4) and then use a Taylor expansion,

$$\ln \hat{p} = \sum_{i=1}^{n} \ln \Big(1 - \frac{\delta_i}{n\bar{H}_n(Z_i)}\Big) = \frac{-1}{n} \sum_{i=1}^{n} \Big(\frac{\delta_i}{\bar{H}_n(Z_i)} + \frac{\delta_i^2}{2n\bar{H}_n^2(Z_i)} + \dots\Big).$$

Under sufficient follow-up assumption [11], $\frac{-1}{n}\sum\limits_{i=1}^{n}(\frac{{\delta_i}^2}{2n\bar{H}_n^2(Z_i)} + \dots) = O(\frac{1}{n})$ is negligible. Therefore,

$$\ln \hat{p} \approx \frac{-1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\bar{H}_n(Z_i)} = -\int_0^\infty \frac{H_{n1}(dt)}{\bar{H}_n(t)} = -\int_0^\infty \frac{H_{n1}(dt)}{\bar{H}(t)} + \int_0^\infty \Big(\frac{1}{\bar{H}(t)} - \frac{1}{\bar{H}_n(t)}\Big) H_{n1}(dt)$$

$$= -\int_0^\infty \frac{H_1(dt)}{\bar{H}(t)} - \int_0^\infty \Big(\frac{H_{n1}(dt)}{\bar{H}(t)} - \frac{H_1(dt)}{\bar{H}(t)}\Big) + \int_0^\infty \frac{\bar{H}_n(t) - \bar{H}(t)}{\bar{H}(t)\bar{H}_n(t)} H_{n1}(dt).$$

By dividing $H_1(t)$ and $\bar{H}(t)$ in Eq.(1) and Eq.(2),

$$\int_0^\infty \frac{H_1(dt)}{\bar{H}(t)} = \int_0^\infty \frac{(1-p)\bar{G}(t)F_0(dt)}{((1-p)\bar{F}_0(t) + p)\bar{G}(t)} = \int_0^\infty \frac{(1-p)F_0(dt)}{((1-p)\bar{F}_0(t) + p)} = -ln(p).$$

Since $\int_0^\infty \frac{H_1(dt)}{\bar{H}(t)} = -\ln p$ under absolute continuity of F_o and G, we have

$$\ln \hat{p} - \ln p = -\int_0^\infty \Big(\frac{H_{n1}(dt)}{\bar{H}(t)} - \frac{H_1(dt)}{\bar{H}(t)}\Big) + \int_0^\infty \frac{\bar{H}_n(t) - \bar{H}(t)}{\bar{H}^2(t)} H_1(dt)$$

$$+ \int_0^\infty \Big(\frac{H_{n1}(dt)}{\bar{H}(t)} - \frac{H_{n1}(dt)}{\bar{H}_n(t)} - \frac{\bar{H}_n(t)H_1(dt)}{\bar{H}^2(t)} + \frac{H_1(dt)}{\bar{H}(t)}\Big).$$

When n goes to infinity, the last term is negligible, since

$$\int_0^\infty \Big(\frac{H_{n1}(dt)}{\bar{H}(t)} - \frac{H_{n1}(dt)}{\bar{H}_n(t)} - \frac{\bar{H}_n(t)H_1(dt)}{\bar{H}^2(t)} + \frac{H_1(dt)}{\bar{H}(t)}\Big)$$

$$= \int_0^\infty \Big(\frac{H_{n1}(dt)(\bar{H}_n(t) - \bar{H}(t))}{\bar{H}^2(t)} - \frac{H_1(dt)(\bar{H}_n(t) - \bar{H}(t))}{\bar{H}^2(t)}\Big)$$

$$= \int_0^\infty \Big(\frac{(H_{n1}(dt) - H_1(dt))(\bar{H}_n(t) - \bar{H}(t))}{\bar{H}^2(t)}\Big)$$

approaches to zero. Consequently,

$$\ln \hat{p} - \ln p \approx -\int_0^\infty \frac{H_{n1}(dt)}{\bar{H}(t)} + \int_0^\infty \frac{\bar{H}_n(t)}{\bar{H}^2(t)} H_1(dt) \approx \frac{-1}{n}\sum_{i=1}^n \frac{\delta_i}{\bar{H}(Z_i)} + \frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{I(Z_i \geq t)H_1(dt)}{\bar{H}^2(t)}.$$

Note that the right side of above equation can be written as

$$\frac{1}{n}\sum_{i=1}^n \psi(Z_i) = \frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{I(Z_i \geq t)H_1(dt)}{\bar{H}^2(t)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\bar{H}(Z_i)},$$

and $E(\psi(Z_i)) = 0$. Hence using the Central Limit Theorem

$$\sqrt{n}(\ln \hat{p} - \ln p) = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{I(Z_i \geq t)H_1(dt)}{\bar{H}^2(t)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\bar{H}(Z_i)}\Big), \tag{5}$$

is asymptotically normal. Taking variance of both sides of Eq.(5) gives us,

$$var(\sqrt{n}(\ln \hat{p} - \ln p)) = \frac{1}{n}\ var\Big(\frac{\delta_i}{\bar{H}(Z_i)} - \int \frac{I(Z_i \geq t)H_1(dt)}{\bar{H}^2(t)}\Big) = \frac{1}{n}E\Big((\frac{\delta_i}{\bar{H}(Z_i)} - \int \frac{I(Z_i \geq t)H_1(dt)}{\bar{H}^2(t)})^2\Big)$$

$$= \frac{1}{n}E\Big(\frac{\delta_i}{\bar{H}(Z_i)}\Big)^2 + \frac{1}{n}\int\int \frac{P(Z_i \geq t \vee t')H_1(dt)H_1(dt')}{\bar{H}^2(t)\bar{H}^2(t')} - \frac{2}{n}\int E\Big(\frac{\delta_i I(Z_i \geq t)}{\bar{H}(Z_i)}\frac{H_1(dt)}{\bar{H}^2(t)}\Big) = \frac{1}{n}E\Big(\frac{\delta_i}{\bar{H}^2(Z_i)}\Big).$$

Therefore $var\,(\ln \hat{p} - \ln p)$ can be estimated by

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\delta_i}{\bar{H}_n^2(Z_i)} = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\delta_{[i:n]}}{\frac{(n-i+1)^2}{n^2}} = \sum_{i=1}^{n}\frac{\delta_{[i:n]}}{(n-i+1)^2}. \tag{6}$$

From the Δ-method, we know that if $\bar{X}_n$ has an asymptotic normal distribution such that

$$\sqrt{n}(\bar{X}_n - \mu) \to N(0, \sigma^2),$$

in distribution as $n \to \infty$, and assuming g has a derivative g' at μ, and $g'(\mu) \neq 0$, then

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \to N(0, (g'(\mu))^2\sigma^2).$$

Hence using the Δ-method and since here $g(x) = e^x$, the estimator of $var\,(\hat{p} - p)$ is

$$\hat{p}^2\sum_{i=1}^{n}\frac{\delta_{[i:n]}}{(n-i+1)^2}. \tag{7}$$

□

2.3 Multivariate Cure Rate Estimator

Let X and Y be two independent m-dimensional random vectors with distribution functions F and G respectively, where G is the censoring distribution. Let $Z_{ji} = min(X_{ji}, Y_{ji})$ and $\delta_{ji} = \mathbf{1}(X_{ji} \leq Y_{ji})$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. The survival function of X is defined as $\bar{F}(x) = \bar{F}(x_1, \ldots, x_m) = P\{X_1 \geq x_1, \ldots, X_m \geq x_m\} = P(X \geq x)$, where the inequality is coordinatewise, here as well as in the rest of the paper. To estimate $\bar{F}$, Sen and Stute [18] employed a mass-shifting technique. In this paper, we use the mass-shifting method to find the estimator for $\bar{F}$. This method gives a positive mass to x_∞ which is defined below, to help us to estimate the cure rate.

Define

$$X_\varepsilon = \begin{cases} X & \text{with probability } 1-\varepsilon \\ x_\infty & \text{with probability } \varepsilon \end{cases}$$

where X is a multivariate vector. x_∞ is a vector, possibly $(\infty, ..., \infty)$ which exceeds all $(x_1, ..., x_m)$ componentwise, $x \leq x_\infty \leq \infty$ and $0 < \varepsilon < 1$. The survival function of X_ε can be written as

$$\bar{F}_\varepsilon(x_1, ...x_m) = \begin{cases} (1-\varepsilon)\bar{F}(x_1, ...x_m) + \varepsilon & \text{if } (x_1, ...x_m) \in \mathrm{R}^m \\ \varepsilon & \text{if } x_\infty = (x_1, ...x_m) \end{cases}$$

and $\bar{F}_\varepsilon(\cdot)$ satisfies the integral equation below, such that $\bar{F}_\varepsilon(0) = 1$,

$$\bar{F}_\varepsilon(x) = \int_0^\infty I(t \geq x, t \neq x_\infty)\bar{F}_\varepsilon(t)\frac{(1-\varepsilon)F(dt)}{(1-\varepsilon)\bar{F}(t) + \varepsilon} + \varepsilon.$$

The survival function of the defective model is $\bar{F}_p(t) = (1-p)P(X \geq t) + p = (1-p)\bar{F}(t) + p$ where $\bar{F}(t)$ is the baseline survival function and $\bar{F}_p(\infty) = p$.

Let $\bar{F}_{\varepsilon,p}(t)$ be a defective survival function with cure rate p which takes the value of x_∞ with a probability of ε,

$$\bar{F}_{\varepsilon,p}(t) = (1-\varepsilon)\bar{F}_p(t) + \varepsilon.$$

Then $(\bar{F}_{\varepsilon,p}(t), \bar{F}_{\varepsilon,p}(\infty))$ for $t \geq 0$, is the unique solution [18] to

$$\bar{F}_{\varepsilon,p}(x) = \int I(t \geq x, t \neq \infty)\bar{F}_{\varepsilon,p}(t)\frac{\bar{F}_{\varepsilon,p}(dt)}{\bar{F}_{\varepsilon,p}(t)} + \bar{F}_{\varepsilon,p}(\infty), \qquad \bar{F}_{\varepsilon,p}(0) = 1. \tag{8}$$

By iteration of Eq.(8), we obtain

$$\begin{aligned}
&\bar{F}_{\varepsilon,p}(x) = \bar{F}_{\varepsilon,p}(\infty) + \bar{F}_{\varepsilon,p}(\infty)\int I(t_1 \geq x, t_1 \neq \infty)\frac{F_{\varepsilon,p}(dt_1)}{\bar{F}_{\varepsilon,p}(t_1)} \\
&+ \int\int I(t_2 \geq t_1 \geq x, t_2 \neq \infty)\bar{F}_{\varepsilon,p}(t_2)\frac{F_{\varepsilon,p}(dt_1)F_{\varepsilon,p}(dt_2)}{\bar{F}_{\varepsilon,p}(t_1)\bar{F}_{\varepsilon,p}(t_2)} \\
&= \bar{F}_{\varepsilon,p}(\infty)\Big(1 + \sum_{k=1}^{\infty}\int \mathbf{1}(t_k \geq ... \geq t_1 \geq x, t_k \neq \infty)\frac{(1-\varepsilon)(1-p)F(dt_1)}{(1-\varepsilon)\bar{F}_p(t_1)+\varepsilon}...\frac{(1-\varepsilon)(1-p)F(dt_k)}{(1-\varepsilon)\bar{F}_p(t_k)+\varepsilon}\Big),
\end{aligned}$$

since $\bar{F}_{\varepsilon,p}(0) = 1$, then

$$\begin{aligned}
&\bar{F}_{\varepsilon,p}(\infty) = p(1-\varepsilon) + \varepsilon \\
&= \frac{1}{1 + \sum_{k=1}^{\infty}\int I(t_k \geq \dots t_1 \geq 0, t_k \neq \infty)\frac{(1-\varepsilon)(1-p)F(dt_k)}{(1-\varepsilon)\bar{F}_p(t_k)+\varepsilon}\dots\frac{(1-\varepsilon)(1-p)F(dt_1)}{(1-\varepsilon)\bar{F}_p(t_1)+\varepsilon}},
\end{aligned} \tag{9}$$

and for $x \neq \infty$,

$$\bar{F}_{\varepsilon,p}(x) = \frac{1 + \sum\limits_{k=1}^{\infty}\int \mathbf{1}(t_k \geq ... \geq t_1 \geq x, t_k \neq \infty)\frac{(1-\varepsilon)(1-p)F(dt_1)}{(1-\varepsilon)\bar{F}_p(t_1)+\varepsilon}...\frac{(1-\varepsilon)(1-p)F(dt_k)}{(1-\varepsilon)\bar{F}_p(t_k)+\varepsilon})}{1 + \sum\limits_{k=1}^{\infty}\int \mathbf{1}(t_k \geq ... \geq t_1 \geq 0, t_k \neq \infty)\frac{(1-\varepsilon)(1-p)F(dt_k)}{(1-\varepsilon)\bar{F}_p(t_k)+\varepsilon}...\frac{(1-\varepsilon)(1-p)F(dt_1)}{(1-\varepsilon)\bar{F}_p(t_1)+\varepsilon})}. \tag{10}$$

