Novel Nonparametric Methods for Event Time Data

by

Dipankar Bandyopadhyay

(Under the direction of Somnath Datta)

Abstract

This dissertation develops a number of nonparametric inference procedures for time to event data under right censoring. The type of event time data that we consider are broader than the usual survival time data. They include some multistage models such as the competing risk and also certain mark variables and covariates. More specifically, we address the issue of testing the equality of two or more survival distributions when the population membership information is not available for the right censored individuals. This can also be regarded as testing the independence of failure time and cause in a competing risk problem or, more generally, as testing the independence of a failure time and a mark variable. We introduce a family of weighted log-rank tests based on the concept of assigning only a fraction of a censored individual to each 'at risk' set of failures. The joint asymptotic normality of our test statistics is explored through an asymptotic linear representation. In addition, two resampling schemes are suggested as alternatives to the asymptotic distribution which might be more useful in practice.

Another major accomplishment of this dissertation is to formulate a class of U-statistics for right censored data problems that are based on the concept of data reweighting and are valid for a kernel of arbitrary order. They are useful for asymptotically unbiased estimation

for certain mean-like functionals of the failure time distributions. We obtain a martingale representation for these statistics which leads to their asymptotic normality. The issue of efficiency gain through the use of a doubly robust version is also discussed. As a motivating application of this estimation methodology, we construct a second test statistic for the testing problem described earlier.

The finite sample properties of our estimators and statistical tests are studied through extensive simulations and two well known data sets from existing literature are used for their illustrations to real data problems.

INDEX WORDS: Competing risk model, Dependent censoring, Doubly-robust, Fractional risk, Kaplan-Meier, Kendall's tau, Log-rank tests, Multistate models, Right-censoring, U-statistic.

Novel Nonparametric Methods for Event Time Data

by

Dipankar Bandyopadhyay

B.Sc., R.K. Mission Residential College (University of Calcutta, India), 1998

M.Sc., University of Calcutta, 2000

M.S., University of Georgia, 2003

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2006

Novel Nonparametric Methods for Event Time Data

by

Dipankar Bandyopadhyay

Approved:

Major Professor: Somnath Datta

Committee: Daniel Hall
Nicole Lazar
Mary Meyer
Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2006

Dedication

To the loving memory of Sri Mrinal Kanti Basu (MKB), ex-lecturer of statistics at Ramakrishna Mission Residential College, Narendrapur, India whose endless inspiration and encouragement had made me proceed this far. I owe more to him than what I realized when he was alive.

Acknowledgments

First and foremost, my sincere thanks goes to my advisor, Professor Somnath Datta. Let me add here that my enthusiasm alone would not have made me choose a career in academics; Somnath is overwhelmingly responsible for that. He was extremely supportive throughout the entire duration of the Ph.D. program even in the last year when he left for his new job at Louisville. I recall once Somnath stepped into my office and told me to get my hands dirty with computations. I had great fear in handling computer codes at the onset but gradually with his encouragement, I became familiar with them in this present boom of research in computational statistics. I am also grateful for several research insights he provided while writing this dissertation and his wise direction that made the completion of this dissertation a successful and rewarding endeavor. His enthusiasm, optimism, and confidence will continue to guide me through good and bad times in my future professional life.

I would also like to thank my committee members Dr. Dan Hall, Dr. Mary Meyer, Dr. Jaxk Reeves and Dr. Nicole Lazar for agreeing to read this dissertation and provide wonderful advice. Special thanks to Dr. Glen Satten for providing me several suggestions through email.

My grateful thanks goes to Professors Uttam Bandyopadhyay, Kalyan Das and Late Adhir Basu at Dept. of Statistics, Calcutta University, India for their encouragement and their recommendations while applying for the Ph.D. program. I also express my sincere thanks to Kaushik Chakraborty and Sudipto Chakraborty for providing me financial support during the application process. It is mainly because of them that I got interested in pursuing my Ph.D program in the US.

I am grateful to Marlow Lemons and Lewis Jordan for providing invaluable support in the transition and adjustment to the new American system. I would also like to thank my wonderful UGA 'statistics' friends, specially Archan, Desale, Tanujit, Katsuo, Xu Hui, Jin-Hong Park, Ling Ling and Haitao for making the department very interesting and international. I acknowledge with sincere thanks the nice time spent with Ananda and Susanta, especially the great Friday night dinners and occassional tea parties that provided enough recreation in the midst of academic boredom.

Special thanks are due to my wife Mridula who remained very patient during the writing phase and provided endless love and mental support.

I am specially thankful to our head Prof. John Stufken for supporting me to attend various statistical meetings with travel grants. He was always there to answer my queries. The smiling faces of our staff Connie, Daphney and Loretta kept me energetic at work.

I am grateful to the UGA Graduate School for awarding me a 'Dissertation Completion' assistantship that gave me enough time in the final year to concentrate on my research.

Finally and most importantly, I recall the support of my parents, their sacrifices during times of thick and thin, their unconditional love towards me and their sincere respects for academics. I am extremely lucky to learn from them the essense of education as an honorable goal of life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Event Time Data

1.1 Event-time data

In general, "event time data" is a term used for describing a category of data that measure the time to some event of interest. In such a study, a collection of individuals moves among a finite (usually small) number of states/stages. In such cases, the event is a transition from one stage to another. For example, it can be the time to an epileptic seizure in a medical study, time to failure of a unit (or component) in reliability, and so on. Possible stages in this context can be described as death, event, failure or transition to denote what happens at the response time. Time to an event is considered a positive real-valued variable having a possibly continuous distribution. The data structure can be (a) univariate or (b) multivariate. When all time variables describe the time to the same type of event for various individuals, the data are univariate, else when time variables describe multiple event times per individual, the data are considered multivariate. In such a case, we do not assume independence between event times. In addition to the survival or event times, there are often observable covariables that can affect both the future events of interest and the censoring variable. While the value of a time independent covariate is fixed and known at the beginning of a study, the modeling of the effect of a time dependent covariate in terms of a stochastic process is much more complex.

1.2 Various incompleteness in data

Although the exact transition times (in continuous time) form the modeling basis of the time-to-event data, often the times are incompletely observed. A host of incompletenesses are encountered in practice most important being censoring (right and left) and truncation (right and left). A brief mathematical description of various forms of censoring and truncation might be helpful here.

Consider, for a specified individual under study, its true lifetime denoted by X and a right censoring time denoted by C_R. Generally, unless covariate information is available, the X's are assumed to be independent and identically distributed. The exact lifetime X of an individual will be known if, and only if, X is less than or equal to C_R. If X is greater than C_R, the only available information is that the individual is a survivor beyond C_R and the individual is thus right-censored. There can be several forms of right censoring such as progressive type I, generalized type I, type II censoring, competing risk censoring, and so on. An individual in a study is considered to be left-censored if the only information we have is that its lifetime X is less than a left censoring time C_L. In other words, the event of interest had already occurred for the individual before the person is observed in the study at time C_L. If a study has both left-censored and right-censored data, then it is called doubly censored data. Another notable feature arising in many survival studies is truncation. It occurs in those survival data when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed. An individual whose event time is not within this interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring where we have at least partial information on each subject, i.e. we have at least the information that an individual has survived till that censored time point. Because we are only aware of individuals with event times in the observational window, the inference for truncated data is restricted to conditional estimation.

Similar to censoring, there can be right truncation, left truncation and interval truncation. Klein and Moeschberger (2003) provides a detailed description of several censoring-truncation mechanisms. More severe forms of censoring deal with current status and interval censored data. The former is a special case of the latter in which individuals are not monitored constantly. Current status data represent the status of individuals inspected at a single inspection time (i.e., a single snapshot) per individual. In the context of pure survival data, it therefore records whether at the inspection time, denoted by C, a given individual failed, or not. A form of interval censoring results in multistage models when each individual is observed at multiple inspection times (i.e. multiple snapshots).

1.3 Multistage models

Multistage models are a type of multivariate time-to-event data in which individuals (or experimental units) move through a succession of 'stages' corresponding to distinct states. For example, in a multistage model for cancer progression, individuals could move from a cancer-free stage through stages of increasing tumor severity, or could move to a state representing death from any of the other stages. Multistage models with a finite number of stages can be divided into two classes. In acyclic networks, each stage is entered at most once; in cyclic networks, reentry into the same stage is allowed. Acyclic networks can be exploded into an equivalent finite network which has a tree topology. Hence we will assume, without loss of generality, that the network is a tree. Two central questions in studying multistage models are (a) What is the probability that a randomly selected person is in stage j at time t? and (b) What is the average length of time a person entering stage j spends before moving to the next stage? More formally, the first of these questions asks what are the stage occupation probabilities and the second question asks what is the mean waiting time in stage j. Further details about multistage models can be found in Datta, Satten and Datta (2000a).

There are several special cases of multistage models. Traditional survival analysis is the simplest example of a multistage model where individuals begin in an initial stage (alive, say) and may move irreversibly to a second stage (death, say). The competing risk model is another important example of a multistage model where a particular individual in a stage can move to a second stage (death, say) because of several (k, say) rival causes. The competing causes can be dependent or independent depending on the specified problem. For an example, consider the data described in Hoel (1972). In a laboratory experiment designed to study the effect of radiation on life length, two groups of RFM strain male mice were given a radiation dose of 300 rad at an age of 5-6 weeks. The first group of 95 mice lived in a conventional laboratory environment while the second group lived in a germ free environment. After the death of the mice, necropsy was performed to ascertain whether the cause of death was thymic lymphoma, reticulum cell sarcoma (both specific types of cancer), or other causes. Another simple but well-known example of a multistage model is the three stage illness-death model in which individuals can move from an initial stage (well) to either the illness or death stages. In the irreversible version of this model, persons in the illness stage may subsequently only move to the death stage; in the reversible version, ill persons may recover and move back to the well stage. Datta, Satten and Datta (2000b) consider nonparametric estimation of stage occupation probabilities in a three stage irreversible model.

1.4 Estimation approaches for event time data

Censored data problems, in various forms, arise frequently, if not always, in a variety of fields, notably in biomedical research, in reliability and life testing, and also in market research. More specifically, we are interested in different forms of censored 'event-time' data stemming from the inability of the investigators to collect complete data due to design constraints or for reasons beyond their control (e.g. loss to follow up). In the beginning, parametric approaches

were developed to explain the survival mechanisms but gradually a transition to nonparametrics occured because the parametric counterparts were often based upon unverifiable assumptions and were thus prone to errors due to model misspecifications. For event history analysis, a major contribution was the 1975 Berkeley dissertation 'Statistical Inference for a family of Counting Processes' written by Odd Aalen, where the basic nonparametric statistical problems for censored data were studied in terms of the conditional intensity of the counting process that records the events as time proceeds. Aalen showed that using the martingale representation of a counting process, one can have a unified treatment to the analysis of censored data models.

In a general multistage model, there can be broadly two types of estimation approaches. We first describe the re-weighting approach which has its roots in sample survey. Generally speaking, all the basic quantities we seek to estimate would be easily estimated in the absence of right censoring. For example, stage occupation probabilities can be estimated by the empirical proportion of persons in each stage at any given time. The first step is to express the uncensored-data estimator in a linear (or average) form. The next step is to model the process that governs the censoring process. Because each individual is censored at most once, standard (univariate) survival process models can be used, such as Cox's proportional hazards models or Aalen's linear hazard models. We use the model of the hazard of being censored to weigh the observed (censored) data to reconstruct what the uncensored experiment would have produced. The details of this approach for the estimation of the marginal survival curve is given in Satten, Datta and Robins (2001). The common thread in this development is 'approximate unbiasedness' which makes the censored data answers match the corresponding complete data answers 'on the average'. This is based on the pioneering work by Robins and Rotnitzky (1992), Robins (1993).

The second approach appears to be more traditional and there one models the hazard of future events of interest. In a general multistage model the past history of the process,

including the current stage, may affect its future evolution. It is theoretically possible to model this dependence to obtain transition hazards between stages that are conditional on an individual's past history. In the Markovian version such dependence is explained only through the currently occupied stage information. However, such conditional analyses do not easily lead to answers to the marginal questions since 'unconditioning' or 'marginalizing' by averaging conditional hazards can be quite difficult.

Multistage models are often analyzed by fitting one of two models: a Markov model or a semi-Markov model. In a Markov model, the past history of the process does not affect its future evolution given the present stage. In a semi-Markov model, waiting times in each stage are independently distributed. Nonparametric estimators of marginal quantities are available for both of these models (Aalen and Johansen, 1978; Fleming, 1978a, 1978b; Lagakos *et al.*, 1978). Recently, nonparametric estimators have been given that do not make these structural assumptions (Datta and Satten, 2001; Satten and Datta, 2002).

Even if there were no 'external' covariates, there are still 'internal' covariates generated by an individual's past history that may affect future transitions and censoring hazards (Cox, 1972). Another feature of the multistage model is that even if censoring time is independent of the failure time, dependent censoring may be induced at the marginal level. This occurs when there is correlation between failure and censoring times. If there is no additional information on the nature of this correlation, then the problem is intractable (as for each individual, only the earlier of the failure and censoring times is observed, Tsiatis, 1975). However, if this dependence is carried by covariates so that for fixed levels of covariates, failure and censoring times are uncorrelated, then it is possible to account for dependent censoring (Robins and Rotnitzky, 1992; Robins, 1993; Robins and Finkelstein, 2000; Satten, Datta and Robins, 2001).

1.5 Models for competing risk

In the next chapter of this dissertation, we focus on a population of event times consisting of various sub-population of event times. Once such important model where this arises is called the competing risk or multiple decrement model. For our purpose, it serves as an example of a multistage model where the terminal stage is determined by the cause of failure and corresponds to various subtypes of individuals.

Competing risk theory dates back to a memoir read in 1760 by Daniel Bernoulli before the French Academy of Sciences and published in 1765. Typically, in real life, we can think of competing causes of failure as several causes in action on a particular system and the occurrence of one inhibits the occurrence of the other. Thus, one observes for each unit simply a failure time and the cause of failure (or possibly censoring if there be any). Gail (1975) provides a review of various models used in competing risk analysis.

Problems involving competing risks are quite common in medical and reliability applications. In cancer studies, common competing risks are relapse and death in remission (or treatment related mortality). Interest often lies in estimating the rate of occurrence of the competing risks, comparing these rates between treatment groups and modeling the effect of covariates on the rate of occurrence of the competing risks. In reliability, competing risks arise in the analysis of series systems of components where the failure of any component will lead to system failure. One observes the time at which the system fails and which component caused the system to fail. A nonparametric maximum likelihood estimator (NPMLE) for the competing risk problem, along with martingale interpretations was proposed by Aalen (1976) under the name 'multiple decrement models'. These models/methods can be thought of as a special case of the Aalen-Johansen theory of estimation of time-inhomogeneous Markov processes (Aalen and Johansen, 1978).

Competing risks are typically represented by a set of positive random variables $X_1, \ldots, X_J$ where X_j is the potential (possibly unobservable) time to occurrence of the jth cause of failure. We observe T=min$(X_1, \ldots, X_J)$ and an indicator δ which tells which of the J risks caused the failure, i.e. $\delta = j$ if $T = X_j$. A basic quantity in competing risk theory is the cause specific hazard rate (CSHR), $\lambda_j(t)$, which is the rate of occurrence of the jth failure cause in the presence of all causes,i.e.