Note that the mass-shifting technique ensures a unique solution [18]. The empirical version of Eq.(8) is

$$\bar{F}_{\varepsilon,n}(x) = \int \mathbf{1}(t \geq x, t \neq \infty)\bar{F}_{\varepsilon,n}(t)\frac{(1-\varepsilon)H_n^1(dt)}{(1-\varepsilon)\bar{H}_n(t)+\varepsilon} + \bar{F}_{\varepsilon,n}(\infty), \qquad \bar{F}_{\varepsilon,n}(0) = 1 \tag{11}$$

where

$$H_n^1(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}...\delta_{mi}\mathbf{1}(Z_{1i} \leq t_1, ..., Z_{mi} \leq t_m), \qquad \bar{H}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_{1i} \geq t_1, ..., Z_{mi} \geq t_m).$$

Following Sen and Stute's work [18], Eq.(11) can be written in the following eigenvector form and has a unique solution,

$$\mathbf{A}\mathbf{B}\bar{\mathbf{F}} = \bar{\mathbf{F}}, \qquad \mathbf{b}^T\bar{\mathbf{F}} = 1,$$

where $\bar{\mathbf{F}} = (\bar{F}_1, \bar{F}_2, ..., \bar{F}_n, \bar{F}_{n+1})$, $\mathbf{A} = ((a_{ij}))_{1\leq i,j\leq n+1}$, $\mathbf{B} = diag(b_1, ..., b_n, b_{n+1})$, $\mathbf{b} = (b_1, ..., b_n, b_{n+1})$, and for $1 \leq i, j \leq n$,

$$a_{ij} = I(Z_j \geq Z_i), a_{i,n+1} = a_{n+1,n+1} = 1, a_{n+1,j} = 0,$$

$$b_i = \frac{(1-\varepsilon)\delta_i}{(1-\varepsilon)\sum\limits_{k=1}^{n} a_{ik} + n\varepsilon}, b_{n+1} = 1,$$

Sen and Stute [18] proposed $\bar{F}_{\varepsilon,n}(.)$ as the multivariate Kaplan-Meier estimator. Note that in 1 dimension (i.e., $m = 1$) with $\varepsilon = 0$, this estimator reduces to the univariate Kaplan-Meier estimator. Here we propose $\bar{F}_{\varepsilon,n}(\infty) = \bar{F}_{n+1}$ as the cure estimator $\hat{p}_n$, extending the Maller-Zhou estimator to the multivariate case. The choice of $\varepsilon = \frac{1}{n+1}$ appears to be a natural choice.

Under random censoring, define

$$\bar{F}_{\varepsilon,p}^o(x) = \int \mathbf{1}(t \geq x, t \neq \infty)\bar{F}_{\varepsilon,p}^o(t)\frac{(1-\varepsilon)H_p^1(dt)}{(1-\varepsilon)\bar{H}_p(t) + \varepsilon} + \bar{F}_{\varepsilon,p}^o(\infty), \tag{12}$$

where $\bar{F}_{\varepsilon,p}^o(0) = 1$, $H_p^1(dt) = \bar{G}(t)F_p(dt)$ and $\bar{H}_p(t) = \bar{G}(t)\bar{F}_p(t)$.

Theorem 2 Assuming Supp F $\subseteq$ Supp G, $\bar{F}_{\varepsilon,n}(x) - \bar{F}^o_{\varepsilon,p}(x)$ has the asymptotic representation

$$\sqrt{n}(\bar{F}_{\varepsilon,n}(x) - \bar{F}^o_{\varepsilon,p}(x)) = L_n(x) + o_p(1), \qquad 0 \leq x \leq \infty$$

where $L_n(x)$ is given by

$$L_n(x) - \int \mathbf{1}(t \geq x, t \neq \infty) L_n(t) \frac{H^1_p(dt)}{\bar{H}_p(t)} - L_n(\infty) = \int \mathbf{1}(t \geq x, t \neq \infty)\alpha_n(dt), \qquad x \neq \infty$$

$$- \int \mathbf{1}(t \geq 0, t \neq \infty) L_n(t) \frac{H^1_p(dt)}{\bar{H}_p(t)} - L_n(\infty) = \int \mathbf{1}(t \geq 0, t \neq \infty)\alpha_n(dt), \qquad x = 0$$

where

$$\alpha_n(dt) := \bar{F}(t)\Big(\frac{H^1_n(dt)}{\bar{H}(t)} - \bar{H}_n(t)\frac{H^1(dt)}{\bar{H}^2(t)}\Big), \ t \geq 0.$$

Proof: By subtracting Eqs. (12) from (11) and multiplying both sides by $\sqrt{n}$, we get

$$\sqrt{n}(\bar{F}_{\varepsilon,n}(x) - \bar{F}^o_{\varepsilon,p}(x)) = \int \mathbf{1}(t \geq x, t \neq \infty)\bar{F}_{\varepsilon,n}(t)\sqrt{n}\Big(\frac{(1-\varepsilon)H^1_n(dt)}{(1-\varepsilon)\bar{H}_n(t)+\varepsilon} - \frac{(1-\varepsilon)H^1_p(dt)}{(1-\varepsilon)\bar{H}_p(t)+\varepsilon}\Big)$$

$$+ \int \mathbf{1}(t \geq x, t \neq \infty)\sqrt{n}(\bar{F}_{\varepsilon,n}(t) - \bar{F}^o_{\varepsilon,p}(t))\frac{(1-\varepsilon)H^1_p(dt)}{(1-\varepsilon)\bar{H}_p(t)+\varepsilon} + \sqrt{n}(\bar{F}_{\varepsilon,n}(\infty) - \bar{F}^o_{\varepsilon,p}(\infty)), \quad (13)$$

since we know $\bar{F}_{\varepsilon,n}(0) = \bar{F}^0_{\varepsilon,p}(0) = 1$ by iterating Eq.(13), we obtain

$$\sqrt{n}(\bar{F}_{\varepsilon,n}(\infty) - \bar{F}^o_{\varepsilon,p}(\infty)) =$$

$$- \frac{b_{\varepsilon,n}(0) + \sum_{k=1}^{\infty} \int \mathbf{1}(0 \le t_1 \le ... \le t_k \ne \infty) b_{\varepsilon,n}(t_k) \prod_{j=1}^{k} \frac{(1-\varepsilon)H^1_p(dt_j)}{(1-\varepsilon)\bar{H}_p(t_j)+\varepsilon}}{1 + \sum_{k=1}^{\infty} \int \mathbf{1}(0 \le t_1 \le ... \le t_k \ne \infty) \prod_{j=1}^{k} \frac{(1-\varepsilon)H^1_p(dt_j)}{(1-\varepsilon)\bar{H}_p(t_j)+\varepsilon}}, \tag{14}$$

where

$$b_{\varepsilon,n}(x) = \int \mathbf{1}(t \ge x, t \ne \infty) \bar{F}_{\varepsilon,n}(t) \sqrt{n} \Big(\frac{(1-\varepsilon)H^1_n(dt)}{(1-\varepsilon)\bar{H}_n(t)+\varepsilon} - \frac{(1-\varepsilon)H^1_p(dt)}{(1-\varepsilon)\bar{H}_p(t)+\varepsilon} \Big),$$

and consequently, by replacing Eq.(14) in Eq.(13) and iterating it, we get

$$\sqrt{n}(\bar{F}_{\varepsilon,n}(x) - \bar{F}^o_{\varepsilon,p}(x)) = b_{\varepsilon,n}(x) - b_{\varepsilon,n}(0) + \sum_{k=1}^{\infty} \int (\mathbf{1}_x - \mathbf{1}_0) b_{\varepsilon,n}(t_k) \prod_{j=1}^{k} \frac{(1-\varepsilon)H^1_p(dt_j)}{(1-\varepsilon)\bar{H}_p(t_j)+\varepsilon},$$

where $\mathbf{1}_u = \mathbf{1}(t_k \ge ... \ge t_1 \ge u, t_k \ne \infty)$.

Now it may be verified that if we choose $\varepsilon = o(\frac{1}{\sqrt{n}})$, (e.g., $\varepsilon = \frac{1}{n+1}$), then as $n \to \infty$,

$$\sqrt{n}(\bar{F}_{\varepsilon,n}(x) - \bar{F}^o_p(x)) = O_p(1) \qquad 0 \le x \le \infty,$$

$$b_{\varepsilon,n}(x) = \sqrt{n} \int \mathbf{1}(t \ge x, t \ne \infty) \alpha_n(dt) + o_p(1),$$

so using Eq.(13), $L_n(x)$ satisfies

$$L_n(x) - \int \mathbf{1}(t \geq x, t \neq \infty) L_n(t)(\bar{F}) \frac{H_p^1(dt)}{\bar{H}_p(t)} - L_n(\infty) = \int \mathbf{1}(t \geq x, t \neq \infty)\alpha_n(dt), \qquad x \neq \infty, \tag{15a}$$

$$- \int \mathbf{1}(t \geq 0, t \neq \infty) L_n(t) \frac{H_p^1(dt)}{\bar{H}_p(t)} - L_n(\infty) = \int \mathbf{1}(t \geq 0, t \neq \infty)\alpha_n(dt). \tag{15b}$$

□

The asymptotic covariance function of L_x is denoted by $V_{xy} = E(L_n(x)L_n(y))$. To obtain the covariance function, first we rewrite Eq.(15a-b) for y,

$$L_n(y) - \int \mathbf{1}(s \geq y, s \neq \infty) L_n(s)(\bar{F}) \frac{H_p^1(ds)}{\bar{H}_p(s)} - L_n(\infty) = \int \mathbf{1}(s \geq y, s \neq \infty)\alpha_n(ds), \tag{16a}$$

$$- \int \mathbf{1}(s \geq 0, s \neq \infty) L_n(s) \frac{H_p^1(ds)}{\bar{H}_p(s)} - L_n(\infty) = \int \mathbf{1}(s \geq 0, s \neq \infty)\alpha_n(ds). \tag{16b}$$

We know

$$\int \mathbf{1}(t \geq x)\alpha_n(dt) = n^{-1} \sum_{i=1}^{n} \Big(\frac{\delta_i \mathbf{1}\{Z_i \geq x\}}{\bar{G}(Z_i)} - \int \mathbf{1}\{Z_i \geq t \geq x\} \frac{F(dt)}{\bar{H}(t)} \Big) \text{ for all } x \geq 0,$$

so multiplying for $x, y \in I\!R_+^2 := [0,\infty) \times [0,\infty)$ and taking expectation give us the following function, call it A_{xy},

$$A_{xy} = n^{-1}\Big(\int \mathbf{1}\{t \ge x \vee y\}\frac{F(dt)}{\bar{G}(t)} - \int \mathbf{1}\{t \ge x\}\bar{F}(t \vee y)\frac{F(dt)}{\bar{H}(t)} - \int \mathbf{1}\{s \ge y\}\bar{F}(s \vee x)\frac{F(ds)}{\bar{H}(s)}$$

$$+ \int\int \mathbf{1}\{t \ge x, s \ge y\}\bar{H}(t \vee s)\frac{F(dt)}{\bar{H}(t)}\frac{F(ds)}{\bar{H}(s)}\Big) = n^{-1}\Big(\int \mathbf{1}\{t \ge x \vee y\}\frac{F(dt)}{\bar{G}(t)}$$

$$+ \int\int \mathbf{1}\{t \ge x, s \ge y\}(1 - \mathbf{1}\{t \ge s\} - \mathbf{1}\{t < s\})\bar{H}(t \vee s)\frac{F(dt)}{\bar{H}(t)}\frac{F(ds)}{\bar{H}(s)}\Big)$$

$$= n^{-1}\Big(\int \mathbf{1}\{t \ge x \vee y\}\bar{F}^2(t)\frac{H^1(dt)}{\bar{H}^2(t)}$$

$$+ \int\int \mathbf{1}\{t \ge x, s \ge y\}(1 - \mathbf{1}\{t \ge s\})(1 - \mathbf{1}\{t < s\})\bar{H}(t \vee s)\bar{F}(t)\bar{F}(s)\frac{H^1(dt)}{\bar{H}^2(t)}\frac{H^1(ds)}{\bar{H}^2(s)}\Big). \quad (17)$$

Multiplying Eq.(15a)× Eq.(16a), Eq.(15a) × Eq.(16b), Eq.(15b) × Eq.(16a), Eq.(16b) × Eq.(16b), then taking expectation give us Eqs.(18a-d) respectively

$$V_{xy} - \int \mathbf{1}(t \geq x, t \neq \infty) V_{ty} \frac{H_p^1(dt)}{\bar{H}_p(t)} - \int \mathbf{1}(s \geq y, s \neq \infty) V_{xs} \frac{H_p^1(ds)}{\bar{H}_p(s)} - V_{x\infty} - V_{y\infty}$$

$$+ \int \int \mathbf{1}(t \geq x, t \neq \infty, s \geq y, s \neq \infty) V_{ts} \frac{H_p^1(dt)}{\bar{H}_p(t)} \frac{H_p^1(ds)}{\bar{H}_p(s)} + \int \mathbf{1}(t \geq x, t \neq \infty) V_{t\infty} \frac{H_p^1(dt)}{\bar{H}_p(t)}$$

$$+ \int \mathbf{1}(s \geq y, s \neq \infty) V_{s\infty} \frac{H_p^1(ds)}{\bar{H}_p(s)} + V_{\infty\infty} = A_{xy}, \tag{18a}$$

$$- \int \mathbf{1}(s \geq 0, s \neq \infty) V_{sx} \frac{H_p^1(ds)}{\bar{H}_p(s)} - V_{x\infty} + \int \int \mathbf{1}(s \geq o, s \neq \infty, t \geq x, t \neq \infty) V_{ts} \frac{H_p^1(ds)}{\bar{H}_p(s)} \frac{H_p^1(dt)}{\bar{H}_p(t)}$$

$$+ \int \mathbf{1}(s \geq 0, s \neq \infty) V_{s\infty} \frac{H_p^1(ds)}{\bar{H}_p(s)} + \int \mathbf{1}(t \geq x, t \neq \infty) V_{t\infty} \frac{H_p^1(dt)}{\bar{H}_p(t)} + V_{\infty\infty} = A_{x0}, \tag{18b}$$

$$- \int \mathbf{1}(t \geq 0, t \neq \infty) V_{ty} \frac{H_p^1(dt)}{\bar{H}_p(t)} - V_{y\infty} + \int \int \mathbf{1}(t \geq o, t \neq \infty, s \geq y, s \neq \infty) V_{ts} \frac{H_p^1(dt)}{\bar{H}_p(t)} \frac{H_p^1(ds)}{\bar{H}_p(s)}$$

$$+ \int \mathbf{1}(t \geq 0, t \neq \infty) V_{t\infty} \frac{H_p^1(dt)}{\bar{H}_p(t)} + \int \mathbf{1}(s \geq y, s \neq \infty) V_{s\infty} \frac{H_p^1(ds)}{\bar{H}_p(s)} + V_{\infty\infty} = A_{0y}, \tag{18c}$$

$$\int \int \mathbf{1}(t \geq 0, t \neq \infty, s \geq 0, s \neq \infty) V_{ts} \frac{H_p^1(dt)}{\bar{H}_p(t)} \frac{H_p^1(ds)}{\bar{H}_p(s)} + \int \mathbf{1}(t \geq 0, t \neq \infty) V_{t\infty} \frac{H_p^1(dt)}{\bar{H}_p(t)}$$

$$+ \int \mathbf{1}(s \geq 0, s \neq \infty) V_{s\infty} \frac{H_p^1(ds)}{\bar{H}_p(s)} + V_{\infty\infty} = A_{00}. \tag{18d}$$

Matrix Form We estimate V_{xy} by $\hat{V}_{xy} = V_{ij}$ at $x = Z_i$ and $y = Z_j$, for $1 \leq i, j \leq n+1$ where $Z_{n+1} = x_\infty$ and $V_{n+1,n+1} = \hat{V}_{\infty,\infty} = \hat{V}(\hat{p}_n)$. Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{b}$ be as before, and let $\bar{I\!\!F} = \text{diag}(\bar{F}_1, \ldots, \bar{F}_n)$. Define the matrix $\mathbf{D} = ((d_{ij}))_{1 \leq i,j \leq n}$ as

$$d_{ij} = (1-a_{ij})(1-a_{ji})\sum_{r=1}^{n} a_{ir}a_{jr}, \qquad 1 \le i,j \le n.$$

Then the sample version of Eqs.(18a-d) can be written as

$$\begin{bmatrix} (\mathbf{I}-\mathbf{AB}) \\ -\mathbf{b}^T \end{bmatrix} \mathbf{V} \begin{bmatrix} (\mathbf{I}-\mathbf{AB})^T & -\mathbf{b} \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{AB})_{n\times n}\bar{\mathbb{F}} \\ \mathbf{b}_n^T\bar{\mathbb{F}} \end{bmatrix} (\mathbf{I}+\mathbf{B}_n\mathbf{D}\mathbf{B}_n) \begin{bmatrix} \bar{\mathbb{F}}(\mathbf{BA})_{n\times n}^T & \bar{\mathbb{F}}\mathbf{b}_n \end{bmatrix},$$

where $\mathbf{b}_n$ is a subvector of $\mathbf{b}$ with first n elements and matrices of $\mathbf{B}_n$ and $(\mathbf{AB})_{n\times n}$ are submatrices of matrix $\mathbf{B}$ and matrix $\mathbf{AB}$, respectively, with the first n rows and first n columns. So $\mathbf{V}$ can be found by

$$\mathbf{V} = \begin{bmatrix} \mathbf{I}-\mathbf{AB} \\ -\mathbf{b}^T \end{bmatrix}^{-} \begin{bmatrix} \mathbf{AB}\bar{\mathbb{F}} \\ \mathbf{b}^T\bar{\mathbb{F}} \end{bmatrix} (\mathbf{I}+\mathbf{BDB}) \begin{bmatrix} \bar{\mathbb{F}}\mathbf{BA}^T & \bar{\mathbb{F}}\mathbf{b} \end{bmatrix} \begin{bmatrix} (\mathbf{I}-\mathbf{AB})^T & -\mathbf{b} \end{bmatrix}^{-},$$

where for any matrix $\mathbf{w}$, $\mathbf{w}^{-}$ denotes the g-inverse of $\mathbf{w}$. Here we use,

$$\begin{bmatrix} \mathbf{I}-\mathbf{AB} \\ -\mathbf{b}^T \end{bmatrix}^{-} = \left(\begin{bmatrix} \mathbf{I}-\mathbf{AB} \\ -\mathbf{b}^T \end{bmatrix}^T \begin{bmatrix} \mathbf{I}-\mathbf{AB} \\ -\mathbf{b}^T \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{I}-\mathbf{AB} \\ -\mathbf{b}^T \end{bmatrix}^T.$$

2.4 Simulation

Univariate Case: In this section, a simulation study is conducted. We first generate n independent random variables, X_i, $1 \le i \le n$ from an exponential distribution with mean of m_1 and n independent random censored times, Y_i, $1 \le i \le n$ from an exponential distribution with mean of m_2. X_i and Y_i are chosen to be dependent of each other. The cure rate p is assumed to be known and u is generated from the uniform (0,1) distribution. δ_1 is generated as a binary variable (if $u > p$ and $X_1 \leq Y_1$ then $\delta_1 \equiv 1$, otherwise $\delta_1 \equiv 0$). Using Eq.(7) and setting $m_1 = 2$, $m_2 = 4$ and different values for n and p, we got Table 1.

	Sample size	n=100	n=200	n=500
p=0	$\hat{p}$	0.01553	0.01000	0.00655
	$v\hat{a}r$	0.00029	0.00001	0.00005
	MSE	0.00095	0.00035	0.00018
p=0.25	$\hat{p}$	0.25846	0.25436	0.25315
	$v\hat{a}r$	0.00422	0.00243	0.00115
	MSE	0.00719	0.00294	0.00152
p=0.5	$\hat{p}$	0.50324	0.50168	0.49890
	$v\hat{a}r$	0.00533	0.00289	0.00120
	MSE	0.00651	0.00337	0.00104

Table 1: Table of univariate cure rate estimators (for different n and p)

Bivariate Case: To run a bivariate simulation study, the independent random vectors of S_{ij}, T_{ij} and D_{ij} for $1 \leq i \leq n$ and $j = 1, 2$ with following densities are generated

$$S_{i1} \sim exp(\lambda_1 - c_1), S_{i2} \sim exp(\lambda_2 - c_1), D_{i1} \sim exp(c_1),$$

$$T_{i1} \sim exp(\lambda_3 - c_2), T_{i2} \sim exp(\lambda_4 - c_2), D_{i2} \sim exp(c_2).$$

Then X_{i1}, X_{i2}, Y_{i1} and Y_{i2} are generated as the following such that X_{i1} and X_{i2}, and Y_{i1} and Y_{i2} are dependent:

$$X_{i1} = min(S_{i1}, D_{i1}) \sim exp(\lambda_1), \qquad X_{i2} = min(S_{i2}, D_{i1}) \sim exp(\lambda_2),$$

$$Y_{i1} = min(T_{i1}, D_{i2}) \sim exp(\lambda_3), \qquad Y_{i2} = min(T_{i2}, D_{i2}) \sim exp(\lambda_4).$$

So, we can easily obtain that the survival function of a defective model is

$$\bar{F}_p(x_{i1}, x_{i2}) = (1-p)exp(-\lambda_1 x_{i1} - \lambda_2 x_{i2} - c_1 max(x_{i1}, x_{i2})) + p, \qquad for \;\; i = 1, ..., n.$$

A random sample with $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, $\lambda_3 = 0.15$, $\lambda_4 = 0.15$, $c_1 = 0.1$ and $c_2 = 0.1$ with different values of n are generated 100 times. After running the simulation, we got the estimated cure rate, its estimated variance and its MSE for different n and p which are shown in Table 2.