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t, \delta_i = j | T \geq t)}{\Delta t} \tag{1.5.1}$$

The cumulative cause-specific hazard function is defined as $\Lambda_j(t) = \int_0^t \lambda_j(u)du$. The cause specific hazard rate can be computed from the joint survival function of the X's, $S(x_1, \ldots, x_j) = P(X_1 > x_1, \ldots, X_j > x_j)$ as

$$\lambda_j(t) = \left[-\frac{\partial ln S(x_1, \ldots, x_j)}{\partial x_j} \right]_{x_1 = \ldots = x_k = t} \tag{1.5.2}$$

An alternative way to summarize the likelihood of a competing risk model occurring is in terms of the cumulative incidence function (CIF) (also called sub-distribution function) which is given as $G_j(t) = P(T \leq t, \delta = j)$. We note that the cumulative incidence depends on all J of the crude incidence rates. The function $G_j(t)$ is a sub-distribution function with the property that $G_j(\infty) = P(\delta = j)$. The cumulative incidence function represents the chance that competing risk j occurs in a world where individuals can fail from any of the causes. An alternative formulation of the competing risk setup is in terms of a multistage model. It was originally proposed by Prentice *et al.* (1978b) and recently discussed by Andersen *et al.* (2002) and doesn't require the construction of potential failure times for each cause of failure. In the multistage model formulation, there are $J+1$ stages a subject may be in at any point of time. The transient state is the one when the subject is alive and the other J states are the absorbing states when the subject is dead for a given cause. The basic parameters are the transition probabilities, $P_{hj}(s, t)$ which are the probabilities that an individual is in

state j at time t given that he was in state h at time s. Here $P_{00}(0,t)$ is the probability a subject is alive at time t and $P_{0j}(0,t) = G_j(t)$ as defined earlier.

1.6 Moment estimation for censored data

The estimation of the population mean by its sample counterpart is no longer valid in the face of data incompleteness, most commonly being right-censoring. In the spirit of the celebrated paper by Koul *et al.* (1981), very recently Datta (2005) illustrated the estimation of the mean life time in case of right censored data through the concept of reweighting. To explain the concept in a basic survival analysis setup, we have lifetimes T^* and censoring times C for a set of n individuals. The observed data now become $T = T^* \wedge C$ with the censoring indicators $\delta = I(T^* \leq C)$. Define $K_c(t)$=P$(C > t)$ to be the survival function of the censoring variable C. Datta proved the identity

$$E\left\{\frac{\delta T}{K_c(T-)}\right\} = E\{T^*\}$$

which is the theoretical basis for estimating the first moment $E(T^*)$ for censored data by the sample average $\{\sum_{i=1}^n w_i T_i\}/n$ where $w_i = 0$ for a censored point and $= \{\hat{K}_c(T_i-)\}^{-1}$ for an observed life time. Hence, in order to compute the mean in a censored sample, we reweight the observed failure times with inverse of the Kaplan-Meier estimator of the censoring times and then compute the mean. The key behind this concept is also the notion of 'approximate unbiasedness' described in Section 1.4. In Chapter 3, using the same idea of reweighting, we introduce a right-censored version for a general kernel U-statistic, which is a cornerstone in nonparametric statistical literature. This helps in asymptotically unbiased estimation of some mean functional (as above) for censored data.

1.7 Application of nonparametric methods for event time data

In this dissertation, we primarily consider the problem of testing the independence of time to failure and cause of failure in a competing risk model. Competing risk data often come with time to failure and cause of failures for a set of n individuals. Additionally there are right censored observations present in the data which complicate the modeling. Dewan *et al.*(2004) study several dependence structures between time to failure and cause of failure through conditional probabilities. They also explain in detail the necessity of such testing procedures in reliability, clinical trials, epidemiological follow-up studies, etc. In Chapters 2 and 3, we describe novel nonparametric methods which provide a unified treatment to this problem of competing risk. We also consider application of our testing methodology to two real life data sets, viz. (a) Cell carcinoma data and (b) Stanford heart transplantation data. Both of these datasets have two competing causes of failure along with right censored observations. More details on these data sets appear in Section 2.6 of Chapter 2. Our methodology will be particularly helpful for medical practitioners to assess the comparative severity of the concerned set of diseases/cause of failure. In future, we plan to apply our testing tools to more complicated data sets such as current status data and microarray data with survival endpoints. More elaborate description of these types of data appears in Chapter 4 of this dissertation.

1.8 Overview of the dissertation

This rest of this dissertation is organized as follows.

Chapter 2 introduces a novel nonparametric approach for testing the equality of two or more survival distributions based on right censored failure times with missing membership information. The standard log-rank test is not applicable here because the population membership information is not available for right censored individuals. The performance of

this test is validated through several simulation schemes and its functionality is assessed by applying it to two real data sets.

Chapter 3 introduces a censored version of U-statistics (Hoeffding, 1948) with a general kernel of size m obtained by a reweighting principle popularized by Robins (1993). Using the Kendall's τ kernel, we introduce a test which is applied to testing independence of time to failure and cause of failure in a competing risk setup. As in Chapter 2, we study the performance of this test statistic through simulation studies using different kernels and by applying it to a real data set.

Chapter 4 provides the concluding note to this dissertation and summarizes the work done so far. It also states a number of open problems which either I am currently working on or plan to consider in the near future.

1.9 References

[1] Aalen, O.O. (1976). Nonparametric inference in connection with multiple decrement models, *Scandinavian Journal of Statistics*, **3**, 15-27.

[2] Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for nonhomogenous Markov chains based on censored observations, *Scandinavian Journal of Statistics*, **5**, 141-150.

[3] Andersen, P.K., Abildstrom, S.Z., Rosthøj, S.(2002). Competing risks as a multi-state model, *Statistical Methods in Medical Research*, **11**,203-215.

[4] Cox, D.R. (1972). The statistical analysis of dependencies in point processes, In: *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P.A.W. Lewis, Ed. Wiley, NY.

[5] Datta, S. (2005). Estimating the mean life time using right censored data, *Statistical Methodology*, **2**, 65-69.

[6] Datta, S., Satten, G.A. and Datta, S. (2000a). Estimation of stage occupation probabilities in multistage models, In *Advances on Theoretical and Methodological Aspects of Probability and Statistics*, N.Balakrishnan, Ed, Gordon and Breach, NY, 493-506.

[7] Datta, S., Satten, G.A. and Datta, S. (2000b). Nonparametric estimation for the three stage irreversible illness-death model, *Biometrics*, **56**, 841-847.

[8] Datta, S. and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-markov models, *Statistics and Probability Letters*, **55**, 403-411.

[9] Dewan, I., Deshpande, J. V. and Kulathinal, S. B. (2004). On testing dependence between time to failure and cause of failure via conditional probabilities, *Scandinavian Journal of Statistics*, **31**, 79-91.

[10] Fleming, T.R. (1978a). Nonparametric estimation for nonhomogenous Markov processes in the problem of competing risk estimation, *Annals of Statistics*, **6**, 1057-1070.

[11] Fleming, T.R. (1978b). Asymptotic distribution results in competing risks estimation, *Annals of Statistics*, **6**, 1071-1079.

[12] Gail, M. (1975). A review and critique of some models in competing risk analysis, *Biometrics*, **31**, 209-222.

[13] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Statistics*, **19**, 293-325.

[14] Hoel, D. G. (1972). A representation of mortality data by competing risks, *Biometrics*, **28**, 475- 488.

[15] Klein, J.P. and Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, NY

[16] Koul, H., Susarla,V., Van Ryzin, J. (1981). Regression analysis of randomly right-censored data, *Annals of Statistics*, **9**, 1276-1288.

[17] Lagakos, S.W., Sommer, C.J. and Zelen, M. (1978). Semi-Markov models for partially censored data, *Biometrika*, **65**, 311-318.

[18] Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T. and Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics*, **34**, 541-554.

[19] Robins, J.M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, *AIDS Epidemiology- Methodological issues*, In: Jewell, N., Dietz, K., Farewell, V. (Eds.). Birkhäuser, Boston, 297-331.

[20] Robins, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, In: *Proceedings of the American Statistical Association - Biopharmaceutical Section*, 24-33.

[21] Robins, J.M. and Finkelstein, D.M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests, *Biometrics*, **56**, 779-788.

[22] Satten, G.A. and Datta, S. (2002). Marginal estimation of multistage models: waiting time distributions and competing risks analysis, *Statistics in Medicine*, **21**, 3-19.

[23] Satten, G.A., Datta, S. and Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates, *Statistics and Probability Letters*, **54**, 397-403.

[24] Seal, H.L. (1977). Studies in the history of probability and statistics. XXXV multiple decrements or competing risks, *Biometrika*, **64**, 429-439.

[25] Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences, USA*, **72**, 20-22.

Chapter 2

Testing Equality of Survival Distributions when the Population Membership Information is Censored[†]

[†]Bandyopadhyay, D. and Datta, S. Submitted to *International Journal of Biostatistics,* 03/23/06.

2.1 Abstract

This paper introduces a novel nonparametric approach for testing the equality of two or more survival distributions when the population membership information is not available for the right censored individuals. Although such data structures arise in practice very often, this problem has received less than satisfactory treatment in the nonparametric testing literature. Currently there is no nonparametric test for this hypothesis in its full generality in the presence of right censored data. We propose to use the imputed population membership for the censored observations leading to fractional weights that can be used with a two sample censored data test. We study a class of weighted log-rank tests obtained this way through simulation. We obtain an asymptotic linear representation of our test statistic and propose two resampling alternatives which might be easier to use in practice. We illustrate our testing methodology using two real data sets.

Key Words: Competing risk model, Equality of survival curves, Fractional risk set, Log-rank tests, Multistate models.

2.2 Introduction

Testing equality of two or more survival curves is a well studied problem in statistics. The resulting methodologies, e.g., log-rank tests and Cox's regression (in case there are covariates) constitute some of the most used statistical techniques for survival or, more generally, time to event data. The setup assumes that one has independent samples from two or more groups whose survival functions are to be compared. Of course, most (if not all!) of the 'time to event' data are subject to some form of censoring, right censoring being the most common. A unified theoretical and methodological treatment of these problems is achieved through counting process and related martingale techniques.

One basic assumption in using the log-rank type test is that every individual including those whose failure times are right censored is clearly identified to have come from one of two or more independent populations. Now consider a situation where the population memberships are also unknown, along with their failure times, for the right censored individuals. For example, consider a setup commonly referred to as a 'competing risk or multiple decrement model' where individuals are subjected to more than one type of 'risk factor' leading to more than one possible type of failures. If we are interested in studying the sub-populations corresponding to different failure types, the membership information can only be found after the individual actually fails. As a result, these important population membership indicators will be unavailable for individuals who were right censored during an observational study. Certainly, the standard multi-sample testing methods mentioned earlier won't be applicable without this information because it won't be clear which 'at risk' set a censored person would contribute to.

For a better explanation of the testing problem, let us consider the following real-life examples. In a study analyzed by Lagakos (1978), the results of a lung cancer clinical trial being conducted by the Eastern Cooperative Oncology Group were summarized. Patients suffering from squamous cell carcinoma died either out of 'local spread' of disease or 'metastatic spread' of disease; for other patients cause of death was unknown (due to right censoring). A pertinent question to address would be whether the survival curves for the two groups subjected to 'different spreads' of disease are the same or not. In another study, popularly known as the Stanford Heart Transplantation Program, which began in October 1967 and was analyzed by Crowley and Hu (1977) and many others, patients were admitted to the Stanford program for heart replacement. The transplant recipients were subjected to mainly two competing causes of failure, (here death), viz., transplant rejection, or other causes. However, this data set also contains right-censored observations for whom the (eventual) cause of failure was not available. One might be interested in comparing the survival behavior of

those patients who died of 'transplant rejection' to those who died of 'other causes' since the heart transplant. We revisit these data sets and answer these questions in Section 2.6.

The competing risk model is just one formulation where this problem arises. Of course, the testing question in that setup translates to testing the independence of failure time and the cause of failure. More generally, this question may arise in a variety of marginal (over covariates) and marginal-conditional analyses (marginal over some covariates and conditional over others) of event time data. For example, one may be interested in knowing whether individuals who suffer at least one heart attack during their life time will have a shorter life span than individuals who do not. Once again we can consider dividing all individuals into two populations: (i) those who will suffer at least one heart attack during the course of their lives and (ii) individuals who will not experience a single heart attack ever during their life span. Certainly this information will not be available for an individual who, till the time of follow-up, has not had a heart attack but was still living. For another example of a situation where a group or population membership is not known in the beginning of a study, consider a toxicologic experiment on animals where one is interested in studying the effect of a certain carcinogenic agent which can also produce other adverse effects killing an animal before it actually produces cancer. Amongst other things, the survival distribution of animals dying from cancer caused by this agent was of interest. Animals are autopsied at death to study the tumor growth and type through which their population membership was determined where the subpopulation of interest is the animal group where the toxicologic agent would cause a cancer death. Suppose, for some reason, the necessary autopsy could not be performed on certain animals although their death times were known. Our approach would apply to this situation where we would assign a fractional mass to each such individual towards the at risk set of cancer death. Next suppose, one is comparing the survival times of animals dying from cancer caused by two such agents applied to two sets of animals. Suppose, once again, for certain animals in each set, we fail to determine whether they died of cancer or some other

adverse effect. Our testing methodology would handle such individuals by assigning them fractional masses. Another approach would be to simply remove the censored observations and compare the survival times of the animals that are known to have died of cancer. However, as shown by a simple example in Section 2.7, removing censored observations may result in a loss of power.

The rest of the article is organized as follows. Section 2.3 provides literature review of existing tests somewhat related to our testing problem, where we also explain what's different about our testing problem. In Section 2.4, we develop our testing methodology. Section 2.5 deals with a number of simulation studies to check the finite sample performance of our test statistic under different alternatives. Section 2.6 illustrates applications of our proposed test to two real life data sets mentioned above. The paper ends with a discussion section (Section 2.7) followed by an appendix containing an outline of an asymptotic linear representation of our test statistic.

2.3 Literature Review

Are the conditional distributions of failure times of individuals failing due to causes 1 and 2 the same ? Even though this question is very natural and arises frequently, it has received very little attention in the nonparametric testing literature based on right censored data. A somewhat related problem, however, has received much attention in the competing risk framework, namely, testing the equality of cause specific hazard rates (CSHR). Approaches to this problem are described in the next paragraph. Another related problem has also received attention, namely testing the equality of the two (or more) sub-distribution functions or cumulative incidence functions (CIF). Sub-distribution functions are defined as $G_j(t) = P(T \leq t, \delta = j)$, where T and δ are the time to failure and cause of failure respectively. It is easy to see that equality of cause-specific hazard rates or the equality of cumulative

incidence functions would imply that same percentage of failures would take place due to different causes. However, there are situations where this is not expected; even then, one may be interested in comparing the conditional distributions of the failure times due to various causes. This is the testing problem that we are considering in this paper and using the existing tests for equality of CSHR's or CIF's may lead to wrong size under our null hypothesis. The existing tests for each of these testing problems are reviewed next.

Tests for equality of cause specific hazards are given in Bagai *et al.* (1989a, 1989b), but they do not handle any censoring. Aly *et al.* (1994) and Sun and Tiwari (1995) proposed tests for testing the equality of two CSHR's in the presence of censoring. Generally speaking, in order to compare the cause-specific hazards between groups, all other risks are equated with censoring. A literature review also reveals at least three tests to compare the cumulative incidence functions directly. The first test is due to Gray (1988) which is based on comparing the (weighted) differences between the estimates of hazard rates of an improper random variable and its pooled sample estimate. Choices of weight functions are also discussed. The second test is due to Lin (1997) and is a variant of the Kolmogorov-Smirnov test with a weight function. The third test is due to Pepe (1991) and is based on the integrated difference between the weighted cumulative incidence functions. Lam (1998) and Carriere and Kochar (2000) also provide some tests for equality of CSHR's or CIF's. Fine and Gray (1999) proposed regression modeling of a hazard corresponding to a sub-distribution function. The resulting methodology can provide tests of equality of CSHR's or CIF's.

The problem of comparing CSHR's between groups is a somewhat different problem in which the group memberships are assumed to be known. Lindkvist and Belyaev (1998) generalized the log-rank statistic to provide a class of tests for comparing cause-specific hazard rates from two competing causes of failure between two groups and Kulathinal and Gasbarra (2002) extended these results to $M(\geq 2)$ groups.

There are only a limited number of papers dealing with the testing problem we are interested in. The papers by Dykstra *et al.* (1998) and Kochar and Proschan (1991) provide some restricted tests for testing the independence of time to failure and cause of failure in a competing risk framework. Testing of dependence structures between failure time and cause of failure expressed in terms of monotonicity properties of the conditional probabilities involving failure times and failure cause are proposed very recently by Dewan *et al.*(2004). They used a U-statistic approach which assumes no censored observations.

2.4 Tests Using Fractional Risk Sets

2.4.1 Notation

We consider the competing risk network as a multistate continuous time stochastic process $\{Z(t), t \in \mathcal{T}\}$ with a finite state space $\mathcal{S} = \{1, \ldots, J, 0\}$ having a tree topology and right-continuous sample paths: $Z(t+) = Z(t)$ where we assume that the states $1, \ldots, J$ are absorbing whereas state 0 is transient (the root node). Here $\mathcal{T} = [0, \mathcal{T}]$ where $\mathcal{T}$ is a large possibly observed time point ($\leq \infty$). Typically, for applications, $\mathcal{T}$ will be taken to be the largest time where some event (failure) took place. Let T_i^* be the (possibly unobserved) time the ith person leaves stage 0 for a failure (stage j, say). Let X_i^* denote the stage occupied by the ith individual at time T_i^* (i.e., its failure type). Let C_i be the censoring time for the ith person. Let $T_i = T_i^* \wedge C_i$ denote the right censored failure time and δ_i denote the failure/censoring indicator

$$\delta_i = \begin{cases} j & \text{if} \quad T_i^* \leq C_i \text{ and } X_i^* = j, \\ 0 & \text{if} \quad T_i^* > C_i. \end{cases} \tag{2.4.1}$$

Note that δ_i is an observable quantity for everybody but X_i^* is observed only for the uncensored people and then the two are equal. It is further assumed that the censoring variable C_i is independent of the entire collection of $\{T_i^*, X_i^*\}$ and all the random variables are i.i.d. across

the n individuals. Our data consist of the pairs (T_i, δ_i), $1 \leq i \leq n$. Define the conditional survival function $S_j(t)$ as

$$S_j(t) = P(T_i^* > t | X_i^* = j), \quad 1 \leq j \leq J.$$

We are primarily concerned with the null hypothesis,

$$H_0 : S_1(t) = \cdots = S_J(t) (\equiv S(t), \text{ say}), \forall\ 0 \leq t \leq L,$$

where $L \leq \mathcal{T}$ is a finite but large time point.

2.4.2 Test Statistic

A standard log-rank test would have been applicable if the individuals could be separated into various groups depending on their failure types. Basically, the Nelson-Aalen estimators of the (conditional) cumulative hazards for various groups could be compared amongst each other. However, as indicated earlier, the difficulty now is that the subpopulation memberships X_i^*'s are not known for the censored people and simply ignoring (or deleting) them from the 'at risk' consideration will lead to a biased comparison which would lead to improper size for the resulting tests. This is so because if the censored observations contribute full mass to all 'at risk' sets then the resulting comparison is equivalent to comparing the CIF's. As explained earlier in Section 2.3, this would correspond to the wrong null hypothesis whenever $P(X_i^* = j)$ are not constant in j. Nevertheless, the Nelson-Aalen representation for various subpopulations will be the starting point of our proposed method. To this end, we introduce the following standard notations. Let N_j be the counting process counting the number of observed failures of type j (i.e., number of transitions into stage j) in the time interval $[0, t]$:

$$N_j(t) = \sum_{i=1}^{n} I(T_i \leq t,\ \delta_i > 0, \quad X_i^* = j),$$

and let $Y_j(t)$ denote the number of individuals at risk of failing due to cause j or of getting censored:

$$Y_j(t) = \sum_{i=1}^{n} I(T_i \geq t,\ X_i^* = j).$$

It is not difficult to see that N_j is computable from the observed data, since $\{\delta_i > 0, X_i^* = j\}$ is the same as $\{\delta_i = j\}$. Y_j, on the other hand, is not computable, once again, due to the unavailability of all the group membership data. Nevertheless, it will be helpful to recall the definition of the Nelson-Aalen estimator (Andersen *et al.*, 1993) of cumulative hazards of failure amongst individuals of subpopulation (or failure type) j:

$$\hat{\Lambda}_j(t) = \int_0^t \frac{I(Y_j(s) > 0)}{Y_j(s)} dN_j(s). \tag{2.4.2}$$

Note that $\hat{\Lambda}_j(t)$ estimates the correct cumulative hazard corresponding to the conditional distribution $S_j(t)$, and not the so called cause specific cumulative hazard. When some individuals are censored, we cannot classify their failure types. In the absence of such an identifier, however, we may still assign a probability of each individual being in one of the J subpopulations (something like an imputed subpopulation identifier). Once these probabilities are known, we proceed with the supposition that the data be divided into J subpopulations, the risk set of each subpopulation now contains fractional observations with the fractional mass specified by an estimate of the probability that the observation belongs to a particular subpopulation. Thus, we estimate $Y_j(t)$ by $Y_j^f(t)$, where $Y_j^f(t)$ denotes the 'fractional risk set' corresponding to the jth cause of failure defined as (Satten and Datta, 1999; Datta *et al.*, 2000)

$$Y_j^f(t) = \sum_{i=1}^{n} \hat{\phi}_{ij} I(T_i \geq t), \tag{2.4.3}$$

where $\hat{\phi}_{ij}$ is the estimated probability that the ith individual belongs to the jth subpopulation according to its failure type. It is not hard to see that a reasonable choice for $\hat{\phi}_{ij}$ is given by

$$\hat{\phi}_{ij} = \begin{cases} \hat{P}_j(T_i, \infty), & \text{if } \delta_i = 0 \\ 1, & \text{if } \delta_i = j \\ 0, & \text{if } \delta_i > 0, \quad \delta_i \neq j, \end{cases} \tag{2.4.4}$$

where $\hat{P}_j(T_i, \infty)$ is the nonparametric maximum likelihood estimator (NPMLE) or the Aalen-Johansen estimator of the transition probability

$$P_j(s,t) = Pr\{T_i^* \leq t, X_i^* = j | T_i^* > s) \text{ at } s = T_i,\ t = \infty$$

See Datta and Satten (2000) for the legitimacy of the above estimator (2.4.4). Specializing Andersen *et al.* (1993) results for a Markov chain to a competing risk setup, we obtain

$$\hat{P}_j(s,t) = \int_{(s,t]} \left\{ \prod_{(s,u)} \left(1 - \frac{dN(v)}{Y(v)} \right) \right\} \frac{dN_j(u)}{Y(u)} \tag{2.4.5}$$

In the expression (2.4.5), $Y(t) = \sum_{i=1}^{n} I(T_i \geq t)$ is the size of the 'at risk' set irrespective of failure types, $N(t) = \sum_{j=1}^{J} N_j(t)$ is the total number of observed failures of all types (total number of stages entered) by time t. Since N_j and N are discrete with jumps only at the failure times T_i's, the above integral can be replaced by sums leading to the following simpler expression

$$\hat{P}_j(s,t) = \sum_{s < T_i \leq t} \left\{ \frac{\hat{S}(T_i-)}{\hat{S}(s)} \right\} \left\{ \frac{\Delta N_j(T_i)}{Y(T_i)} \right\}$$

where $\hat{S}$ is the Kaplan-Meier estimator of failure due to all causes and $\Delta N_j(T_i)$ is the number of failures of type j at time T_i. Note that for a typical individual who had not failed up to and including time s, the first term $\hat{S}(T_i-)/\hat{S}(s)$ in the above summand computes the probability that such an individual had not failed until time T_i and the second term $\Delta N_j(T_i)/Y_j(T_i)$ is the probability of its failing at time T_i and the failure is of type j given survival until that time.

A natural class of nonparametric test process $\{\Delta_j(t), t \geq 0\}$ that are applicable to such a setup is the class of weighted log-rank tests based on the weighted departures of the estimated Λ_j compared to a pooled estimate of the common Λ under the null hypothesis given by

$$\Delta_j(L) = \int_0^L W(t) \left\{ dN_j(t) - \frac{Y_j^f(t)}{Y^f(t)} dN(t) \right\} \tag{2.4.6}$$

where $W(t)$ is a non-negative, locally bounded, predictable weight process that provides a flexible way to control the relative importance attached to differences in the estimated conditional hazards at various time points, $Y_j^f(t)$ is as defined in (2.4.3) and $Y^f(t) = \sum_{j=1}^{J} Y_j^f(t)$ is the total fractional risk set at time t. The components of the vector $\boldsymbol{\Delta} = (\Delta_1(L), \ldots, \Delta_J(L))$ are linearly dependent since $\sum_{j=1}^{J} \Delta_j(L) = 0$. Let $\hat{\Sigma}$ be an estimated variance covariance matrix of the vector test statistic $\boldsymbol{\Delta}$ and let $\hat{\Sigma}^-$ be its spectral (generalized) inverse. Our test statistic is given by the quadratic form

$$T = \boldsymbol{\Delta}^T \hat{\Sigma}^- \boldsymbol{\Delta} \tag{2.4.7}$$

which has an asymptotic chi-squared distribution with $J - 1$ degrees of freedom under the null hypothesis provided suitable regularity conditions are met.

2.4.3 Asymptotic Normality of $\boldsymbol{\Delta}$ and the Construction of $\hat{\Sigma}$

The most commonly used tool for obtaining asymptotic distributions of estimators and test statistics with right censored data is the counting process formulation and the associated martingale techniques pioneered by Odd Aalen (Aalen, 1978) and used by numerous authors since then. Certainly they work for the regular log-rank tests when subpopulation memberships are known. However, in the present context, the imputed subpopulation memberships are based on the entire data set and in a sense the fractional at risk set at time t are functions of future information as well (i.e., they are not predictable with respect to the natural filtration process one considers). However we can still obtain asymptotic normality of the test statistic by establishing the following asymptotic linear representation:

$$\Delta_j(L) = \sum_{i=1}^{n} V_i + o_P(\sqrt{n}), \tag{2.4.8}$$

where

$$V_i = \int_0^L w(t) dM_{j,i}(t) - \int_0^L w(t) \frac{y_j(t)}{y(t)} dM_{.,i}(t) +$$
$$\int_0^L \frac{w(t)}{y(t)} \{ I(T_i \geq t,\ X_i = j,\ \delta_i = 0) - P_j(T_i, \infty) I(T_i \geq t,\ \delta_i = 0)$$
$$+ \int_t^\infty \int_z^\infty \frac{S(u)}{S(z)y(u)} P_j(u, \infty) dM_{.,i}(u) dn^C(z) \Big\} dn(t)$$
$$- \int_0^L \frac{w(t) y_j(t)}{y^2(t)} \sum_{j'=1}^{J} \{ I(T_i \geq t, X_i = j',\ \delta_i = 0) - P_{j'}(T_i, \infty) I(T_i \geq t, \delta_i = 0)$$
$$+ \int_t^\infty \int_z^\infty \frac{S(u)}{S(z)y(u)} P_{j'}(u, \infty) dM_{.,i}(u) dn^C(z) \Big\} dn(t),$$

with

$$M_{j,i} = I(T_i \leq t, \delta_i > 0, X_i = j) - \int_0^t I(T_i \geq s, X_i = j) d\Lambda(s), M_{.,i} = \sum_j M_{j,i},$$

$$y_j(t) = P(T_i \geq t, X_i = j); y(t) = \sum_j y_j(t),$$

$$n(t) = P(T_i \leq t, \delta_i > 0), n^C(t) = P(T_i \leq t, \delta_i = 0),$$

and w is the (in-probability) limit of the weight function W. We present an outline of the derivation in the appendix. This representation also yields a closed form expression for $\hat{\Sigma}$ obtained by the empirical variance covariance of $\hat{V}_i$, where $\hat{V}_i$ is obtained by replacing all the population quantities in V_i by their estimates. However, a close examination of the formulas reveals that the resulting expression is computationally cumbersome and requires $O(n^3)$ order of calculations. A more practical solution will be to use a resampling (bootstrap) scheme to compute the estimated variance-covariance matrix $\hat{\Sigma}$. Asymptotic validity (consistency) of this bootstrap based estimate can be established deriving a similar linear representation (2.4.8) for the resampled statisitc. The theoretical details are not pursued here.

2.4.4 Bootstrap Based Tests

For testing problems, a resample scheme has to incorporate the null hypothesis. This is achieved as follows:

(i) Let $\delta_i' = I(\delta_i > 0)$ denote the true failure indicators irrespective of the failure types, $1 \le i \le n$. Generate i.i.d. resamples $(\hat{T}_i^*, \hat{\delta}_i^{*\prime})$, $1 \le i \le n$, from the empirical distribution of the pairs $\{(T_i, \delta_i') : \ 1 \le i \le n\}$.

(ii) We let $\hat{\delta}_i^* = 0$, if $\hat{\delta}_i^{*\prime} = 0$; otherwise, for $\hat{\delta}_i^{*\prime} > 0$, generate $\hat{\delta}_i^*$ from a J-point distribution with $\text{Prob}\{\hat{\delta}_i^* = j\} = \hat{\Phi}_j / \sum_{k=1}^{J} \hat{\Phi}_k$, with

$$\hat{\Phi}_j = \sum_{i=1}^{n} \hat{\phi}_{ij}, \quad 1 \le j \le J.$$

A typical bootstrap sample is given by $(\hat{T}_i^*, \hat{\delta}_i^*)$, $i = 1, \ldots, n$.

(iii) Repeat steps (i) and (ii), a large number of times, say B, to obtain B sets of bootstrap samples $(\hat{T}_{i,b}^*, \hat{\delta}_{i,b}^*)$ and compute the corresponding test statistics $\hat{\boldsymbol{\Delta}}_b^*$, $b = 1, \ldots, B$.