	Sample size	n=100	n=200	n=500
p=0	$\hat{p}$	0.05802	0.03523	0.01956
	$v\hat{a}r$	0.00021	0.00079	0.00023
	MSE	0.00397	0.00139	0.00045
p=0.25	$\hat{p}$	0.28098	0.26836	0.25746
	$v\hat{a}r$	0.00648	0.00360	0.00176
	MSE	0.00756	0.00340	0.00205
p=0.5	$\hat{p}$	0.53117	0.49564	0.50668
	$v\hat{a}r$	0.00794	0.00439	0.00173
	MSE	0.00672	0.00390	0.00211

Table 2: Table of bivariate cure rate estimators (for different n and p)

Both univariate and bivariate results show that our estimator's performance is great for moderate to large sample size. The $\hat{p}$ is very close to p and the variances and MSE's are very close to each other.

In Fig. 1, the plots show the bivariate surface of $\bar{F}$ (left) and its estimation $\bar{F}_n$ (right) for one simulated data set. p is assumed to be 0.25.

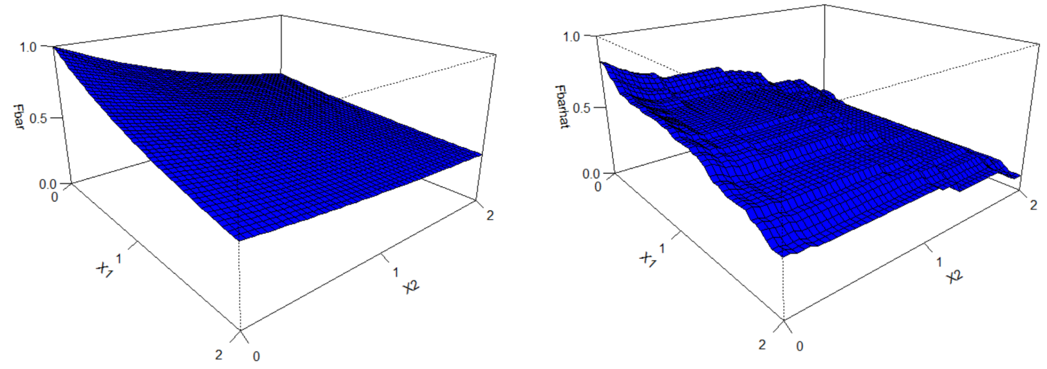

Figure 1: The bivariate surface of $\bar{F}$ (left) and its estimation $\bar{F_n}$ (right)

Fig. 1 shows that $\bar{F}$ and $\bar{F}_n$ are very similar and the values are very close to each other.

2.5 Illustration of the Model through Real Data

We applied our proposed model to the data from the Litter-matched tumorigenesis experiment [14]. In this study, 150 rats were divided into three different groups, control 1, control 2 and drug-treated. The event of interest is considered as tumor appearance. Any death due to other causes are considered as censoring. The length of the study is 104 weeks. For our model, we applied some modifications to their data to use it for a bivariate model. For our study, the drug-treated group is ignored, the control 1 is considered as X_1 and the control 2 is considered as X_2. Our goal is to estimate the probability of a cure among rats with a tumor. The Kaplan-Meier survival estimate is shown in Fig. 2 to show the possible presence of cured patients. The cure rate is estimated around 0.9024 and the estimated variance is equal to 0.003.

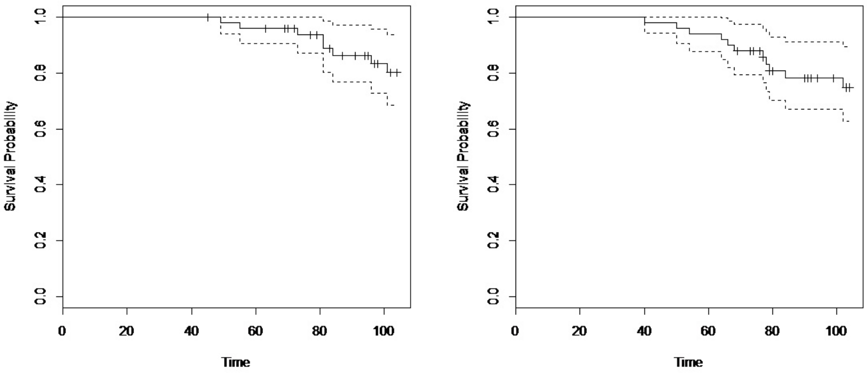

Figure 2: The Kaplan-Meier plot for control 1 (left) and control 2 (right).

2.6 Conclusion

In this research work, we have proposed a new multivariate cure rate model and have examined several of its properties. In this new approach, the tail of the Sen-Stute estimator has been used for the estimation of the cure rate. The asymptotic normality of this estimator has been obtained under the sufficient follow-up condition.

One of the advantages of this model is that the proposed estimator is quite robust, its calculation is not complicated and it can handle both univariate and multivariate cases. For ease and clarity of exposition, we conducted a simulation study for univariate and bivariate cure rate estimators.

Acknowledgments

This research was partially supported by NSERC.

Chapter 3

Nonparametric Multivariate Cure Rate Estimation with Covariates

3.1 Introduction

One of the popular and common studies in the survival analysis is survival data with cure rate. In survival analysis, a proportion of the population may be *cured*, i.e., will not experience the event of interest under study, such as death, disease onset, etc. On the other hand, survival data is often subject to *random censoring* which can happen due to reasons such as loss to follow-up, drop-out and limited study period. In this chapter, we consider the estimation of *the conditional cure-rate*, i.e., cure-probability as a function of covariates, under multivariate random censoring.

Let U_i, $1 \leq i \leq n$, be independent and identically distributed (iid), non-negative random vectors, each having a marginal distribution function $F(u) = P(U \leq u)$ and survival function $\bar{F}(u) = P(U \geq u)$. A proportion of individuals cured because of covariates, are considered in the model. Since a proportion of individuals are cured, the conditional survival function is defective and it has the following form

$$\bar{F}_0(u|c) = P(U \geq u | U < \infty, C = c),$$
$$\pi(c) = P(U = \infty | C = c),$$

where C is the vector of covariates and $\pi(c)$ is the conditional cure rate which depends on the covariates. The overall conditional survival function therefore has the following mixture model form:

$$\bar{F}(u|c) = P(U \geq u | C = c) = \pi(c) + (1 - \pi(c))\bar{F}_0(u|c).$$

Beran [1] considered univariate conditional Kaplan-Meier estimators in the presence of covariates and censoring. Beran [1] and Dabrowska [6] studied the asymptotic distribution and uniform consistency for the kernel and nearest neighbour estimates of the survival function.

Nieto-Barajas and Yin [15] studied a Bayesian semiparametric model with cure rates. They proposed a model in which each individual's cure time can be different depending on the parametrically modelled covariates.

Tsodikov [21] used the proportional hazard model to find the effect of covariates on the cure rate in a non-parametric framework. One of the disadvantages of his method is that it is only applicable on discrete covariates.

Xu and Peng [24] proposed a fully non-parametric estimator for the cure rate with covariates. Following Maller and Zhou [11], they used the conditional Kaplan-Meier estimator of Beran [1] at the largest uncensored failure time to find a consistent and asymptotically normal estima for conditional cure rate. They considered the effect of one and more than one covariates on the model, although the survival time is univariate.

Chen et al. [5] developed a Bayesian estimator for multivariate cure rates in the presence of right censoring, based on a parametric model. In their proposed model, the proportional hazard model is used to estimate the effect of covariates on the cure rate. However, no non-parametric cure-rate estimator seems to exist in the multivariate set-up.

In this chapter, we propose a new non-parametric multivariate cure rate estimator in the presence of multivariate censoring and covariates. Our estimator thus extends the estimator of Xu and Peng [24] to the case of multivariate survival time and reduces to the latter in the univariate case. The estimator is obtained by introducing kernel smoothing into the fundamental eigenvector equation of the previous chapter.

In Section 3.2, the smoothing equation is introduced and a nonparametric multivariate cure rate estimator in the presence of covariates is obtained. In Section 3.3, the asymptotic distribution of the proposed estimator is obtained. In Section 3.4, the estimated covariance function has been obtained along with optimal order and choice of bandwidth via cross-validation. A simulation study of the proposed model is conducted in Section 3.5. A brief conclusion and discussion are provided in Section 3.6.

3.2 Nonparametric Multivariate Cure Rate Estimator with Covariates

Let X and Y be independent m and r-dimensional random vectors with distribution functions F and G, respectively, where $X = (U, C) = (U_1, \ldots, U_r, C_1, \ldots, C_s)$, $r + s = m$, $r \geq 1$, $s \geq 1$, are random variables such that U is subject to censoring and C is the s dimensional vector of covariates which is not subject to censoring. Let $Y = (Y_1, \ldots, Y_r)$ be the censoring variable. We assume the conditional distribution of $U|C = c$, is given by

$$\begin{array}{lll} (1-\pi(c))f_{U|C}(u|c) & 0 \leq u < \infty & \text{[conditional density]}, \\ \pi(c) = P(u = \infty|c) & u = \infty & \text{[mass point]}. \end{array}$$

In other words, the finite (i.e., non-cured) part of the conditional distribution is given by a conditional density. Further, assuming C has a density, the joint distribution of U and C is given by

$$\begin{array}{ll} (1-\pi(c))f_{U,C}(u,c) & 0 \leq u < \infty, \quad 0 \leq c < \infty, \\ \pi(c)f_C(c) & u = \infty, \quad 0 \leq c < \infty, \end{array}$$

and the marginal distribution of U is given by

$$\begin{array}{ll} \int (1-\pi(c))f_{U|C}(u|c)f_C(c)dc & 0 \leq u < \infty, \\ \int \pi(c)f_C(c)dc = p & u = \infty. \end{array}$$

Our aim is to estimate the conditional cure-rate function $\pi(c)$.

Assume $W_j = min(U_j, Y_j)$ and $\eta_j = I(U_j \leq Y_j)$ for $j = 1, \ldots, r$ so (Z, δ) are observable under random censoring, where $\delta = (\eta_1, \ldots, \eta_r, 1, \ldots, 1)$ and $Z = (W, C) = (W_1, \ldots, W_r, C_1, \ldots, C_s)$ are m-dimensional vectors. Note again that $C = (C_1, ..., C_s)$ is uncensored. Our estimation is based on sample size of n.

Let $\bar{F}_\pi(u|c)$ be the defective conditional survival function with cure rate $\pi(c)$, i.e.,

$$\bar{F}_\pi(u|c) = (1-\pi(c))\bar{F}_{U|C}(u|c) + \pi(c),$$

then it is the unique solution to the following equation

$$\bar{F}_\pi(u|c) = \int I(u \leq t < \infty)\bar{F}_\pi(t|c)\frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} + \bar{F}_\pi(\infty|c) \qquad 0 \leq u < \infty,$$

where $\bar{F}_\pi(\infty|c) = \pi(c)$ and $\bar{F}_\pi(0|c) = 1$.

To find smooth structure in the data, kernel smoothing is commonly used in the literature. The general form of s-dimensional kernel density estimator is

$$f(x, h_n) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i),$$

where $K_n(x-X_i) = \frac{1}{h_n^s}K(\frac{x-X_i}{h_n})$ and $K(.)$ is a non-negative, s-variate kernel function such as Gaussian ($K(x) = (2\pi)^{-s/2}\exp(-\frac{x'x}{2})$), satisfying $\int K(x)dx = 1$ and h_n is the bandwidth satisfying, as $n \to \infty$, $h_n \to 0$ and $nh_n^s \to \infty$ [23].