A bootstrap estimate of $\hat{\Sigma}$ is given by the empirical variance-covariance matrix of the $\hat{\boldsymbol{\Delta}}_b^*$, $1 \le b \le B$. Alternatively, we could use bootstrap to compute the p-value of a supremum test that avoids computation of the variance-covariance matrix as

$$\hat{p} = B^{-1} \sum_{b=1}^{B} I\{\max_{1 \le j \le J} |\hat{\Delta}_{j,b}^*| \ge \max_{1 \le j \le J} |\Delta_j|\} \tag{2.4.9}$$

and rejecting the null hypothesis for small values of $\hat{p}$. The test based on (2.4.9) will be the same as that based on (2.4.7) when J=2 since in this case $\Delta_1 = -\Delta_2$ and $\hat{\Delta}_{1,b}^* = -\hat{\Delta}_{2,b}^*$. In the next section, we perform a number of simulation studies to assess the performance of both these tests under a number of alternative hypotheses models.

2.4.5 The Alternative Hypothesis

Log-rank tests for comparing survival curves are popular in the literature because of their easy interpretability and availability in various 'survival analysis' computer packages. Differences in hazards (or the survival curves) leading to alternative hypotheses could occur in a variety of ways, sometimes in early period of follow-up, or late or in the middle. It has been established that the standard log-rank tests can have very low power with respect to some alternative hypotheses. In addition, the choice of an 'optimal' test also depends upon the sample sizes and the censoring patterns. When equal weights are given to all points on the entire curve and proportional hazards assumption holds (under Lehmann alternatives), the log-rank is the most powerful. However, they are insensitive to non-Lehmann alternatives and under heavy censoring (Fleming *et al.*, 1981). Similar properties are expected for our log-rank tests as well. Other customized tests such as a test of trend (Klein and Moeschberger, 2003) can be proposed using the fractional risk set idea as well.

2.4.6 Choice of the Weight Process

In principle, one can use any of the huge arsenal of weights that are recommended for a weighted log-rank test. The choice of $W(t_i) = 1$ reduces the statistic similar to the popular version of log-rank test. Several other choices can be considered such as that proposed by Fleming and Harrington (1981), Harrington and Fleming (1982) popularly known as the G^{ρ} family which assigns more weights to early and late differences between the hazard rates in the J populations. One can also give an extension analogous to the $G^{\rho,\gamma}$ family which is more efficient for survival differences in the middle of the study (Fleming and Harrington, 1991). Recent choices of weight functions for testing when events are rare as of Buyske *et al.* (2000), flexible weights for detecting early and/or late survival differences as of Wu and

Gilbert (2000) or incorporating quality adjusted lifetime as of Zhao and Tsiatis (2001) can also be considered with certain modifications to our test statistic.

2.5 Simulation Studies

To assess the performance of our test statistic, we conducted two different simulation schemes to be described next each resulting in a two-cause ($\delta = 1, 2$) competing risk model. We used sample sizes of $n = 50, 200$ and a nominal level $\alpha = 0.05$ was used in each case. Furthermore, for each sample, a bootstrap replication size of $B = 1000$ was used. Finally the power was computed by the proportion of rejection out of 2000 Monte-Carlo replications. The weight function was taken to be unity and the largest observed failure time was used for L in all cases.

2.5.1 Simulation Schemes

(i) Simulating causes prior to event times

Under this scheme, we generate data in such a way that the fraction of people under the two groups remains constant no matter what sample size is chosen. This is done to illustrate the effect of sample size on power while other settings remain comparable. We first simulate the event (failure) indicators δ_i by setting the arm probability ϕ to be 0.25, 0.5 or 0.75 and letting δ be 1 if a randomly generated uniform (0,1) random variable falls below ϕ, and 2 otherwise. Next, lifetimes T_i are generated from a standard log-normal ($= \exp(N(0, 1))$ distribution for both subpopulations under the null hypothesis. For the alternative hypothesis, lifetimes for the subpopulation $\delta = 1$ are generated as in the null case; however, samples for the $\delta = 2$ subpopulations are generated as $\exp(N(-a, 1))$ with $a = 0(0.3)1.5$, where the value in parentheses indicate increments. The censoring times C_i are generated from the lognormal distributions with the mean parameters 0.954 and 0, while keeping the variance parameter

to be 1 in both cases, representing light (about 25%) and heavy (about 50%) censoring, respectively. In all cases, we take the nominal level α to be 5%.

(ii) Simulating event times from a bivariate distribution

Consider a unit being exposed to two risks and let the notional (or latent) lifetimes of the unit under the two risks be denoted by X and Y. In general, X and Y are dependent and being lifetimes, they should be non-negative. In this simulation setup, we only observe (T, δ), where $T = \min(X, Y)$ is the failure time and $\delta = 2 - I(X \leq Y)$ is the cause of failure. For the null hypothesis, we generate (X', Y') from a bivariate normal distribution with mean vector (0,0), and variance-covariance matrix $((1, \rho), (\rho, 1))$, where ρ is chosen to be -0.5, 0 or 0.5. For the alternative, we generate (X', Y') from a bivariate normal distribution with mean vector $(0, (1 - a))$ and variance-covariance matrix $((1, \rho a), (\rho a, a^2))$, where $a = 1(0.3)2.5$. Finally, we let $X = \exp(X')$ and $Y = \exp(Y')$. The censoring times are generated from log-normal distributions with variance parameter 1 and mean parameters 0.954 and 0 leading to light (8% - 17%) and moderately heavy (28% - 45%) censoring, respectively.

2.5.2 Results

First, we report the empirical sizes for the two simulation schemes in Table 2.1 and 2.2. The empirical sizes of our tests are close to the nominal level 0.05 and within a 95% confidence interval for the larger sample size although. The empirical sizes for $n = 50$ are also fairly close to the nominal although they are marginally inflated in a few cases.

Figures 2.1 and 2.2 display arrays of plots illustrating the power curves under the two schemes as a function of the alternative parameter a which was defined earlier and has different interpretation under the two schemes. Once again, the set up is identical to the size study above and includes two choice of sample sizes, namely, 50 and 200. The upper two curves are for sample size 200 and the lower two for sample size 50. In each figure, we overlay

Table 2.1: Size of imputed log-rank tests using $\alpha = 0.05$ under simulation scheme (i)

	Bootstrapped p-value test						Chi-squared test					
	Light censoring			Heavy censoring			Light censoring			Heavy censoring		
	ϕ			ϕ			ϕ			ϕ		
	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
n												
50	0.049	0.061	0.063	0.062	0.059	0.058	0.045	0.054	0.058	0.053	0.049	0.055
200	0.051	0.053	0.049	0.054	0.053	0.047	0.047	0.046	0.047	0.048	0.043	0.045

n = sample size; ϕ = arm-probability.

the power curves obtained by our method with the power curves of a regular log-rank test if the group information were known for all individuals. The solid lines correspond to the standard log-rank tests with the knowledge of the group membership for everyone and the dotted lines correspond to our chi-squared test. The bootstrap p-value approach was seen to yield very similar power profiles as the Chi-squared test and hence we only included the Chi-squared plots in our figures.

It is seen that a considerable increase in power is observed with an increase in sample size. The power curves reveal that for low censoring and larger sample size, our imputed test produces a power profile that is very close to the log-rank test with the additional information about group membership for the censored observation. Heavy censoring has some adverse effect on the empirical powers particularly under simulation setup (i). These tend to get better with the larger sample size.

We have also studied the sensitivity of our results with respect to the selection of the bootstrap replication size B. As mentioned earlier, all the values reported in Tables 2.1 and 2.2 were based on $B = 1000$, which was judged sufficient after balancing accuracy

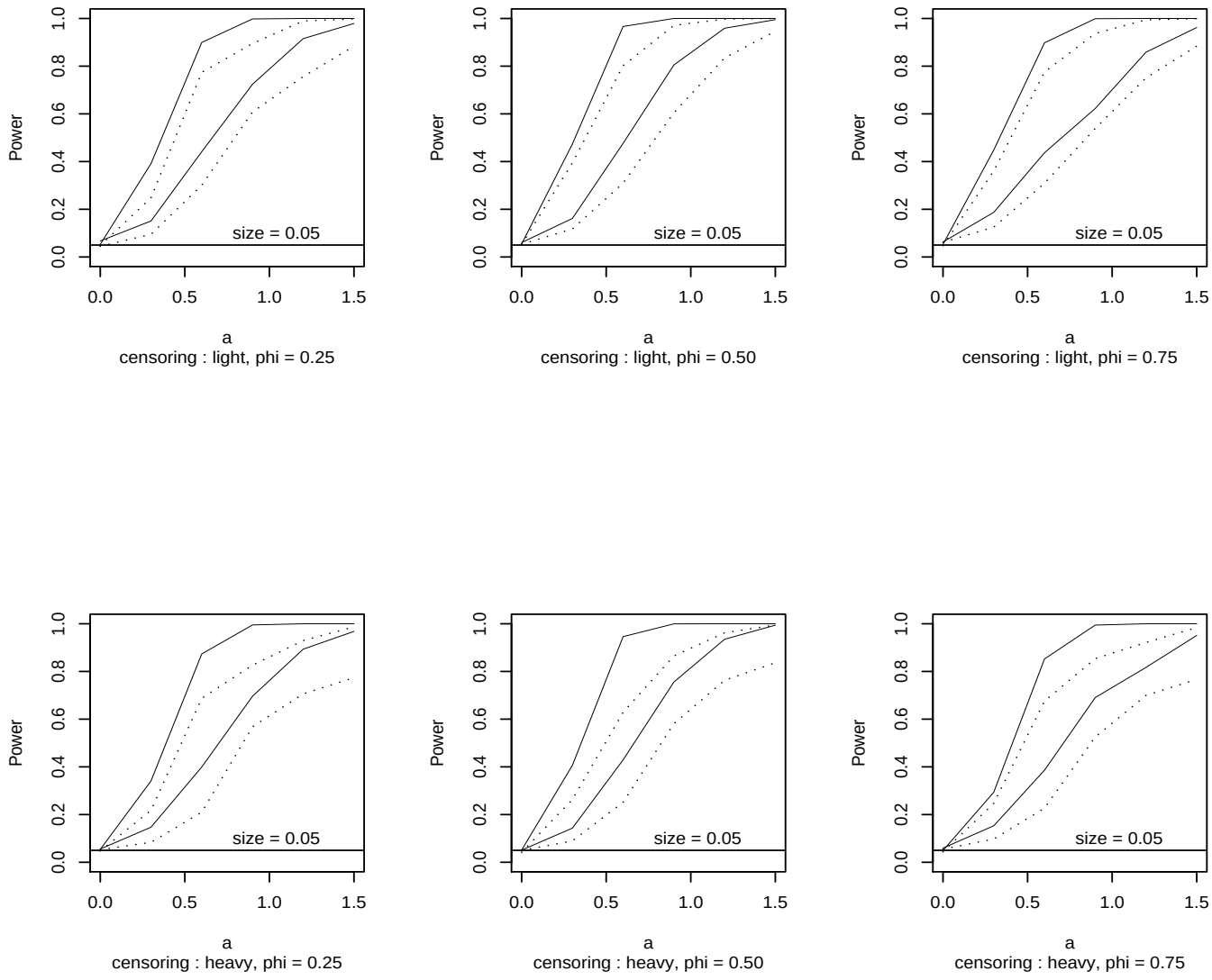

Figure 2.1: Power curves of our tests for Simulation scheme (i). The solid lines correspond to the log rank tests for known population memberships and the dotted lines corresponds to our Chi-squared test.

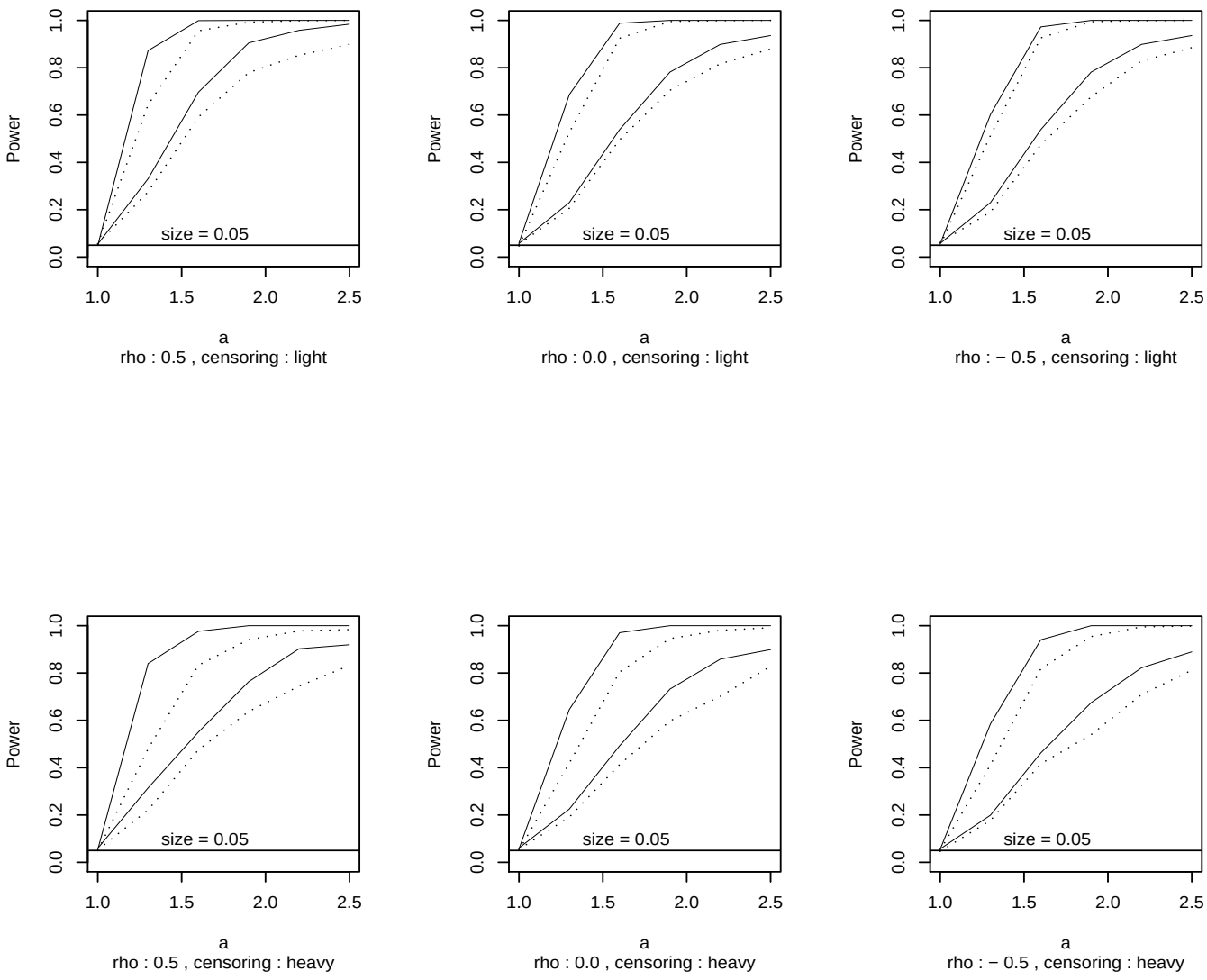

Figure 2.2: Power curves of our tests for Simulation scheme (ii).The solid lines correspond to the log rank tests for known population memberships and the dotted lines corresponds to our Chi-squared test.

Table 2.2: Size of imputed log-rank tests using $\alpha = 0.05$ under simulation scheme (ii)

		Bootstrapped p-valued test		Chi-squared test	
		Censoring			
		Light	Moderately heavy	Light	Moderately heavy
ρ	n				
0.5	50	0.061	0.060	0.056	0.053
	200	0.056	0.052	0.054	0.049
0	50	0.062	0.061	0.054	0.056
	200	0.049	0.054	0.047	0.051
-0.5	50	0.053	0.064	0.047	0.061
	200	0.051	0.057	0.049	0.051

n = sample size; ρ = correlation.

with computational burden. Table 2.3 reports the power values obtained using a variety of bootstrap iterations for the simulation setup (ii) with $\rho = 0.5$ and $n = 200$. We here report the power calculations based on three choices of B, namely, 200, 500 and 1000. In each case, a Monte Carlo replication size of 2000 was used as before. As can be seen from Table 2.3, the values were reasonably stable; in particular, the values corresponding to $B = 500$ and $B = 1000$ differ by no more than 5×10^{-3}.

2.6 Examples

We now illustrate our methods using the Cell Carcinoma data and the Stanford Heart Transplant data mentioned in the introduction.

Table 2.3: Power values for different bootstrap iterations under simulation scheme (ii)

Censoring	B	Size	Power				
		a =1	a = 1.3	a = 1.6	a = 1.9	a = 2.2	a = 2.5
Low	200	0.052	0.630	0.957	0.992	0.998	1.000
	500	0.052	0.643	0.958	0.993	0.997	0.999
	1000	0.056	0.644	0.956	0.992	0.998	1.000
Moderately heavy	200	0.056	0.487	0.858	0.945	0.974	0.991
	500	0.054	0.485	0.828	0.945	0.976	0.987
	1000	0.052	0.480	0.833	0.941	0.978	0.984

2.6.1 Cell Carcinoma Data

The Cell Carcinoma data published in Lagakos (1978) has 194 affected patients with squamous cell carcinoma out of which 83 patients failed with local spread (LS) of disease (Cause 1), 44 failed with metastatic spread (MS) of disease (Cause 2) and 67 have right-censored failure times, i.e. about 35% censoring. The value of τ (the maximum value of event time) was 101 days and the minimum event time was 1 day. The survival times of the LS patients ranged from 1 day to 88 days and that of the MS patients from 2 days to 84 days. There were a number of covariates in the data set but presently we ignore them. We are interested in testing the hypothesis that the lifetime distributions in the groups are the same. We use the bootstrap technique as described before with bootstrap replication size of 5000 to compute the standard error of the test statistic. The value of Δ (computed considering $W(t_i) = 1$ and $L = \tau$) for this data set turned out to be 2.014; the standard error was 6.75 and the χ^2 test statistic equaled 0.09. The bootstrapped p-value was 0.76. Thus, the null hypothesis of equality of survival curves in the two disease categories was not rejected at the 5% level.

Figure 2.3 shows the nonparametrically estimated cumulative hazard rates (2.4.2) for the two types of disease spreads computed using fractional risk sets. The results are consistent with our hypothesis testing findings in a sense that both the estimated conditional cumulative hazards remain very close to each other.

2.6.2 Stanford Heart Transplant Data

We consider the subset of $N = 65$ patients of the Stanford Heart Transplant data published in Crowley and Hu (1977) as considered in Larson and Dinse (1985). Among these 65 patients who received transplants, there were 29 (45%) rejection deaths (RD, coded as Cause 1), 12 (18%) deaths from other causes (OC, coded as Cause 2), and 24 (37%) right censored observations. The value of τ was 1775 days and the minimum event time was 0 day. The survival times of RD patients ranged from 10 days to 1350 days and that for the OC patients from 0 to 551 days. Once again, a resample based on 5000 replications was used to compute the standard error of the test statistic. The value of Δ (computed similarly) for this data set was 7.07 with a bootstrap estimated standard error of 3.58 leading to a $Z(= \Delta/SE(\Delta)$, SE = standard error) test statistic value of 1.98 for the test statistic that was significant at 5%. The bootstrap estimated p-value was 0.0426. We can conclude that the conditional hazards for cause RD is lower than that of cause OC which is also supported by the plot of the cumulative hazards of the two disease types considering fractional risk sets. See Figure 2.4 for the illustration.

2.7 Discussion

A weighted log-rank test for testing the equality of J survival functions has been proposed when the subpopulation memberships of the right censored individuals in the study are unknown. Such a structure arises naturally in practice. The concept of fractional risk was the

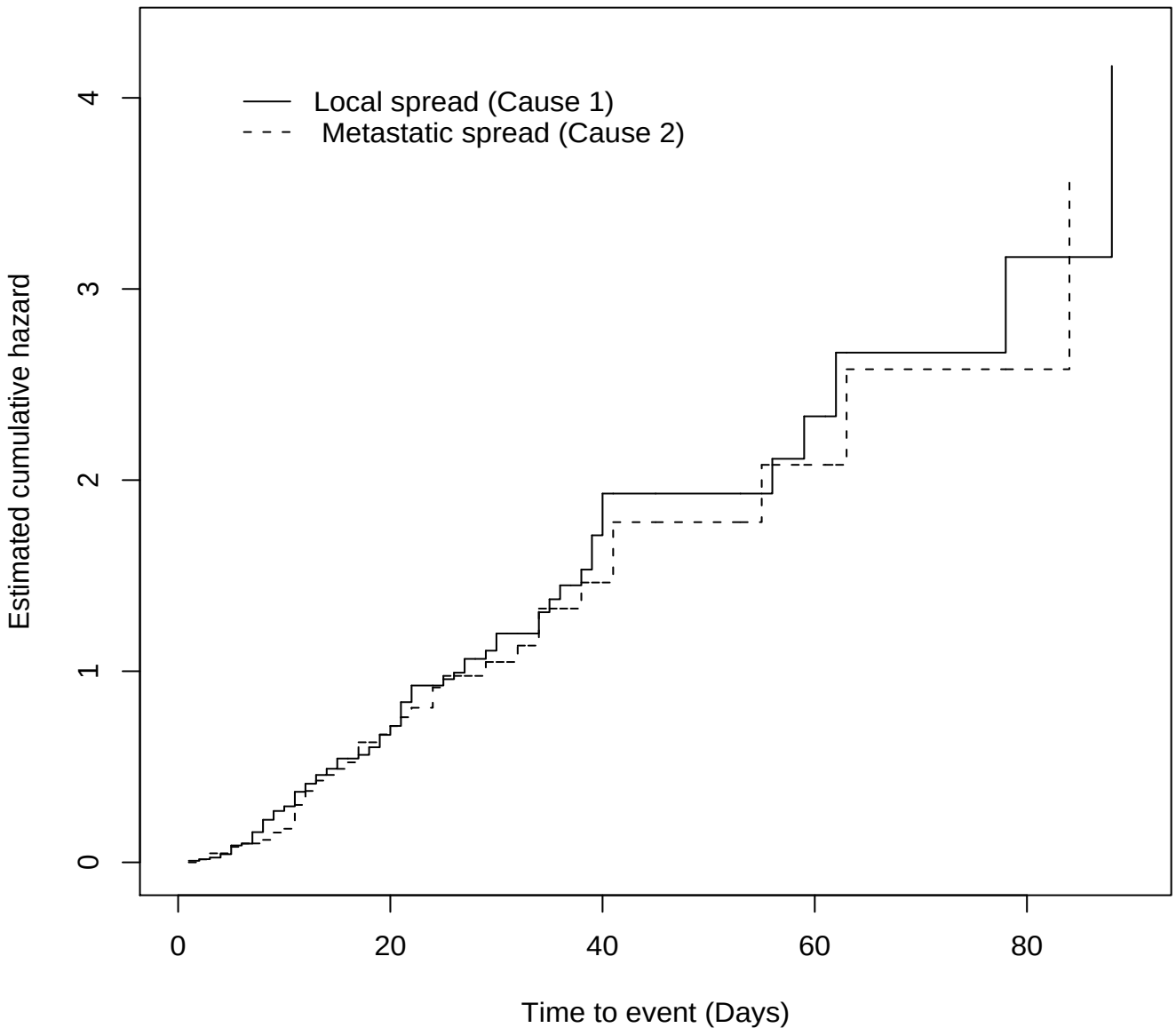

Figure 2.3: Estimated cumulative hazard rates for the two groups in the Cell Carcinoma Data

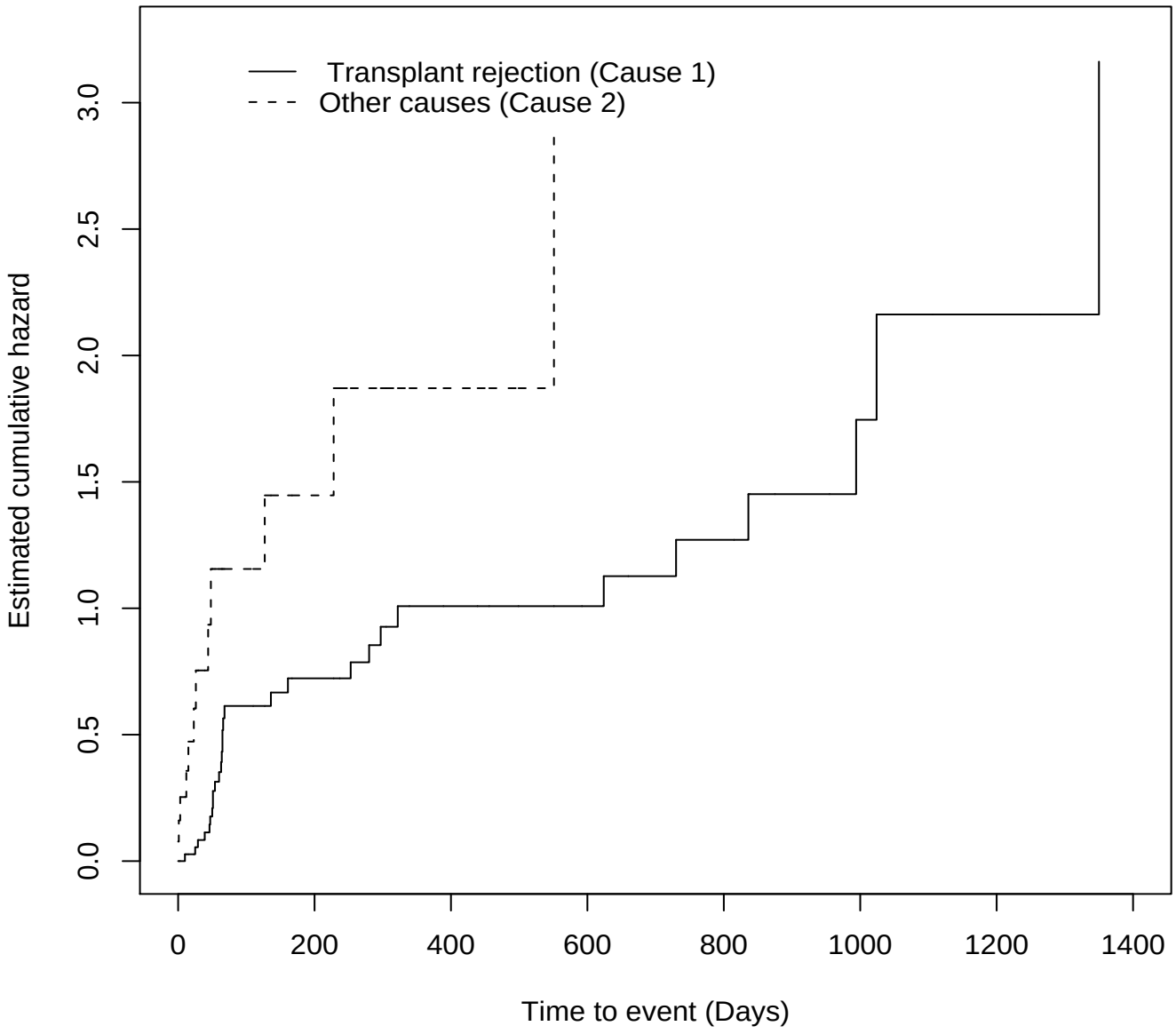

Figure 2.4: Estimated cumulative hazard rates for the two groups in the Stanford Heart Transplant Data

key in breaking up the total risk set into J sub risk-sets (corresponding to J sub populations) at every event time point and it thereby connects this testing problem to the traditional log-rank testing methodology.

Due to the lack of a suitable martingale representation, the form of the estimated variance-covariance matrix of the test statistics is computationally cumbersome and as a result we propose a bootstrap scheme to carry out the test either by computing the bootstrap variance-covariance matrix or by a bootstrap approximation of the p-values directly. While the two procedures are asymptotically equivalent, based on a simulation study, the second approach appears to be one notch better in terms of size and power in small to moderate samples.

As explained earlier in Section 2.2 and 2.3, if one choses to ignore the subpopulation information for censored data and let them contribute full mass (whole observations) to all 'at risk' sets then the resulting test will have wrong size for our null hypothesis. If on the other hand, the censored observations are thrown away (equivalent to contributing zero mass to each 'at risk' set), the size would be maintained but there may be a substantial loss in power. This is illustrated by a simple simulation scheme as follows. Suppose we generate equal number of observations from two populations. Failure times in population 1 are generated from a Uniform(0.9, 1.2) distribution and that for population 2 from Uniform(0.2+i, 3.1) distribution, $i = 0(0.1)0.3$. The four values of i create four alternatives for the testing problem that we are interested in. Censoring times are independently generated from a Uniform(0.9, 3.0) distribution. We choose a sample size of $n = 50$, a targeted nominal level of $\alpha = 0.5$ and a bootstrap size of $B = 1000$ and an iteration size of $N = 1000$ to compute powers empirically. Table 2.4 reports the power of our log-rank test based on fractional masses for censored observations and that of a log-rank test where censored observations are removed from the sample. We see up to nearly 30% power gain by imputing population membership rather than removing the censored observations.

Table 2.4: Power of imputed log-rank test and that of a standard log-rank after throwing away the censored observations, each at 5% level

Population 1	Unif(0.9, 1.2)			
Population 2	Unif(0.2,3.1)	Unif(0.3, 3.1)	Unif(0.4, 3.1)	Unif(0.5,3.1)
Log-rank test	0.581	0.698	0.780	0.869
Log-rank (no censored)	0.449	0.551	0.645	0.781

Although in this paper we restrict our attention to the class of log-rank tests, the concept of fractional risk set goes much further. In essence, it will enable us to propose versions of other tests for testing equality of survival curves based on independent right-censored samples that are based on risk sets, where the group membership is known for everyone, to the present situation. Examples of such tests would include tests based on weighted difference of Kaplan-Meier survival estimates of conditional survival functions for the J groups. The conditional survival functions for the J groups can be obtained by normalizing the Aalen-Johansen estimators or equivalently by Kaplan-Meier formulas using the fractional risk sets. See Satten and Datta (1999) for this equivalence.

Our methodology can be applied to survival data in which the curves cross or differ from each other in more general ways with simple modifications. The log-rank type of tests are based upon weighted integrals of estimated differences between survival curves. Hence for crossing survival curves, the positive differences will be negated by negative differences leading to substantial loss of power. In such situations, we can modify our test statistic

(2.4.6) as

$$\Delta_j(L) = \sum_{t_k \le L} W(t_k)|U_{k,j}|$$

or

$$\Delta_j(L) = \sum_{t_k \le L} W(t_k)U_{k,j}^2$$

where

$$U_{k,j} = \Delta N_j(t_k) - \frac{Y_j^f(t_k)}{Y^f(t_k)}\Delta N(t_k).$$

Here $\Delta N_j(.)$ and $\Delta N(.)$ denote jumps in the respective processes. One can now resort to the resampling technique introduced before to compute the null mean and variance of these test statistics and turn them into asymptotically standard normal (or chi-squared). Alternatively, one can compute the p-values directly via resampling as before without studentizing. Performance of such tests statistics will be studied elsewhere.