If $\varphi(U)$ be an arbitrary function of U, we have

$$E(\varphi(U)K_n(c - C)) = \int \varphi(u)I(0 \leq u < \infty)K_n(c - c')(1 - \pi(c'))f_{U|C}(u|c')f_C(c')dudc'$$

$$+ \varphi(\infty)\int K_n(c - c')\pi(c')f_C(c')dc'. \tag{19}$$

If the bandwidth approaches zero, the equation above becomes (*all the limit statements below assume appropriate smoothness and integrability conditions, to be specified later*):

$$E(\varphi(U)K_n(c - C)) \xrightarrow{h_n \to 0} (1 - \pi(c))f_C(c)\int \varphi(u)I(0 \leq u < \infty)f_{U|C}(u|c)du + \varphi(\infty)\pi(c)f_C(c)$$

$$= f_C(c)E(\varphi(U)|C = c).$$

Further, by taking $\varphi(U) = I(U \geq u)$,

$$E(I(U \geq u)K_n(c - C)) \xrightarrow{h_n \to 0} (1 - \pi(c))f_C(c)\bar{F}_{U|C}(u|c) + \pi(c)f_C(c) = f_C(c)\bar{F}_\pi(u|c). \tag{20}$$

For the censored data, define the empirical processes as

$$\int \varphi(w)H_{n1}(dw|c, h_n) = \frac{1}{n}\sum_{i=1}^{n} \eta_i K_n(c - C_i)\varphi(W_i),$$

$$\bar{H}_n(w|c, h_n) = \frac{1}{n}\sum_{i=1}^{n} K_n(c - C_i)I(W_i \geq w),$$

which are kernel weighted versions of $H_{n1}(.)$ and $\bar{H}_n(.)$, respectively, of the previous chapter. Here, as before, $\eta = \eta_1\eta_2...\eta_r$. Note that

$$\int \varphi(w)H_1(dw|c, h_n) = E\Big(\int \varphi(w)H_{n1}(dw|c, h_n)\Big)$$

$$= \int \bar{G}(u)\varphi(u)I(0 \leq u < \infty)K_n(c - c')(1 - \pi(c'))f_{U|C}(u|c')f_C(c')dudc', \tag{21}$$

so that

$$\int \varphi(w)H_1(dw|c, h_n) \xrightarrow{h_n \to 0} f_C(c)E\Big(\varphi(U)\bar{G}(U)I(0 \leq U < \infty)|C = c\Big), \tag{22}$$

and

$$\bar{H}(w|c, h_n) := E(\bar{H}_n(w|c, h_n)) \xrightarrow{h_n \to 0} f_C(c)\bar{G}(w)\bar{F}_\pi(w|c). \tag{23}$$

Hence

$$\int \varphi(w)\frac{H_1(dw|c,h_n)}{\bar{H}(w|c,h_n)} \xrightarrow{h_n\to 0} \int \varphi(w)I(0\le w<\infty)\frac{F_\pi(dw|c)}{\bar{F}_\pi(w|c)}, \tag{24}$$

and it follows that we may use the unique solution $(\bar{F}_n(u|c,h_n), 0\le u<\infty, \pi_n(c))$ of the following equation as estimators of $(\bar{F}_\pi(u|c), 0\le u<\infty, \pi(c))$:

$$\bar{F}_n(u|c,h_n) = \int I(u\le t<\infty)\bar{F}_n(t|c,h_n)\frac{(1-\varepsilon)H_{n1}(dt|c,h_n)}{(1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon} + \pi_n(c) \qquad 0\le u<\infty, \tag{25}$$

where $\bar{F}_n(0|c,h_n) = 1$ and $\varepsilon > 0$ is a sufficiently small mass-shifting parameter (below we take $\varepsilon = \frac{1}{n+1}$, as before). For convenience of comparison, we recall the equation for $(\bar{F}_\pi(u|c), 0\le u<\infty, \pi(c))$:

$$\bar{F}_\pi(u|c) = \int I(u\le t<\infty)\bar{F}_\pi(t|c)\frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} + \pi(c) \qquad 0\le u<\infty, \qquad \bar{F}_\pi(0|c) = 1. \tag{26}$$

Eq.(25) can be re-written as the following eigenvector equation,

$$\mathbf{A}\mathbf{B}_C\bar{\mathbf{F}} = \bar{\mathbf{F}}, \qquad \mathbf{b}_C^T\bar{F} = 1, \tag{27}$$

where for $1\le i,j\le n, \bar{\mathbf{F}} = (\bar{\mathbf{F}}_1,...,\bar{\mathbf{F}}_n,\bar{\mathbf{F}}_{n+1}), \mathbf{A} = ((a_{ij})), \mathbf{B}_C = diag(b_1(c),...,b_n(c),1)$, $\mathbf{b}(c) = (b_1(c),...,b_n(c),1)$ and $a_{ij} = I(Z_j\ge Z_i) = I(W_j\ge W_i, C_j\ge C_i), a_{i,n+1} = a_{n+1,n+1} = 1, a_{n+1,j} = 0$,

$$b_i(c) = \frac{(1-\varepsilon)\frac{\eta_i}{nh^s}K(\frac{c-C_i}{h})}{(1-\varepsilon)\sum_{k=1}^n a_{ik}\frac{1}{nh^s}K(\frac{c-C_k}{h})+\varepsilon}. \tag{28}$$

Note that $\pi_n(c) = \bar{F}_{n+1}$ and the inequalities are coordinate-wise, i.e., $I(Z_j\ge Z_i) = I(Z_{jk}\ge Z_{ik}, 1\le k\le m)$.

3.3 Asymptotic Properties

In this section, we first aim to find the expressions for $\bar{F}_n(u|c,h_n) - \bar{F}_\pi(u|c)$ and $\pi_n(c) - \pi(c)$ as below. Note that

$$\bar{F}_n(u|c,h_n) - \bar{F}_\pi(u|c) =$$

$$\int I(u \leq t < \infty)\bar{F}_n(t|c,h_n)\Big(\frac{(1-\varepsilon)H_{n1}(dt|c,h_n)}{(1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon} - \frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon}\Big)+$$

$$\int I(u \leq t < \infty)[\bar{F}_n(t|c,h_n) - \bar{F}_\pi(t|c)]\frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon}+$$

$$\int I(u \leq t < \infty)\bar{F}_\pi(t|c)\Big(\frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|h_n)+\varepsilon} - \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)}\Big) + [\pi_n(c) - \pi(c)]. \tag{29}$$

Eq.(29) can be re-written as

$$[\bar{F}_n(u|c,h_n) - \bar{F}_\pi(u|c)] - \int I(u \leq t < \infty)[\bar{F}_n(t|c,h_n) - \bar{F}_\pi(t|c)]\Lambda(dt|c,h_n)(c) - [\pi_n(c) - \pi(c)]$$

$$= \int I(u \leq t < \infty)\bar{F}_n(t|c,h_n)\alpha_n(dt|c,h_n) + \int I(u \leq t < \infty)\bar{F}_\pi(t|c)\alpha(dt|c,h_n),$$

$$\bar{F}_n(0|c,h_n) - \bar{F}_\pi(0|c) = 0, \tag{30}$$

where

$$\alpha_n(dt|c,h_n) = \frac{(1-\varepsilon)H_{n1}(dt|c,h_n)}{(1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon} - \frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon},$$

$$\alpha(dt|c,h_n) = \frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon} - \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)},$$

$$\Lambda(dt|c,h_n) = \frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon}.$$

By iteration Eq.(30) becomes,

$$\begin{aligned}
&\bar{F}_n(u|c,h_n) - \bar{F}_\pi(u|c)\\
&= \sum_{k=1}^{\infty}\int ... \int I(u \le t_1 \le ... \le t_k < \infty)\bar{F}_n(t_k|c,h_n)\Lambda(dt_1|c,h_n)...\Lambda(dt_{k-1}|c,h_n)\alpha_n(dt_k|h_n)\\
&+ \sum_{k=1}^{\infty}\int ... \int I(u \le t_1 \le ... \le t_k < \infty)\bar{F}_\pi(t_k|c)\Lambda(dt_1|c,h_n)...\Lambda(dt_{k-1}|c,h_n)\alpha(dt_k|c,h_n)\\
&+ (\pi_n(c)-\pi(c))\Big(1+\sum_{k=1}^{\infty}\int ... \int I(u \le t_1 \le ... \le t_{k-1} < \infty)\Lambda(dt_1|c,h_n)...\Lambda(dt_{k-1}|c,h_n)\Big).
\end{aligned} \tag{31}$$

Now using $\bar{F}_n(0|c,h_n) - \bar{F}_\pi(0|c) = 0$, we get

$$\begin{aligned}
\pi_n(c)-\pi(c) = -\Big(&\sum_{k=1}^{\infty}\int ... \int I(0 \le t_1 \le ... \le t_k < \infty)[\bar{F}_n(t_k|c,h_n)\alpha_n(dt_k|c,h_n)\\
&+ \bar{F}_\pi(t_k|c)\alpha(dt_k|c,h_n)]\Lambda(dt_1|c,h_n)...\Lambda(dt_{k-1}|c,h_n)\Big).
\end{aligned}$$

$$\Big(1+\sum_{k=1}^{\infty}\int ... \int I(0 \le t_1 \le ... \le t_k < \infty)\Lambda(dt_1|c,h_n)...\Lambda(dt_k|c,h_n)\Big)^{-1}. \tag{32}$$

Eqs. (31) and (32) are the key to establishing asymptotic properties –consistency and asymptotic normality– of $(\bar{F}_n(.|c,h_n) - \bar{F}_\pi(.|c))$ and $(\pi_n(c) - \pi(c))$.

First we list some basic convergence results, involving $H_{n1}(.|c,h_n), \bar{H}_n(.|c,h_n)$, in the following lemma whose proof is easy:

Lemma 1 Assume that

(A1) $K(.)$ is a *product-kernel*, i.e., of the form $K(x_1,...,x_s) = K_0(x_1)...K_0(x_s)$, where $K_0(.)$ is a *symmetric*, univariate density function satisfying

$$\int x^2 K_0(x)dx =: \sigma^2(K_0) < \infty, \int K_0^2(x)dx =: R(K_0) < \infty.$$

(A2) The s-variate marginal and conditional densities $f_C(c)$ and $f_{U|C}(.|c)$, respectively, as well as the cure-rate function $\pi(c)$ are twice continuously differentiable at c.

Suppose $E(\varphi^2(W)) < \infty$ and $n \to \infty$,

(a) if $h_n \to 0$ and $\sum_{n=1}^{\infty} exp(-\rho n h_n^s) < \infty$ for all $\rho > 0$,

$$(i) \int \varphi(w)H_{n1}(dw|c,h_n) - f_C(c)E\Big(\varphi(U)\bar{G}(U)I(0 \le U < \infty)|C=c\Big) \to 0,$$

and

$$(ii) \max_{w \ge 0} |\bar{H}_n(w|c,h_n) - f_C(c)\bar{G}(w)\bar{F}_\pi(w|c)| \to 0,$$

for each $0 \leq c < \infty$, *with probability one.*

(b) if $h_n \to 0, nh_n^s \to \infty, n(h_n)^{s+4} \to 0$,

$$(nh_n^s)^{1/2}\Big(\int \varphi(w)H_{n1}(dw|c,h_n) - f_C(c)E\big(\varphi(U)\bar{G}(U)I(0 \leq U < \infty)|C=c\big)\Big) \to N(0,\sigma^2(\varphi,c))$$

in distribution, where

$$\sigma^2(\varphi,c) := (R(K_0))^s f_C(c)E\Big(\varphi^2(U)\bar{G}(U)I(0 \leq U < \infty)|C=c\Big),$$

and

$$(nh_n^s)^{1/2}(\bar{H}_n(w|c,h_n) - f_C(c)\bar{G}(w)\bar{F}_\pi(w|c)) \to N\big(0,(R(K_0))^s f_C(c)\bar{G}(w)\bar{F}_\pi(w|c)\big),$$

in distribution, for each $0 \leq c < \infty, 0 \leq w$.

Proof: Note that both $\int \varphi(w)H_{n1}(dw|c,h_n)$ and $\bar{H}_n(w|c,h_n)$ are averages of iid random variables, and the condition $\sum_{n=1}^{\infty} exp(-\rho nh_n^s) < \infty$ for $\rho > 0$ implies $nh_n^s \to \infty$.

The proofs of Parts (a) and (b) follow by standard methods used in curve smoothing literature, i.e., splitting each difference into random and bias terms, then calculating variance of the random part and finally using Taylor's expansion on both variance and bias terms. See, for instance, Wand and Jones [23]. In particular, the uniform convergence in Part (a), (ii), follows from Theorem 1 of Stute [20].

□

Lemma 2 Under the conditions of **Lemma 1(a)**, the empirical process

$$D_n(u|c) = \int I(u \le t < \infty)\bar{F}_\pi(t|c)(nh_n^s)^{1/2}\Big(\frac{(1-\varepsilon)H_{n1}(dt|c,h_n)}{(1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon} - \frac{(1-\varepsilon)H_1(dt|c,h_n)}{(1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon}\Big)$$

$$= \int I(u \le t < \infty)\bar{F}_\pi(t|c)(nh_n^s)^{1/2}\alpha_n(dt|c,h_n) \qquad 0 \le u < \infty, \tag{33}$$

converges in distribution, as $n \to \infty$, to the mean-zero Gaussian process $D(u|c), 0 \le u < \infty$, with covariance function

$$\sigma(u,v|c) := \frac{(R(K_0))^s}{f_C(c)}\Big(E\big(\frac{\eta_1\bar{F}_\pi^2(W_1)}{\bar{H}^2(W_1)}I(max(u,v) \le W_1 < \infty)|C=c\big)$$

$$+ E\big((\frac{\eta_1\eta_2\bar{F}_\pi(W_1)\bar{F}_\pi(W_2)}{\bar{H}^2(W_1)\bar{H}^2(W_2)}\bar{H}(max(W_1,W_2))(1-I(W_1 \ge W_2))(1-I(W_1 < W_2))).$$

$$I(u \le W_1 < \infty, v \le W_2 < \infty)|C=c\big)\Big) \qquad 0 \le u,v < \infty,$$

where $W_i = min(U_i,Y_i), \eta_i = I(U_i \le Y_i), i = 1,2$, are independent, as above.

Proof: Note that

$$D_n(u|c) = \int I(u \le t < \infty)\bar{F}_\pi(t|c).$$

$$\frac{(nh_n^s)^{(1/2)}(1-\varepsilon)\Big(((1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon)H_{n1}(dt|c,h_n) - ((1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon)H_1(dt|c,h_n)\Big)}{((1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon)((1-\varepsilon)\bar{H}(t|c,h_n)+\varepsilon)}.$$

Next, note that tightness of $D_n(.|c)$ follows from that of $D_n(0|c)$, and the latter follows by the facts that $\bar{F}_\pi(.|c)$ is bounded and $\bar{H}_n(t|c,h_n)$ in the denominator of the integrand can be replaced, by **Lemma 1(a)**, (ii), by $f_C(c)\bar{G}(t)\bar{F}_\pi(t|c.)$ Further, once we make this replacement the expression for the limiting covariance function follows by **Lemma 1(b)** and elementary covariance calculations.

□

Theorem 3 Denote

$$L_n(u|c) := (nh_n^s)^{1/2}(\bar{F}_n(u|c, h_n) - \bar{F}_\pi(u|c)),$$

$$L_n(\infty|c) := (nh_n^s)^{1/2}(\pi_n(c) - \pi(c)).$$

Then under the conditions of **Lemma 1(a)**, $(L_n(u|c), u \geq 0, L_n(\infty|c))$ converges in distribution, as $n \to \infty$, to the Gaussian process $(L(u|c), u \geq 0, L(\infty|c))$ determined by the equations

$$L(u|c) - \int I(u \leq t < \infty) L(t|c) \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} - L(\infty|c) = D(u|c) \qquad u \geq 0,$$

$$L(0|c) = 0 \leftrightarrow - \int I(0 \leq t < \infty) L(t|c) \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} - L(\infty|c) = D(0|c). \tag{34}$$

Proof: Multiplying both sides of Eq.(30) by $(nh_n^s)^{1/2}$, we get

$$L_n(u|c) - \int I(u \leq t < \infty) L_n(t|c) \Lambda(dt|c, h_n) - L_n(\infty|c)$$

$$= \int I(u \leq t < \infty) \bar{F}_n(t|c)(nh_n^s)^{1/2} \alpha_n(dt|c, h_n)$$

$$+ \int I(u \leq t < \infty) \bar{F}_\pi(t|c)(nh_n^s)^{1/2} \alpha(dt|c, h_n) \qquad 0 \leq u < \infty,$$

$$L_n(0|c) = 0 \qquad n \geq 1. \tag{35}$$

It then follows by **Lemma 2** that the process $(L_n(u|c), u \geq 0, L_n(\infty|c))$ is tight, since $\bar{F}_n(.|c, h_n)$ is bounded, and Eq.(31) and Eq.(32) show that Eq.(35) is invertible. Note that the second term on the right hand side of Eq.(35) is non-random and tends to 0, as $n \to \infty$ by Taylor expansion arguments similar to those in the proof of **Lemma 1**. Further, the latter also means that Eq.(24) holds, i.e.,

$$\Lambda(dt|c, h_n) \to \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} \qquad as \quad n \to \infty.$$

By the above arguments, we conclude from Eq.(35), Lemma 2 and Theorem 3.3.1, p.310, of van der Vaart and Wellner [22] that the limit $(L(u|c), 0 \leq u < \infty, L(\infty|c))$ satisfies Eq.(34).

□

Corollary 1 Denote by $\tau(., .|c)$, the covariance of the limiting Gaussian process $(L(u|c), u \geq 0, L(\infty|c))$ in **Theorem 3**, i.e.,

$$\tau(u,v|c) = E(L(u|c)L(v|c)) \qquad 0 \le u, v \le \infty.$$

Further, define the measure

$$\Lambda_\pi(dt|c) = \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} \qquad 0 \le t < \infty, \qquad\qquad \Lambda_\pi(\infty|c) = 1.$$

Then $\tau(.,.|c)$ is determined by the following equations:

$$\tau(u,v|c) - \int I(t \ge u)\tau(t,v|c)\Lambda_\pi(dt|c) - \int I(s \ge v)\tau(u,s|c)\Lambda_\pi(ds|c)$$

$$+ \int\int I(t \ge u, s \ge v)\tau(t,s|c)\Lambda_\pi(dt|c)\Lambda_\pi(ds|c) = \sigma(u,v|c),$$

$$- \int I(s \ge 0)\tau(u,s|c)\Lambda_\pi(ds|c) + \int\int I(t \ge u, s \ge 0)\tau(t,s|c)\Lambda_\pi(dt|c)\Lambda_\pi(ds|c) = \sigma(u,0|c),$$

$$- \int I(t \ge 0)\tau(t,v|c)\Lambda_\pi(dt|c) + \int\int I(t \ge 0, s \ge v)\tau(t,s|c)\Lambda_\pi(dt|c)\Lambda_\pi(ds|c) = \sigma(0,v|c),$$

$$\int\int I(t \ge 0, s \ge 0)\tau(t,s|c)\Lambda_\pi(dt|c)\Lambda_\pi(ds|c) = \sigma(0,0|c), \tag{36}$$

where $\sigma(.,.,c)$ is the covariance function in **Lemma 2**.

Proof: Follows by writing Eq.(34) once each for $u \ge 0$ and $v \ge 0$, then multiplying the two and taking expectation on both sides. Note that Eq.(36) can in fact be explicitly solved for $\tau(.,.|c)$ by the iteration technique leading to Eqs.(31) and (32). □

3.4 Estimated Covariance Function

We now obtain the empirical version of Eq.(36) by replacing every function by its empirical version, which necessarily has jumps at sample values, referring to Eqs. (27) and (28).

On the left-hand side of Eq.(36), note that

$$\Lambda_\pi(dt|c) = \frac{F_\pi(dt|c)}{\bar{F}_\pi(t|c)} \quad 0 \le t < \infty, \qquad \Lambda_\pi(\infty|c) = 1,$$

can obviously be estimated, at $t = Z_i$ and $t = \infty$, by

$$\frac{(1-\varepsilon)H_{n1}(dt|c,h_n)}{(1-\varepsilon)\bar{H}_n(t|c,h_n)+\varepsilon} = b_i(c), \qquad 1 \le i \le n, \qquad \mathit{and} \quad b_{n+1}(c) = 1,$$

respectively. On the right-hand side, $\sigma(u,v|c)$ can be estimated at $u = Z_i$ and $v = Z_j$, by

$$\sum_{k=1}^{n} a_{ik}a_{jk}\bar{F}_k^2 b_k^2(c) + \sum_{k=1,l=1}^{n} a_{ik}a_{jl}d_{kl}\bar{F}_k^2\bar{F}_l^2 b_k^2(c) b_l^2(c),$$

whereas $\sigma(u,0|c)$ at $u = Z_i$ by

$$\sum_{k=1}^{n} a_{ik}\bar{F}_k^2 b_k^2(c) + \sum_{k=1,l=1}^{n} a_{ik}d_{kl}\bar{F}_k^2\bar{F}_l^2 b_k^2(c) b_l^2(c),$$

and so on, where

$$d_{kl} = (1-a_{kl})(1-a_{lk})\sum_{r=1}^{n} a_{kr}a_{lr} \qquad 1 \le k,l \le n.$$

Hence, denoting the matrices

$$\mathbf{V} = ((\hat{\tau}(Z_i, Z_j|c)))_{1\le i,j\le n+1}, \mathbf{D} = ((d_{ij}))_{1\le i,j\le n}, \bar{\mathbb{F}} = diag(\bar{F}_1, ..., \bar{F}_n),$$

where $Z_{n+1} = \infty$, and using the notation of Eqs. (27)-(28), we get the equation

$$\begin{bmatrix} \mathbf{I} - \mathbf{A}\mathbf{B}_c \\ -\mathbf{b}_c^T \end{bmatrix} \mathbf{V} \begin{bmatrix} (\mathbf{I} - \mathbf{A}\mathbf{B}_c)^T & -\mathbf{b}^T \end{bmatrix} = \begin{bmatrix} \mathbf{A}_0\mathbf{B}_{0c}\bar{\mathbb{F}} \\ \mathbf{b}_{0c}^T\bar{\mathbb{F}} \end{bmatrix} (\mathbf{I} + \mathbf{B}_{0c}\mathbf{D}\mathbf{B}_{0c}) \begin{bmatrix} \bar{\mathbb{F}}\mathbf{B}_{0c}A_0^T & \bar{\mathbb{F}}\mathbf{b}_{0c} \end{bmatrix},$$

where ^{T} denotes the matrix transpose, $\mathbf{A}_0$, $\mathbf{B}_{0c}$ are obtained by deleting the $(n+1)$-st row and column of $\mathbf{A}$, $\mathbf{B}_c$, respectively, and $\mathbf{b}_{0c}$ is obtained from $\mathbf{b}_c$ by deleting its $(n+1)$-st component (which is 1).

3.4.1 Optimal Order of Bandwidth h_n

From Eq.(30), it is clear that the convergence of $(\bar{F}_n(.|c, h_n) - \bar{F}_\pi(.|c), \pi_n(c) - \pi(c))$ is controlled by the random part ('variance' term)

$$v_n := \int I(u \le t < \infty)\bar{F}_\pi(t|c, h_n)\alpha_n(dt|c, h_n),$$

and the deterministic part ('bias' term)

$$B_n = \int I(u \le t < \infty)\bar{F}_\pi(t|c, h_n)\alpha(dt|c, h_n).$$

Since, under assumption **A1**, **A2** of **Lemma 1**,

$$var(V_n) = O(\frac{1}{nh_n^s}), \qquad B_n^2 = O(h_n^4),$$

it follows that the optimal order of h_n, minimizing the mean squared error of a fixed c, is

$$h_n = O(\frac{1}{n^{\frac{1}{s+4}}}),$$

which is the usual optimal order in kernel-smoothing.