The present methodology does not consider any covariates. It might be of interest in some applications to ask whether the conditional survival functions in various groups are equal given some available covariates. This will be pursued elsewhere.

2.8 APPENDIX

2.8.1 Outline of linear representation

Using two triangulations and the fact that the conditional hazard rates are the same under the null hypothesis we obtain from (2.4.6)

$$n^{-1/2}\Delta_j(L) = n^{-1/2}\int_0^L w(t)dM_j(t) - n^{-1/2}\int_0^L w(t)\frac{y_j(t)dM(t)}{y(t)}$$
$$+n^{1/2}\int_0^L w(t)\left\{\frac{Y_j(t)}{Y(t)} - \frac{Y_j^f(t)}{Y^f(t)}\right\}dn(t) + o_p(1), \tag{2.8.1}$$

under H_0, where $dM_j(t) = dN_j(t) - Y_j(t)d\Lambda(t)$, $dM(t) = \sum_j dM_j(t)$, Λ denotes the common conditional cumulative hazard rate under the null hypothesis and $n(t)$, $y(t)$, $y_j(t)$ are the in probability limit of the averages $n^{-1}N(t)$, $n^{-1}(Y(t)$ and $n^{-1}Y_j(t)$, respectively.

By Corollary 3.1 in Datta and Satten (2000), we have $n^{-1}\{Y_j(t) - Y_j^f(t)\} \xrightarrow{P} 0$. Therefore, we can express the third term on the RHS of (2.8.1) as

$$n^{-1/2} \int_0^L \left\{ \frac{w(t)}{y(t)} \{Y_j(t) - Y_j^f(t)\} - \frac{w(t)y_j(t)}{y^2(t)} \{Y(t) - Y^f(t)\} \right\} dn(t) + o_p(1). \tag{2.8.2}$$

Next, by definition,

$$\begin{aligned} & n^{-1/2} \left\{ Y_j(t) - Y_j^f(t) \right\} \\ &= n^{-1/2} \left\{ \sum_{i=1}^n I(T_i \geq t,\ X_i = j,\ \delta_i = 0) - \sum_{i=1}^n \hat{P}_j(T_i, \infty) I(T_i \geq t,\ \delta_i = 0) \right\} \\ &= n^{-1/2} \sum_{i=1}^n \{ I(T_i \geq t,\ X_i = j,\ \delta_i = 0) - P_j(T_i, \infty) I(T_i \geq t,\ \delta_i = 0\} + \end{aligned}$$

$$n^{-1/2} \sum_{i=1}^n \{P_j(T_i, \infty) - \hat{P}_j(T_i, \infty)\} I(T_i \geq t,\ \delta_i = 0). \tag{2.8.3}$$

Next express the second term on the RHS of (2.8.3) as

$$-n^{1/2} \int_0^\infty \{\hat{P}_j(z, \infty) - P_j(z, \infty)\} I(z \geq t) dn^C(z) + o_p(1). \tag{2.8.4}$$

The expression within the braces of (2.8.4) is the $(0, j)$-th element of the matrix $\hat{\boldsymbol{P}}(s,t) - \boldsymbol{P}(s,t)$ for $s = z$ and $t = \infty$. Therefore by Andersen *et al.* (1993, 4.4.6), we have after some algebra

$$n^{1/2}\{\hat{P}_j(z, \infty) - P_j(z, \infty)\} = -n^{-1/2} \sum_{i=1}^n \int_z^\infty \frac{S(u)}{S(z)y(u)} P_j(u, \infty) dM_{\cdot,i}(u) + o_p(1). \tag{2.8.5}$$

Combining (2.8.3), (2.8.4), and (2.8.5), we obtain the linear representation for the first part of the expression with braces of RHS (2.8.2). A linear representation for the second part follows in the same way by summing over j' since $Y(t) - Y^f(t) = \sum_{j'} \{Y_{j'}(t) - Y_{j'}^f(t)\}$.

2.9 References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics*, **6**, 701-726.

[2] Aly, E.A.A., Kochar, S.C. and McKeague, I.W. (1994). Some tests for comparing cumulative incidence functions and cause-specific hazard rates, *Journal of the American Statistical Association*, **89**, 994-999.

[3] Andersen, P.K., Borgan, ⊘, Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag: New York, 1993.

[4] Bagai, I, Deshpande, J. V. and Kochar, S. (1989a). Distribution-free tests for stochastic ordering in the competing risks model, *Biometrika*, **76**, 775-781.

[5] Bagai, I., Deshpande, J. V. and Kochar, S. (1989b). A distribution-free test for the equality of failure rates due to two competing risks, *Communications in Statistics, Theory and Methods*, **18**, 107-120.

[6] Buyske, S., Fagerstrom, R. and Ying, Z. (2000). A Class of Weighted Log-Rank Tests for Survival Data when the Event is Rare, *Journal of the American Statistical Association*, **95**, 249-258.

[7] Carriere, K. C. and Kochar, S. C. (2000). Comparing sub-survival functions in a competing risks model, *Lifetime Data Analysis*, **6**, 85-97.

[8] Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant survival Data, *Journal of the American Statistical Association*, **72**, 27-36.

[9] Datta, S. and Satten G. A. (2000). Estimating future stage entry and occupation probabilities in a multistage model based on randomly right-censored data, *Statistics and Probability Letters*, **50**, 89-95.

[10] Datta, S., Satten, G. A. and Datta, S. (2000). Nonparametric estimation for three-stage irreversible illness-death model, *Biometrics*, **56**, 841-847.

[11] Dewan, I., Deshpande, J. V. and Kulathinal, S. B. (2004). On Testing Dependence between Time to Failure and Cause of Failure via Conditional Probabilities, *Scandinavian Journal of Statistics*, **31**, 79-91.

[12] Dykstra, R., Kochar, S. C. and Robertson, T. (1998). Restricted tests for testing independence of time to failure and cause of failure in a competing risks model, *Canadian Journal of Statistics*, **26**, 57-68.

[13] Fleming, T. R. and Harrington, D. P. (1991) *Counting Processes and Survival Analysis*, Wiley, New York.

[14] Fleming, T. R. and Harrington, D. P. (1981). A class of hypothesis tests for one and two sample censored survival data, *Communications in Statistics: Theory and Methods A*, **10**, 763-794.

[15] Fleming, T. R., O'Fallon, J. R., O'Bien, P. C. and Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data, *Biometrics*, **36**, 607-625.

[16] Gray, R. J. (1998). A class of K-sample tests for comparing the cumulative incidence of a competing risk, *Annals of Statistics*, **16**, 1141-1154.

[17] Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data, *Biometrika*, **69**, 553-566.

[18] Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.

[19] Kochar, S. C. and Proschan, F. (1991). Independence of time and cause of failure in the multiple dependent competing risks model, *Statistica Sinica*, **1**, 295-299.

[20] Kulathinal, S. B. and Gasbarra, D.(2002). Testing equality of case-specific hazard rates corresponding to m competing risks among k gorups, *Lifetime Data Analysis*, **8**, 147-161.

[21] Lagakos, S. W. (1978). A Covariate Model for Partially Censored Data subject to Competing Causes of Failure, *Applied Statistics*, **27**, 235-241.

[22] Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of the competing risk data, *Applied Statistics*, **34**, 201-211.

[23] Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing-risk studies, *Statistics in Medicine*, **16**, 901-910.

[24] Lindkvist, H. and Belyaev, Y. (1998). A class of nonparametric tests in the competing risk model for comparing two samples, *Scandinavian Journal of Statistics*, **25**, 143-150.

[25] Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies, *Journal of the American Statistical Association*, **86**, 770-778.

[26] Satten, G. A. and Datta, S. (1999). Kaplan-Meier representation of competing risk estimates, *Statistics & Probability Letters*, **42**, 299-304.

[27] Sun, Y. and Tiwari, R. C. (1995). Comparing cause-specific hazard rates of a competing risks model with censored data, *Analysis of Censored data. IMS Lecture Notes-Monograph Series*, **27**, 255-270.

[28] Wu, L. and Gilbert, P. (2000). Flexible Weighted Log-Rank Tests Optimal for detecting Early and/or Late Survival Differences, *Biometrics*, **58**, 997-1004.

[29] Zhao, H. and Tsiatis, A. A. (2001). Testing Equality of Survival Functions of Quality-Adjusted Lifetime, *Biometrics*, **57**, 861-867.

Chapter 3

U - statistics for right censored data [†]

[†]Bandyopadhyay, D., and Datta, S. To be submitted to *Journal of the American Statistical Association*

3.1 Abstract

A right-censored version of a U-statistic with a general kernel of size m is introduced by the principle of a mean preserving reweighting scheme popularized by Jamie Robins and others. Its extension to handle dependent censoring is also proposed. A doubly-robust version of this reweighted U-statistic is also introduced to preserve consistency in the face of model misspecifications. Using two different kernels, we study the performance measures of our U-statistic by simulation. Its asymptotic normality and an expression of its standard error are obtained through a martingale argument. The asymptotic normality is also assessed by using probability-probability plots. Using a Kendall's τ kernel, we obtain a test statistic for testing independence of time to failure and cause of failure in a competing risk problem. Using extensive simulation we study its performance by plotting its power curve. Its functionality is also assessed by applying it on a real data set.

Key Words: Dependent censoring, Doubly-robust, Kaplan-Meier, Kendall's tau, Right-censoring, U-statistic.

3.2 Introduction

Consider a sequence of independent random variables $X_1, \ldots, X_n$ with a common distribution function F. We estimate the population mean $\mu = EX_1$ by the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Hoeffding (1948) suggested the name U-statistic and generalised the notion of averaging over observations $X_1, \ldots, X_n$ in the following way. Let h be a measurable function such that

$$h : \mathbb{R}^m \to \mathbb{R} : (x_1, \ldots, x_m) \mapsto h(x_1, \ldots, x_m)$$

symmetric in its m arguments and $Eh^2(X_1, \ldots, X_m) < \infty$. Let P_{nm} be the collection of ordered m indices out of n indices drawn without replacement such that $P_{nm} = \{\underline{i} : (i_1, \ldots, i_m) \in \mathbb{N}^m : 1 \leq i_1 < \ldots i_m \leq n\}$. A U-statistic is obtained by averaging the

'terms' $h(x_{i_1}, \ldots, x_{i_m})$, $\underline{i} \in P_{nm}$, as

$$U_n(h) = \frac{1}{\binom{n}{m}} \sum_{P_{nm}} h(X_{i_1}, \ldots, X_{i_m}) \tag{3.2.1}$$

The U-statistic (3.2.1) is the nonparametric uniformly minimum variance unbiased estimator of the regular functional

$$\theta : \mathcal{L}_0 \to \mathbb{R} : F \longmapsto \theta_F = Eh(X_1, \ldots, X_m)$$

$$= \int_{\overset{\cdots}{R^m}} \int h(x_1, \ldots, x_m) dF(x_1) \ldots dF(x_m) \tag{3.2.2}$$

where $\mathcal{L}_0$ is a subset of the set $\mathcal{L}$ of one-dimensional distribution functions. It is also the minimizer with respect to α of

$$\sum_{P_{nm}} (h(X_{i_1}, \ldots, X_{i_m}) - \alpha)^2$$

The function h is called a kernel of degree m and we assume $E|h(X_1, \ldots, X_m)| < \infty$. Asymptotic properties of these statistics can be found in the work of Serfling (1980), Lee (1990), Bickel and Lehmann (1979), Randles and Wolfe (1979) etc. The study of U-statistics is important in several ways. In particular, under a positive variance condition, a U-statistic is asymptotically linear and normally distributed. Many statistical functionals and estimators are approximately U-statistics and so the theory provides an unified paradigm for study of distributional properties in the field of nonparametrics. The simple structure of U-statistics makes them ideal for studying general estimation processes like bootstrapping (Janssen, 1997) and jackknifing and generalising asymptotic theory that concerns behaviors of sequences of sample means. Also, application of the theory generates new statistics relevant to practical estimation problems.

In survival analysis of time-to-event data, subjects are followed from an initiating event and observed till a failure has taken place. In such studies, the study of the response variable

(here time to failure) becomes complicated due to the presence of several data incompleteness mechanisms. Perhaps the most common of such mechanisms is right-censoring. This means that apart from a set of survival times T_i^*, $i = 1, \ldots, n$ with a common distribution F, there exists another sequence of independent and identically distributed (iid) random variables (here time) $C_1, \ldots, C_n$, with a common distribution G such that, in reality, one observes the pairs $T_i = T_i^* \wedge C_i$, and $\delta_i = I(T_i^* \leq C_i)$, $i = 1, \ldots, n$. Most of the theoretical development in survival analysis with right censored data is based on the assumption of 'random censoring' (which is a stronger form of independent censoring). Under this model, one assumes that the true failure time T^* is statistically independent of the corresponding censoring time C. Furthermore, for simplicity of exposition we assume F is absoultely continuous.

A natural question is how the above definition and distribution theory of U-statistics could be modified under this right-censoring model. The simplest case of a U-statistic is a sample mean for a kernel of degree one; this amounts to using the empirical distribution of the T's as an estimator of F for averaging that puts equal weight of $1/n$ at each data value. But in case of censoring, using empirical distribution weights will lead to biased answers. Several estimators of F exists in the survival analysis literature, the most notable being the Kaplan-Meier estimator (Kaplan and Meier, 1958). In the usual counting process notation, a Kaplan-Meier estimator of F is given by

$$1 - \hat{F}(x) = \prod_{\tau_i \leq x} \left(1 - \frac{\Delta N(\tau_i)}{Y(\tau_i)}\right)$$

where

$$N(t) = \sum_{i=1}^{n} I(T_i \leq t, \delta_i = 1) \tag{3.2.3}$$

counts the number of observed failures in the time interval [0, t] and

$$Y(t) = \sum_{i=1}^{n} I(T_i \geq t) \tag{3.2.4}$$

counts the number of individuals that are at risk of failure at time t; the τ are the distinct ordered observed failures. Let $W_i = \hat{F}(T_i) - \hat{F}(T_i-)$ be the weight assigned to T_i by the Kaplan-Meier estimator. Define for a degree one kernel h,

$$U_{1n}(h) = \int h d\hat{F} = \sum_{i=1}^{n} h(T_i)W_i \tag{3.2.5}$$

This could be taken as a U-statistic of degree one for right censored data. For uncensored random variables, it reduces to the usual sample mean. Stute and Wang (1993a) and Stute (1995) provide a complete extension of the SLLN to this random censorship model. Akritas (1986) and Gijbels and Veraberbeke (1991) consider one-sample U-statistics where the kernel is of bounded variation. Stute and Wang (1993b) extend those results to their full generality for multisample U-statistics. Bose and Sen (1999, 2002) introduce the following U-statistic of degree two under random censorship where they normalized the weights:

$$U_{2n}(h) = \frac{\sum\limits_{1\le i_1<i_2\le n} h(T_{i_1}, T_{i_2})W_{i_1}W_{i_2}}{\sum\limits_{1\le i_1<i_2\le n} W_{i_1}W_{i_2}} \tag{3.2.6}$$

They name this U-statisitc a "Kaplan-Meier U-statistic of degree two for randomly right censored data" and proved a strong law for it. They also show that the U-statistic defined via this estimator is asymptotically normal. Their approach avoids the stringent assumptions of Gijbels and Veraverbeke (1991) who consider similar functionals. But they wonder whether similar limit theorems could be established for kernel of degree greater than two without encountering formidable algebra. Another aspect of research in this direction is the estimation of Kendall's τ under random right censoring. Kendall's τ has become a cornerstone kernel of the U-statistic literature. It measures deviations among concordances and discordances. Brown, Hollander and Korwar (1974), Weier and Basu (1980) and Oakes (1980) all propose estimation under right-censoring but none of the estimators are consistent when the true value of τ equals zero. Wang and Wells (2000) estimate τ using a suitable bivariate survival estimator into the integral form that defines τ. Betensky and Finkelstein (1999) consider

estimation of τ in a bivariate interval-censored data. In a somewhat relevant work, Martin and Betensky (2005) consider testing the quasi-independence of failure time and truncation times using conditional Kendall's τ.