3.4.2 Optimal Choice of h_n via Cross-Validation

One of the common ways to find the optimal h_n is minimizing the mean integrated squared error (MISE). The general form for MISE is

$$MISE = \int E(f_n(x) - f(x))^2 dx,$$

where f_n is the density estimator of f. Rudemo [17] and Bowman [2] proposed least squares cross-validation,

$$CV_n = \int \hat{f}^2(x) - \frac{1}{n}\sum_{j=1}^{n} \hat{f}_j,$$

where $\hat{f}_j$ is the kernel density estimator with j-th observation deleted from sample.

We propose to make an optimal, data-based choice of h_n via cross-validation, namely by minimizing the criterion

$$CV_n(h|c) := \sum_{i=1}^{n}(\pi_{in}(c) - \pi_n(c))^2,$$

where $\pi_{in}(c)$ is the estimator of $\pi(c)$ obtained from the same sample, but with $i-$th data-point deleted, i.e., $\pi_{in}(c)$ is based on $(\delta_j, Z_j, j \neq i)$ for $1 \leq i \leq n$.

3.5 Simulation

In this section of thesis, a simulation study is conducted to see how the theory works on the data. For the ease of calculation and make it more visual, we assume that there is only one covariate in the model. A sample of $n = 100$ independent random variables C_i (covariates) from Exponential(1) has been generated. Variables X_i's are dependent on the covariates and they follow the distribution function Exponential(c). Censoring variables Y_i's are generated from the distribution of Exponential(0.2). The cure rate function is taken to be $\pi(c) \equiv \exp(-\exp(-c))$. Figure 3 shows the behaviour of cure rate estimator versus different values of covariate. The red and blue dots show the cure rate and the cure rate estimator using the proposed nonparametric model, respectively.

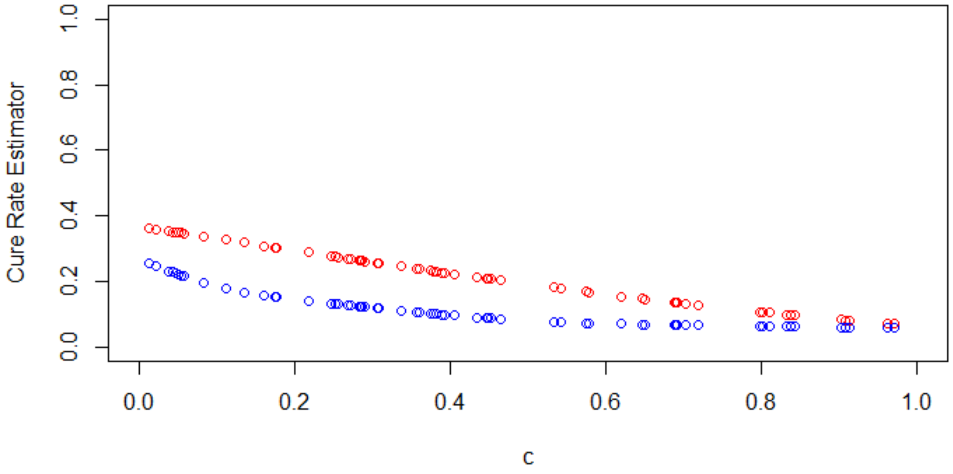

Figure 3: Cure rate estimation for different value of covariates

Using the cross-validation method, for a fixed value of $c = 0.2952$, different bandwidths have been chosen for the model. Based on the smallest CV_n, the best bandwidth is chosen as $h = 0.2$.

Bandwidth	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
CV_n	0.2772	0.2067	0.2144	0.2359	0.2712	0.3066	0.3322	0.3470	0.3543	0.3584

Table 3: Table of CV_n values based on different bandwidths

3.6 Conclusion

In this research work, we proposed a new multivariate cure rate estimator under random censoring when the cure rate is a function of covariates. This estimator in fact is an extension to Xu and Peng [24] estimator. In this research work, the kernel smoothing is used into the eigenvector equation to give us a smoother estimator.

One of the advantages of this estimator, compared to Tsodikov method [21], is that it is applicable on discrete and continuous covariates. In this chapter, not only the asymptotic distribution of the model has been found but also the optimal order of bandwidth and the optimal bandwitdth choice have been obtained through the cross-validation method.

Acknowledgments

This research was partially supported by NSERC.

Chapter 4

A Nonparametric Test for the Presence of Immunes in the Univariate Case

4.1 Introduction

In time-to-event data, there is a possibility that immunes exist. The presence of immune individuals has an important role in the analysis of survival data. Cured or immune individuals are defined as the ones who are not subject to the event of interest (e.g. death). Since we consider only a specific period of time, there is censoring in the study, which is one of the most important issues in survival analysis.

Let $X_i, 1 \leq i \leq n$, be independent and identically distributed (iid), non-negative random variables, each having distribution function $F(x) = P(X \leq x)$ and survival function $S(x) = \bar{F}(x) = P(X \geq x)$. There are also an independent set of censoring variable $Y_i, 1 \leq i \leq n$, with distribution function G. Assume that we observe

$$Z_i = min(X_i, Y_i), \qquad \delta_i = I(X_i \le Y_i) \qquad\quad 1 \le i \le n,$$

Consider the following model

$$F(t) = (1-p)F_0(t),$$

where $F_0(t)$ is the baseline survival function and p is the probability of cure. As discussed in Chapter 2, Maller and Zhou [11] proposed $\hat{p} = 1 - \hat{F}_n(t_n)$ as a consistent estimator for p, where $\hat{F}_n(\cdot)$ is the Kaplan-Meier estimator of $F(\cdot)$ and t_n is the maximum observed failure or censored time. They also proved that under some conditions,

$$\sqrt{n}(\hat{p} - p) \to N\left\{0, p^2 \int_0^t \frac{dF(s)}{(1-F(s))^2(1-G(s))}\right\} \qquad for \quad t > 0.$$

Since the limiting variance depends on p, a question arises here that what happens if there is no cure rate in the model. If $p = 0$, the limiting distribution is degenerate so it is important to test the existence of cure rate in the model. Maller and Zhou [12] addressed this problem with the following null hypothesis,

$$H_0 : p = 0, \quad \text{against} \quad H_1 : p > 0.$$

They assumed that baseline survival function is the exponential survival function with parameter λ. $\hat{\theta} = (\hat{p}, \hat{\lambda})$ is the maximum likelihood estimates (MLE) under the mixture model and $\hat{\theta_0}$ is the MLE under H_0. They introduced $d_n = -2(l_n(\hat{\theta_0}) - l_n(\hat{\theta}))$ statistic to test the null hypothesis, where l_n is the log-likelihood function.

The limiting null distribution of the test-statistic was found to be a 50-50 mixture distribution of chi-square random variable with 1 degree of freedom and a probability distribution degenerate at 0. If $d_n \geq C_{1-\alpha}$, the null hypothesis of no immunes is rejected, where

$$\frac{1}{2} + \frac{1}{2} P(\chi_l^2 \leq C_{1-\alpha}) = 1 - \alpha.$$

In most of the studies the presence of cured patients was studied under the gamma distribution. Peng et al. [16] did a simulation study to find the asymptotic null distribution of the likelihood ratio test (LRT) for presence of cured patients under Weibull and log-normal mixture models. They found that the results from Weibull, log-normal and gamma are very close to each other and they have approximately the same asymptotic null distribution.

In 2007, Sen and Tan [19] considered a nonparametric estimator of cure rate under the mixture model and Case-1 interval censoring. Their proposed estimator is based on the non-parametric MLE and degenerates like Maller and Zhou's [11], when there is no cure in the model.

In this chapter, we propose a test-statistic for the hypothesis of no cure, based on Poisson convergence of censored empirical processes when the F is in the max domain of attraction of some extreme-value distribution.

4.2 The Proposed Test-statistic

First we need to consider some assumptions.

Assumption 1 Suppose there exists a sequence of constants $a_n > 0$ and b_n such that $\frac{\max(X_1,\ldots,X_n)-b_n}{a_n}$ has a non-degenerate limit distribution G_0 as $n \to \infty$. This means that for every continuity point x of G_0 we have

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_0(x),$$

where G_0 is an extreme value distribution (EVD). Consequently,

$$\lim_{n\to\infty} (1 - \frac{n(1-F(a_n x + b_n))}{n})^n = G_0(x),$$

for every continuity point x of G_0. It follows that

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -\log G_0(x).$$

Theorem 4 If $(1-G) = (1-F)^{\alpha}$ (Koziol-Green model) for some $\alpha > 0$, and if there are constants $a_n(\alpha) > 0, b_n(\alpha)$ such that

$$n\Big(1 - F(a_n(\alpha)x + b_n(\alpha))\Big)^{\alpha+1} \to \Big(-\log G_0(x)\Big)^{\alpha+1}.$$

Then under the assumption of no cure rate in the model $(H_0 : p = 0)$,

i) $N_n(x) = \sum_{j=1}^{n} I(Z_j \geq a_n(\alpha)x + b_n(\alpha)) \xrightarrow{d} N'(x)$ where $N'(.)$ is a Poisson process with mean $(-\log G_0(.))^{\alpha+1}$,

ii) $N_{1n}(x) = \sum_{j=1}^{n} \delta_j I(Z_j \geq a_n(\alpha)x + b_n(\alpha)) \xrightarrow{d} N_1'(x)$ and $N_1'(x) \overset{d}{=} \sum_{j=1}^{N'(x)} \eta_j$, where $\eta_1, \eta_2, \ldots$ are iid with Bernoulli distribution with mean $\frac{1}{\alpha+1} = E(\delta)$ and it can be estimated by $\frac{1}{n}\sum_{i=1}^{n} \delta_i$.

iii) $\eta_1, \eta_2, \ldots$ and $N'(.)$ are independent.

Proof:

i) For proof, see Charras and Lezaud [3] and Embrechts et al. [7].

ii) and iii) For proof, see Sen and Tan [19].

□

Theorem 5 Under the null hypothesis of no cure rate and previous assumptions, we construct the following test statistic which has asymptotic standard normal distribution,

$$T_n(x_n) = \Big(\frac{\sum_{i=1}^n \delta_i I(Z_i \geq x_n)}{\sum_{i=1}^n I(Z_i \geq x_n)} - \hat{c}\Big)\sqrt{\frac{\sum_{i=1}^n I(Z_i \geq x_n)}{\hat{c}(1-\hat{c})}} \to N(0,1),$$

where $\hat{c} = \frac{1}{\hat{\alpha}+1} = \frac{1}{n}\sum_{i=1}^n \delta_i$, provided x_n is chosen to satisfy as $n \to \infty, x_n \to \infty$ and $n(1 - F(x_n)) \to \infty$. Furthermore, if $T_n < -z_\alpha$, where $P(N(0,1) < -z_\alpha) = \alpha$, the null hypothesis is rejected.

Note that we are yet to find a method for optimal choice of x_n.

Proof: Note that the weak convergence in Theorem 4 can actually be strengthened to a strong convergence as follows.

Based on the Theorem 2 in Sen and Tan [19], one can construct a sequence of Poisson processes N_1' and N' on $\mathbb{R} \times \mathbb{R}$ such that as $n \to \infty$,

$$\sup_x |N_{1n}(x) - N_1'(x)| \to 0,$$
$$\sup_x |N_n(x) - N'(x)| \to 0.$$

Now

$$\lim_{n\to\infty}\sqrt{\frac{\sum_{i=1}^{n} I(Z_i \geq x_n)}{\hat{c}(1-\hat{c})}}\Big(\frac{\sum_{i=1}^{n}\delta_i I(Z_i \geq x_n)}{\sum_{i=1}^{n} I(Z_i \geq x_n)} - \hat{c}\Big)$$

$$= \lim_{n\to\infty}\sqrt{\frac{N(x_n)}{\hat{c}(1-\hat{c})}}\Big(\frac{N_1(x_n)}{N(x_n)} - \hat{c}\Big)$$

$$= \lim_{n\to\infty}\sqrt{\frac{N(x'_n)}{\hat{c}(1-\hat{c})}}\Big(\frac{N_1(x'_n)}{N(x'_n)} - \hat{c}\Big), \tag{37}$$

where $x'_n = \frac{x_n - b_n}{a_n}$.

By Theorem 3, part (a) of Sen and Tan [19], as $n \to \infty, (-\log G_0(x'_n))^{\alpha+1} \to \infty$ and $\frac{N'(x'_n)}{(-\log G_0(x'_n))^{\alpha+1}} \to 1$. Since $N'_1(x) = \sum_{j=1}^{N'(x)} \eta_j$ and $(\eta_1, \eta_2, ...)$ are iid Bernoulli $(\frac{1}{\alpha+1})$, independent of $N'(.)$, using Assumption 2 and the random central limit theorem, we have the result.

□

4.3 Simulation

A sample of $n = 200$ independent random variables, X_i from Exponential(1) and Y_i from Exponential(0.5) are randomly selected. The following graph shows how T_n behaves for different values of p when $F \approx Exp(1)$ and $G \approx Exp(0.5)$.

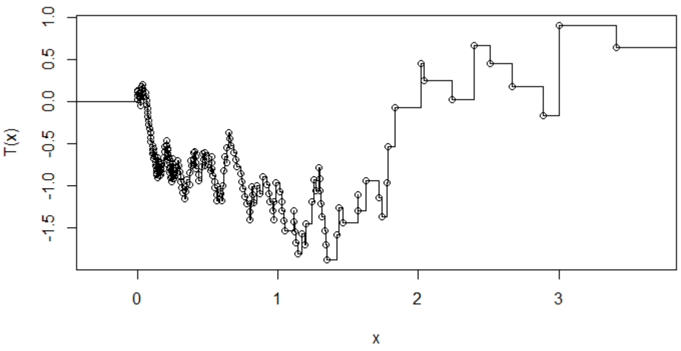

Figure 4: The behaviour of $T_n(x)$ when $p = 0, F \approx Exp(1)$ and $G \approx Exp(0.5)$

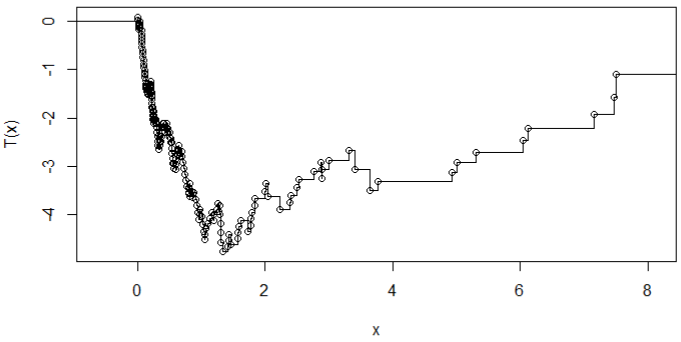

Figure 5: The behaviour of $T_n(x)$ when $p = 0.25, F \approx Exp(1)$ and $G \approx Exp(0.5)$

We also applied the proposed statistic to the data from AMS study [8]. Twenty-six eligible patients are randomly assigned to receive either maintenance chemotherapy or to receive no maintenance therapy. The proposed statistic is illustrated on this data as follows:

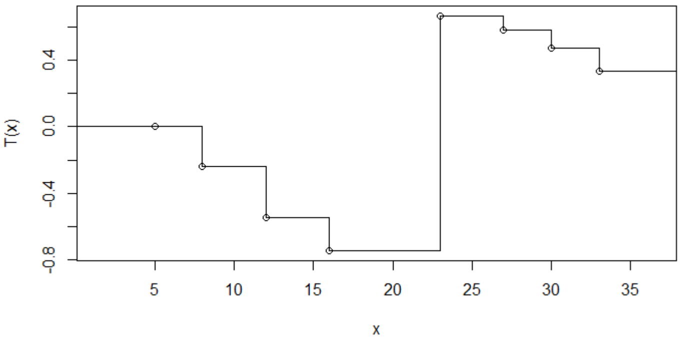

Figure 6: The behaviour of $T_n(x)$ for AML data (Non-Maintained)

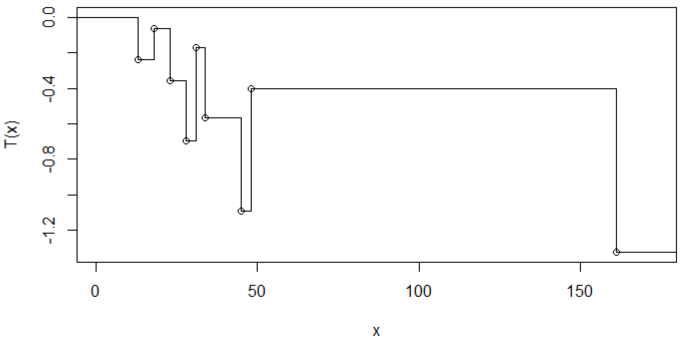

Figure 7: The behaviour of $T_n(x)$ for AML data (Maintained)

In Figure 7, although all the values of T_n are greater than -1.96, for $\alpha = 0.05$, and the null hypothesis cannot be rejected for the maintained AML data, we should note that all T_n values are below zero, as is expected when cure-rate is positive. In both AML maintained and non-maintained data, sample sizes are too small, so nothing can be concluded conclusively.

4.4 Conclusion

In this chapter, we proposed a test for the absence of a cure rate. Our approach is based on Poisson convergence of censored empirical processes. We proposed a new test-statistic to test the existence of cure in the study. Our proposed statistic has a normal distribution with mean of zero and variance of one. A simulation is also conducted to see the behaviour of proposed statistic.

Acknowledgments

This research was partially supported by NSERC.

Chapter 5

Conclusion

5.1 Concluding Remarks

In Chapter 2, we proposed a new non-parametric multivariate cure rate estimator under random censoring. We found that its asymptotic distribution is normal and we proposed a variance estimator applicable under the sufficient follow-up condition. It is demonstrated that the proposed model is robust and can be easily calculated. A simulation study is conducted to support the theoretical results. We applied our proposed model to the data from the Litter-matched tumorigenesis experiment [14]. The estimation of cure rate and variance for the Litter-matched tumorigenesis experiment are obtained.

In Chapter 3, a new multivariate cure rate estimator with covariates under random censoring is considered. The proposed estimator is based on the non-parametric approach and in fact, it is the extension of Xu and Peng [24] estimator to the multivariate survival time. Using cross-validation, the asymptotic distribution, variance estimator and the optimal order of bandwidth for the proposed multivariate cure rate estimator are obtained. A kernel smoothing method has been used to smooth the proposed estimator.

In Chapters 2 and 3, the asymptotic distribution and covariance functions of the estimator have been obtained assuming that immunes exist. If the immunes do not exist, limiting distribution is degenerate, so it is very important to find out if the cure rate exists. In Chapter 4, a test-statistic regarding the presence of immunes in the univariate case is proposed and the limiting distribution of the test-statistic is obtained based on Poisson convergence of censored empirical processes and extreme value theory.

5.2 Suggestion for Future Research

One of the suggested future work is to use Chaubey-Sen Poisson smoothing method [4] instead of kernel smoothing in estimation of the nonparametric multivariate cure rate with covariates. Their proposed method has solved many issues which exist in the other methods.

The kernel smoothing method can be used for estimating functions with non-negative random variables such as survival function and hazard function. In kernel smoothing, some modifications have to be applied on the kernel to avoid the possible probability of negative values. These lead to affect the bias and rate of convergence. However, in Chaubey-Sen smoothing method, this condition is eliminated.

Another restriction of using kernel smoothing is the boundedness of the second derivative of the density function which in Chaubey-Sen smoothing model, is not necessary.

In the future, the proposed estimator can be applied on the first-hand survival-analysis data where cure is a possibility, such as criminal recidivism, smoking cessation, etc.

In this thesis, simulations are presented for illustrative purposes and the strength of the whole thesis lies on the theoretical aspects. Potential tests such as finite sample normality assessment and coverage tests for confidence intervals could be added to the list of potential future work. Another potential work could be to use other distributions than the exponentials. The results could be compared with other estimators.

Bibliography

[1] Beran, R., 1981. Nonparametric regression with randomly censored survival data. Tech. rep., Technical Report, Univ. California, Berkeley.

[2] Bowman, A. W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360.

[3] Charras-Garrido, M., Lezaud, P., 2013. Extreme value analysis: an introduction. Journal de la Société Française de Statistique 154 (2), 66–97.

[4] Chaubey, Y. P., Sen, P. K., 1996. On smooth estimation of survival and density functions. Statistics & Risk Modeling 14 (1), 1–22.

[5] Chen, M.-H., Ibrahim, J. G., Sinha, D., 2002. Bayesian inference for multivariate survival data with a cure fraction. Journal of Multivariate Analysis 80 (1), 101–126.

[6] Dabrowska, D. M., 1987. Non-parametric regression with censored survival time data. Scandinavian Journal of Statistics 14 (3), 181–197.

[7] Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling extremal events: for insurance and finance. Vol. 33. Springer Science & Business Media.

[8] Embury, S. H., Elias, L., Heller, P. H., Hood, C. E., Greenberg, P. L., Schrier,

S. L., 1977. Remission maintenance therapy in acute myelogenous leukemia. Western Journal of Medicine 126 (4), 267.

[9] Kalbfleisch, J. D., Prentice, R. L., 2011. The statistical analysis of failure time data. Vol. 360. John Wiley & Sons.

[10] Kaplan, E. L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American statistical association 53 (282), 457–481.

[11] Maller, R. A., Zhou, S., 1992. Estimating the proportion of immunes in a censored sample. Biometrika 79 (4), 731–739.

[12] Maller, R. A., Zhou, S., 1995. Testing for the presence of immune or cured individuals in censored survival data. Biometrics 51 (4), 1197–1205.

[13] Maller, R. A., Zhou, X., 1996. Survival analysis with long-term survivors. Wiley New York.

[14] Mantel, N., Bohidar, N. R., Ciminera, J. L., 1977. Mantel-haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. Cancer Research 37 (11), 3863–3868.

[15] Nieto-Barajas, L. E., Yin, G., 2008. Bayesian semiparametric cure rate model with an unknown threshold. Scandinavian Journal of Statistics 35 (3), 540–556.

[16] Peng, Y., Dear, K. B., Carriere, K., 2001. Testing for the presence of cured patients: a simulation study. Statistics in medicine 20 (12), 1783–1796.

[17] Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics 9 (2), 65–78.

[18] Sen, A., Stute, W., Under revision. The multivariate kaplan-meier estimator.

[19] Sen, A., Tan, F., 2008. Cure-rate estimation under case-1 interval censoring. Statistical Methodology 5 (2), 106–118.

[20] Stute, W., 1986. On almost sure convergence of conditional empirical distribution functions. The Annals of Probability 14 (3), 891–901.

[21] Tsodikov, A., 2001. Estimation of survival based on proportional hazards when cure is a possibility. Mathematical and Computer modelling 33 (12-13), 1227–1236.

[22] Van Der Vaart, A. W., Wellner, J. A., 1996. Weak convergence. In: Weak convergence and empirical processes. Springer, New York.

[23] Wand, M., Jones, M., 1995. Kernel smoothing. 1995. Chapman & Hall, London.

[24] Xu, J., Peng, Y., 2014. Nonparametric cure rate estimation with covariates. Canadian Journal of Statistics 42 (1), 1–17.