The rest of the paper is organised as follows. In Section 3.3, we discuss the concept of weighted estimation with respect to survival analysis. In Section 3.4, we introduce our censored U-statistic under a random censoring setup. We also discuss very briefly (avoiding theoretical calculations) its extension to handle time dependent/independent covariates under the dependent censoring paradigm. A doubly-robust version of the censored U-statistic is also proposed to preserve consistency under certain model misspecifications. Section 3.5 studies the performance of this test statistic through simulation studies with varying kernels. In Section 3.6, we study the performance of this test for testing the independence between time to failure and cause of failure in a competing risk model by using Kendall's τ. The test is also applied on a real data set. Section 3.7 presents some discussion and future work. We defer the outline of the proof of asymptotic normality of the censored U-statistic to the Appendix.

3.3 Weighted estimation in survival analysis

The concept of weighted estimation seems to have its roots in the field of sample survey with the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) where the population parameter of interest is estimated by a BLUE (Best Linear Unbiased Estimator) after assigning specific weights to each unit of the population. In the context of survival analysis, similar weighting schemes are considered by Koul, Susarla and Van Ryzin (KSV hereafter) (1981) and in the pioneering work by Robins (Robins and Rotnitzky, 1992; Robins, 1993). To describe it briefly, the usual approach in survival analysis is to equate the hazard of failure in a censored experiment to that in the uncensored experiment and model it through a parametric, semi-parametric or non-parametric family. KSV (1981) was the first to observe

and make use of the following mean-preserving identity

$$E\left(\frac{\delta_i T_i}{1-G(T_i)}\right) = E(T_i^*) \tag{3.3.1}$$

in the context of deriving regression models for right censored data. Here $G(.)$ is the distribution function of the censoring times. Note that the expression within parenthesis in the left hand side of (3.3.1) is observable only if G is known. In reality this is not the case and hence we replace G with its Kaplan-Meier estimator $\hat{G}$. Thus, the weighted estimation approach replaces a sample average in the uncensored experiment by a weighted sum in the censored experiment. Furthermore, the weights are fairly simple to formulate; e.g., in the simple case of a survival function estimation under random censorship, a true failure time will be inversely weighted by the survival function of the censoring variable at the time of failure, whereas a censored observation will receive zero weight. This is called inverse probability of censoring weighted (IPCW hereafter) estimation. The numerical calculation of these weights requires a model for the hazard of censoring rather than the hazard of failure as in the usual approach. The advantages of this approach are (i) specification of censoring hazard is an easier problem at least in principle and more importantly the resulting estimator will be relatively robust with respect to its misspecification since it enters the calculation in an indirect way, (ii) formulation of the appropriate censored version of an estimator is simpler since one does not need to connect it through the hazard rate and (iii) the resulting estimator retains the sum form and is therefore suitable for theoretical analysis. In this context, Satten and Datta (2001) showed that the Kaplan-Meier estimator of the survival function can be represented as an IPCW average of identically distributed terms, the weights being related to the survival function of the censoring times. Satten, Datta and Robins (2001) extend the marginal survival function to dependent censoring in the presence of covariates through this IPCW approach. More detail on this IPCW estimation scheme appears in the works of Satten and Datta (2002) and Rotnitzky and Robins (2005). Satten and Datta (2001) and

Datta (2005) showed the equivalence of IPCW approach with other approaches in simple estimation problems without covariates.

3.4 U-statistics for Right Censored Data

3.4.1 Random censoring

For the bulk of this paper, we assume the standard right censoring model generating i.i.d data $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$, $1 \leq i \leq n$. Under this model, the possibly unobserved failure times T and the censoring time C are assumed to be independent. In the next subsection, we briefly state how our approach can be extended to more general censoring models where the dependence between failure and censoring times is carried through a set of observed covariables.

Following the mean-preserving reweighting approach of Koul *et al.* (1981), we define a censored data U-statistic based on kernel h as

$$U = \frac{1}{\binom{n}{m}} \sum_{P_{nm}} \frac{h(T_{i_1}, \ldots, T_{i_m}) \prod_{\ell \in \underline{i}} \delta_\ell}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)} \tag{3.4.1}$$

where K_c is the survival function of the censoring variable C. It is easy to see that U itself is a U-statistic based on the pairs (T_i, δ_i),$1 \leq i \leq n$; its kernel $\mathcal{H}$ is of order m and is given by

$$\mathcal{H}(T_{i_1}, \delta_{i_1}; \ldots; T_{i_m}, \delta_{i_m}) = \frac{h(T_{i_1}, \ldots, T_{i_m}) \prod_{\ell \in \underline{i}} \delta_\ell}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)} \tag{3.4.2}$$

It is easy to see that U is mean preserving since

$$E\left[\frac{h(T_{i_1}, \ldots, T_{i_m}) \prod_{\ell \in \underline{i}} \delta_\ell}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)}\right] = E\left[\frac{h(T_{i_1}^*, \ldots, T_{i_m}^*) \prod_{j=1}^{m} I(T_{i_j}^* \leq C_{i_j})}{\prod_{j=1}^{m} K_c(T_{i_j}^* -)}\right]$$

$$= E\left[E\left(\frac{h(T_{i_1}^*, \ldots, T_{i_m}^*) \prod_{j=1}^{m} I(T_{i_j}^* \leq C_{i_j})}{\prod_{j=1}^{m} K_c(T_{i_j}^* -)} \middle| T_{i_1}^*, \ldots, T_{i_m}^*\right)\right]$$

$$= E\left[\frac{h(T^*_{i_1},\ldots,T^*_{i_m})}{\prod_{j=1}^{m} K_c(T^*_{i_j}-)}\prod_{j=1}^{m} E\left(I(T^*_{i_j}\le C_{i_j})\Big|T^*_{i_1},\ldots,T^*_{i_m}\right)\right]$$

$$= E\left[\frac{h(T^*_{i_1},\ldots,T^*_{i_m})}{\prod_{j=1}^{m} K_c(T^*_{i_j}-)}\prod_{j=1}^{m} K_c(T^*_{i_j}-)\right] = E\left(h(T^*_{i_1},\ldots,T^*_{i_m})\right) = \theta \tag{3.4.3}$$

Note however, K_c is unknown in practice; it needs to be estimated by the Kaplan-Meier formula where we reverse the role of censored and failed observations. Substituting this estimator $\hat{K}_c$ into (3.4.1) we get the following U-statistic that is calculable from right censored data

$$\hat{U} = \frac{1}{\binom{n}{m}}\sum_{P_{nm}} \frac{h(T_{i_1},\ldots,T_{i_m})\prod_{\ell\in\underline{i}}\delta_\ell}{\prod_{\ell\in\underline{i}}\hat{K}_c(T_\ell-)} \tag{3.4.4}$$

The expression in (3.4.4) is very similar to (3.2.6) above considered by Bose and Sen (2002) for a kernel of size two except for the normalization. More generally the above U statistic is related to the Kaplan-Meier U statistic of order m (by generalizing (3.2.6)) as follows:

$$\hat{U} = \frac{n^m}{\binom{n}{m}}\sum_{P_{nm}} h(T_{i_1},\ldots,T_{i_m})W_{i_1}\ldots W_{i_m} \tag{3.4.5}$$

where P_{nm} and W_{i_j}'s are defined earlier. This follows from the Satten and Datta (2001) result who showed that the Kaplan Meier estimator is equivalent to an IPCW estimator.

The weighted average form makes the asymptotic analysis of the U-statistic for right censored data possible. For the theoretical analysis we introduce the following counting process notations. Let $N^c_i(t) = I(T_i \le t, \delta_i = 0)$ be the counting process of censoring for the i th individual and $M^c_i(t) = N^c_i(t) - \int_0^t \alpha^c(u)Y_i(u)du$ be the associated martingale where α^c is the censoring hazard and $Y_i(t) = I(T_i \ge t)$. Also let $\bar{n}(t) = P(T_i \le t, \delta_i = 1)$, $y(s) = P(T_i \ge t)$ and

$$w(s) = \frac{1}{y(s)} \int_{s+}^{\infty} \frac{h_1(u)}{K_c(u-)} d\bar{n}(u).$$

Theorem 3.1: We have, as $n \to \infty$,

$$\sqrt{n}(\hat{U} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

where

$$\sigma^2 = m^2 Var\left(\frac{h_1(T_1)\delta_1}{K_c(T_1-)} + \int w(s) dM_1^c(s) \right).$$

with

$$h_1(t_1) = E(h(t_1, T_2^*, \ldots, T_m^*) | T_1^* = t_1).$$

An outline of the proof of the theorem is given in the Appendix. Note that the asymptotic variance of $\hat{U}$ is easy to estimate. Let

$$\hat{\sigma}^2 = \frac{m^2}{n-1} \sum_{i=1}^{n} (V_i - \bar{V})^2,$$

$$V_i = \frac{\hat{h}_1(T_i)\delta_i}{\hat{K}_c(T_i-)} + \hat{w}(T_i)\delta_i^c - \sum_{j=1}^{n} \frac{\hat{w}(T_j) I(T_i \geq T_j) \delta_j^c}{Y(T_j)}$$

where the formulas for the estimated functions $\hat{h}_1$ and $\hat{w}$ are given in the Appendix. Then the asymptotic standard error (SE) of $\hat{U}$ will be given by $n^{-1/2}\hat{\sigma}$.

3.4.2 Dependent censoring

In many applications, one may observe failures along with a collection of covariables (fixed or time varying) $\underline{Z_i} = \{Z_{ij}(s) : 0 \leq s < t, j \leq J\}$ for the ith individual. We assume that the $\underline{Z}$ are observable for all individuals including those whose failures are censored at least up to time T. Dependent censoring in survival analysis occurs when there is correlation between failure and censoring times. If this dependence is carried out by covariates such that for

fixed levels of covariates failure and censoring times are uncorrelated, then it is possible to account for dependent censoring. One can now postulate a model for that censoring hazard conditional on the covariables which accounts for the dependence between the failure times and the censoring times. This is implicit in the following technical condition on the hazard for censoring:

$$\begin{aligned}&\lim_{dt\to 0}\frac{P\{t\le C_i<t+dt,\delta_i=0|\underline{Z_i}(u),0\le u<t,T_i\ge t,T_i^*\}}{dt}\\&=\lim_{dt\to 0}\frac{P\{t\le C_i<t+dt,\delta_i=0|\underline{Z_i}(u),0\le u<t,T_i\ge t\}}{dt}\end{aligned}\tag{3.4.6}$$

The above equation means that given $\underline{Z}$, the (future) failure time does not affect the current hazard of being censored. We denote this censoring hazard by $\lambda_c(t)$ and the corresponding integrated hazard by $\Lambda_c(t)$, i.e., $\Lambda_c(t)=\int_0^t\lambda_c(s)ds$. Under this setting, K_c can be defined as $K_c(t)=\exp\{-\Lambda_c(t)\}$.

A flexible hazard model increases the chance of obtaining an estimate of K_c that is close to its true value. One such model is Aalen's linear hazard model (Aalen 1980, 1989) where one expresses the censoring hazard as

$$\lambda_c(t)=\sum_{j=1}^{p}\beta_j(t)U_{ij}(t)\tag{3.4.7}$$

where each $\beta_j(t)$ is an unknown function, and where we assume that the first components are $U_{i0}\equiv 1$, $U_{ij}(t)=\phi_j(\underline{Z_i}(u):0\le u<t)$ are predictable functions of the covariates, $1\le j\le p$. We let $B_j(t)=\int_0^t\beta_j(s)ds$ and let $U_i(t)$ be the vector $(U_{i0}(t),\ldots,U_{ip}(t))^T$. Then Aalen's estimator of the vector $B(t)=(B_0(t),\ldots,B_p(t))^T$ is given by

$$\hat{B}(t)=\sum_{i=1}^{n}I(T_i\le t)(1-\delta_i)A^{-1}(t)U_i(t)\tag{3.4.8}$$

where the matrix $A(t)$ is given by

$$A(t)=\sum_{i=1}^{n}I(T_i\ge t)U_i(t)U_i^T(t)\tag{3.4.9}$$

Given the estimator $\hat{B}(t)$, we can write

$$\hat{\Lambda}_c(t|\underline{Z_i}(t)) = \sum_{j=0}^{p} \int_0^t U_{ij}(t) d\hat{B}_j(t)$$

$$= \sum_{i'=1}^{n} I(T_{i'} \leq t)(1-\delta_{i'}) U_i^T(T_{i'}).A^{-1}(T_{i'}).U_{i'}(T_{i'}), t \leq T_i \tag{3.4.10}$$

This leads to an estimate of $K_c(t)$ as $\hat{K}_c(t) = \exp\{-\hat{\Lambda}_c(t)\}$. The formula for the censored data U-statistic (3.4.4) remains intact; the only change needed is reflected in the definition of $\hat{K}_c$. Aalen's model is especially flexible because it fits a function $\beta_j(t)$ to describe the effect of each covariate, even time-independent. This recipe was carried out in detail in Satten *et al.* (2001) for extending the Kaplan-Meier in the case of dependent censoring. However, this flexibility has limited the use of Aalen's linear hazard model as a model for understanding the effect of variables on survival times for many reasons, viz. (a) estimates of $\hat{\Lambda}_c(t|\underline{Z_i}(t))$ may not be monotone increasing and (b) the matrix $A(t)$ may fail to have full rank at some time τ. Details of the estimation of $K_c(t)$ using Aalen's linear hazard model while remaining unaffected by the small sample difficulties stated above is described in Satten and Datta (2004). Using similar martingale arguments as in Satten *et al.* (2001), the U defined above with the choice of K_c remains unbiased for θ even if T and C are not independent but (3.4.6) holds. As a part of future work, we would establish its asymptotic normality by combining proof of Theorem 3.1 with the martingale arguments as in Satten *et al.* (2001).

3.4.3 A doubly-robust censored U-statistic

In general, estimators based on inverse probability weighting (IPW) may not be efficient. Hence, a plausible proposal for the estimator in (3.4.4) to gain efficiency is to modify the estimator into its doubly-robust (DR) (Robins *et al.*(1999), Section 7) version, say U_{DR}. The estimator is called 'doubly-robust' (DR) or 'doubly-protected' in this case because $E(U_{DR}) = \theta$, if either (i) the model for λ_c is correctly specified or (ii) the model for $T^*|(T, \delta, \underline{Z}(T))$ is

correctly specified, although consistency is not preserved against the simultaneous misspecification of both models. In reality, we cannot expect always the correctness of both the models, hence the best one can achieve is a DR estimator. For more on doubly-robustness, see the monograph by van der Laan and Robins (2003). We propose a DR estimator as

$$\begin{aligned} U_{DR} &= \frac{1}{\binom{n}{m}} \sum_{P_{nm}} \Bigg[\frac{h(T_{i_1}, \ldots, T_{i_m}) \prod_{\ell \in \underline{i}} \delta_\ell}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)} \\ &- \frac{\prod_{\ell \in \underline{i}} \delta_\ell - \prod_{\ell \in \underline{i}} K_c(T_\ell -)}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)} g\Big((T_{i_1}, \delta_{i_1}, \underline{Z}_{i_1}(T_{i_1})), \ldots, (T_{i_m}, \delta_{i_m}, \underline{Z}_{i_m}(T_{i_m})) \Big) \Bigg] \end{aligned} \quad (3.4.11)$$

where g is a suitable chosen function of the observed data. Note that it follows by simple conditional expectation argument that U_{DR} remains unbaised for $\theta = Eh$, no matter what function g we choose; furthermore it is also a statistic of order m with the kernel $\mathcal{H}_C$, say. Using the variance reduction property of a projection, one can see that the best choice of m is such that the new kernel $\mathcal{H}_C\Big((T_{i_1}, \delta_{i_1}), \ldots, (T_{i_m}, \delta_{i_m}) \Big)$, say is the residual of the projection of the original kernel $\mathcal{H}$ onto the space of mean zero functions

$$\Lambda_{i_0} = \left\{ \frac{\prod_{\ell \in \underline{i}} \delta_\ell - \prod_{\ell \in \underline{i}} K_c(T_\ell -)}{\prod_{\ell \in \underline{i}} K_c(T_\ell -)} g\Big((T_{i_1}, \delta_{i_1}, \underline{Z}_{i_1}(T_{i_1})), \ldots, (T_{i_m}, \delta_{i_m}, \underline{Z}_{i_m}(T_{i_m})) \Big) \right\} \quad (3.4.12)$$

where g is arbitrary L_2 function. By direct minimization of an L_2 risk, we conclude

$$g_{opt} = E\Big(h(T^*_{i_1}, \ldots, T^*_{i_m}) \Big| (T_{i_1}, \delta_{i_1}, \underline{Z}_{i_1}(T_{i_1})), \ldots, (T_{i_m}, \delta_{i_m}, \underline{Z}_{i_m}(T_{i_m})) \Big)$$

We can see more formally that indeed U_{DR} will have smaller variance than U for each n, since

$$Var(U_{DR}) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k^2(\mathcal{H}_C),$$

(see for example, lemma A in Serfling, 1980), where $\zeta_k^2(\mathcal{H}_C)$ is the variance of $\mathcal{H}_{C,k}$ that is the projection of $\mathcal{H}_C$ onto the first k coordinates. However, it follows from definitions that

$\mathcal{H}_{C,k}$ is the residual of the projection of $\mathcal{H}_k$ onto a corresponding set of zero mean functions. Therefore, $\zeta_k^2(\mathcal{H}_C) \leq \zeta_k^2(\mathcal{H})$ leading to $Var(U_{DR}) \leq Var(U)$.

Once again, in practice, we need to substitute estimators $\hat{K}$ and $\hat{g}_{opt}$ in (3.4.11). Modeling λ_c and estimation of $\hat{K}_c$ has been discussed in the previous subsection. A working strategy for computing $\hat{g}_{opt}$ in the case of fixed covariates will be as follows. First fit a Cox's model to the observed data $(T_i, \delta_i, \underline{Z}_i), 1 \leq i \leq n$ leading to an estimated hazard of failure $\hat{\lambda}_T(t)$= $\exp(\beta.\underline{Z})\hat{\lambda}_0(t)$. The following Monte Carlo calculations can then be used to compute $\hat{g}_{opt}$. For each censored individual i, generate B independent replicates of T^* values $T^*_{i_1}, \ldots, T^*_{i_B}$ from the estimated $\hat{\lambda}_T(t), t \geq T_i$. This can be easily done by successive Bernoulli sampling with success probabilities calculated using this hazard on a fine grid of time points. For an uncensored individual i, let $T^*_{ij} = T^*_i, 1 \leq j \leq B$. Then $\hat{g}_{opt}$ is given by

$$\hat{g}_{opt} = B^{-1} \sum_{j=1}^{B} h(T^*_{i_1 j}, \ldots, T^*_{i_m j})$$

It can be easily checked using the equivalence of a reweighting and mean imputation (eg., Datta 2005) for estimating a mean using right censored data that the correction term will be identically equal to zero when the order of the kernel is one and there is no covariates (e.g., independent right censoring), there won't be any gain in asymptotic efficiency since the asymptotic variance of both U and $\hat{U}_{DR}$ will be $\frac{m^2}{n}\zeta_1^2(\mathcal{H})$. It is interesting to note however that the correction term is not algebraically equal to zero for higher order kernels even in absence of covariates as we show in Example 3.4.1.

Example 3.4.1

We consider computing the correction term for survival data without covariates which is given as

$$C = \frac{1}{\binom{n}{m}} \sum_{P_{nm}} \left(\frac{\prod\limits_{\ell \in \underline{i}} \delta_\ell - \prod\limits_{\ell \in \underline{i}} K_c(T_\ell -)}{\prod\limits_{\ell \in \underline{i}} K_c(T_\ell -)} \right) g\Big((T_{i_1}, \delta_{i_1}), \ldots, (T_{i_m}, \delta_{i_m})\Big)$$

$$= \frac{1}{\binom{n}{m}} \sum_{P_{nm}} \left(\frac{\prod\limits_{\ell \in \underline{i}} \delta_\ell}{\prod\limits_{\ell \in \underline{i}} K_c(T_\ell -)} - 1 \right) E\Big(h(T^*_{i_1}, \ldots, T^*_{i_m}) \Big| (T_{i_1}, \delta_{i_1}), \ldots, (T_{i_m}, \delta_{i_m})\Big) \qquad (3.4.13)$$

We consider a second order kernel $h(x_1, x_2) = (x_1 - x_2)^2$.

Data 1

$T : 1, 3, 4^+, 5$

Here T denotes the failure times and T^+ denotes a censored observation. Denote $\hat{S}$ and $\hat{K}_c$ as the Kaplan-Meier(KM) estimators of the failure times and censoring times. $\hat{K}_c-$ denotes the KM shifted one unit to the right. Table 3.1 displays the values of the KM at different time points. Note that for computing m, 4^+ takes value 5 with probability 1. We now try to compute the different terms that contribute to the summation in (3.4.13).

Table 3.1: Kaplan Meier table for Data 1

T	1	3	4^+	5
δ	1	1	0	1
δ_c	0	0	1	0
$\hat{S}$	3/4	1/2	1/2	0
$\hat{K}_c$	1	1	1/2	1/2
$\hat{K}_c-$	1	1	1	1/2

For $(1, 3)$: $(\frac{1}{1} - 1).(1 - 3)^2 = 0$

For $(1, 4^+)$: $-(1 - 5)^2.1 = -16$

For $(1, 5)$: $(\frac{1}{1/2} - 1).(1 - 5)^2 = 16$

For $(3, 4^+)$: $-(3-5)^2 = -4$

For $(3, 5)$: $(\frac{1}{1/2} - 1).(3-5)^2 = 4$

For $(4^+, 5)$: $-(5-5)^2.1 = 0$

Adding all these terms gives us the value of the correction term $C = 0$.

Data 2

$T : 1, 3^+, 4, 5^+, 6$

We now construct the table as above which in this case is given in Table 3.2. The jump in the Kaplan-Meier value (for failure times) at 4 is $(4/5 - 8/15) = 4/15$ and at 6 is $(8/15 - 0) = 8/15$. The total jump is $4/5$. Thus, 3^+ takes value 4 with prob. $(\frac{4}{15}/\frac{4}{5}) = 1/3$ and 6 with prob. $(\frac{8}{15}/\frac{4}{5}) = 2/3$. Again, 5^+ takes value 6 with probability 1. We now compute the different terms that contribute to the summation in (3.4.13) as was done for Data 1.

Table 3.2: Kaplan Meier table for Data 2

T	1	3^+	4	5^+	6
δ	1	0	1	0	1
δ_c	0	1	0	1	0
$\hat{S}$	4/5	4/5	8/15	8/15	0
$\hat{K}_c$	1	3/4	3/4	3/8	3/8
$\hat{K}_c-$	1	1	3/4	3/4	3/8

For $(1, 3^+)$: $-(1-4)^2.(1/3) - (1-6)^2.(2/3) = -\frac{9+50}{3}$

For $(1, 4)$: $(\frac{1}{3/4} - 1).(1-4)^2 = \frac{9}{3}$

For $(1, 5^+)$: $-(1-6)^2.1 = -25$

For $(1, 6)$: $(\frac{1}{3/8} - 1).(1-6)^2 = \frac{125}{3}$

For $(3^+, 4)$: $-(4-4)^2.(1/3) - (6-4)^2.(2/3) = -\frac{8}{3}$

For $(3^+, 5^+)$: $-(4-6)^2.(1/3) - (6-6)^2.(2/3) = -\frac{4}{3}$

For $(3^+, 6)$: $-(4-6)^2.(1/3) - (6-6)^2.(2/3) = -\frac{4}{3}$

For $(4, 5^+)$: $-(4-6)^2.1 = -4$

For $(4, 6)$: $(\frac{1}{9/32} - 1).(4-6)^2 = \frac{92}{9}$

For $(5^+, 6)$: $-(6-6)^2.1 = 0.$

Adding these terms gives us:

$(-\frac{9+50}{3} + \frac{9}{3} + \frac{50}{3}) - \frac{8+4+4}{3} + \frac{92-36}{9} - 0$

$= -\frac{16}{3} + \frac{56}{9} = \frac{56-48}{9} = \frac{8}{9} \neq 0.$

Hence the correction term $C = \binom{5}{2}^{-1}.\frac{8}{9}$ is not equal to zero.

Example 3.4.2

We illustrate the computation of the DR U-statistic in presence of a binary covariate Z. A natural model to assume in this case is that T^* and C are conditionally independent given Z. Under this assumption, we could carry out the necessary computation by calculating the Kaplan-Meier estimators of the failure and censoring times in the two groups separately. This is illustrated through the following data set.

Data 3

$T : 1, 3^+, 4, 5, 6^+, 7, 8$

$Z : 0, 0, 0, 1, 1, 1, 1$

Here T denotes the failure time as above and Z denotes the binary covariate taking values 0 and 1. We construct the Kaplan-Meier table (Table 3.3) separately for the observations in the two groups. The calculations are tabulated in Table 3.3. From the table, the jump in the Kaplan-Meier value at 4 is $(2/3 - 0) = 2/3$, at 7 is $(3/4 - 3/8) = 3/8$ and at 8 is $(3/8 - 0) = 3/8$. Thus, 3^+ takes value 4 with prob. 1. Again, 6^+ takes value 7 with prob. 1/2 and 8 with prob 1/2. We now compute the different terms that contribute to the summand in (3.4.13) for Data 3.

Table 3.3: Kaplan Meier table for Data 3

T	1	3^+	4	5	6^+	7	8
Z	0	0	0	1	1	1	1
δ	1	0	1	1	0	1	1
δ_c	0	1	0	0	1	0	0
$\hat{S}$	2/3	2/3	0	3/4	3/4	3/8	0
$\hat{K}_c$	1	1/2	1/2	1	2/3	2/3	2/3
$\hat{K}_c-$	1	1	1/2	1	1	2/3	2/3

For $(1, 3^+)$: $-(1-4)^2.1 = -9$

For $(1, 4)$: $(\frac{1}{1/2} - 1).(1-4)^2 = 9$

For $(1, 5)$: $(\frac{1}{1} - 1).(1-5)^2 = 0$

For $(1, 6^+)$: $-(1-7)^2.1/2 - (1-8)^2.1/2 = -\frac{36+49}{2}$

For $(1, 7)$: $(\frac{1}{2/3} - 1).(1-7)^2 = \frac{36}{2}$

For $(1, 8)$: $(\frac{1}{2/3} - 1).(1-8)^2 = \frac{49}{2}$

For $(3^+, 4)$: $-(4-4)^2.1 = 0$

For $(3^+, 5)$: $-(4-5)^2.1 = -1$

For $(3^+, 6^+)$: $-(4-7)^2.1/2 - (4-8)^2.1/2 = -\frac{9+16}{2}$

For $(3^+, 7)$: $-(4-7)^2.1 = -9$

For $(3^+, 8)$: $-(4-8)^2 = -16$

For $(4, 5)$: $(\frac{1}{1/2} - 1).(4-5)^2 = 1$

For $(4, 6^+)$: $-(4-7)^2.1/2 - (4-8)^2.1/2 = -\frac{9+16}{2}$

For $(4, 7)$: $(\frac{1}{1/3} - 1).(4-7)^2 = 18$

For $(4, 8)$: $(\frac{1}{1/3} - 1).(4-8)^2 = 32$

For $(5, 6^+)$: $-(5-7)^2.1/2 - (5-8)^2.1/2 = -\frac{4+9}{2}$

For $(5,7)$: $(\frac{1}{2/3}-1).(5-7)^2 = 2$

For $(5,8)$: $(\frac{1}{2/3}-1).(5-8)^2 = \frac{9}{2}$

For $(6^+,7)$: $-(7-7)^2.1/2 - (8-7)^2.1/2 = -1/2$

For $(6^+,8)$: $-(7-8)^2.1/2 - (8-8)^2.1/2 = -\frac{1}{2}$

For $(7,8)$: $(\frac{1}{4/9}-1).(7-8)^2 = \frac{5}{4}$

Adding these terms gives us $\frac{1}{4} \neq 0$. Hence the correction term C is not equal to zero.

3.5 Simulation setup

In order to validate the performance of our test statistic for finite samples, we conducted simulation studies based on randomly right-censored data with two different choices of the kernel.

Example 1

For the kernel $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$; x_1, $x_2 \in \mathbb{R}^1$, the corresponding U-statistic equals s^2, the sample variance. Here $\theta(F) =$ variance of $F = \sigma^2(F) = \int (x-\mu)^2 dF(x)$. We choose T^* from a Weibull distribution with shape parameter 2.3 and scale parameter 2.0. The censoring times (C) are generated from a Weibull distribution with shape 1.5 and scale 4 to represent 25% censoring and Weibull with shape 0.3 and scale 5.0 to represent 50% censoring. We choose sample sizes n=200, 500 and 1000 with iteration size 5000 for our simulations. The theoretical variance of T^* is 0.6674 and we compare this with the mean of our re-weighted U-statistic. We also compare the variance of the censored U-statistic with the asymptotic variance in Table 3.5. Values within parenthesis denote asymptotic variance, all values being multiplied by their respective sample sizes.

It is seen from Table 3.4 that the estimated sample variance from the censored data (at both censoring levels) remains close to the true variance of the uncensored data. The

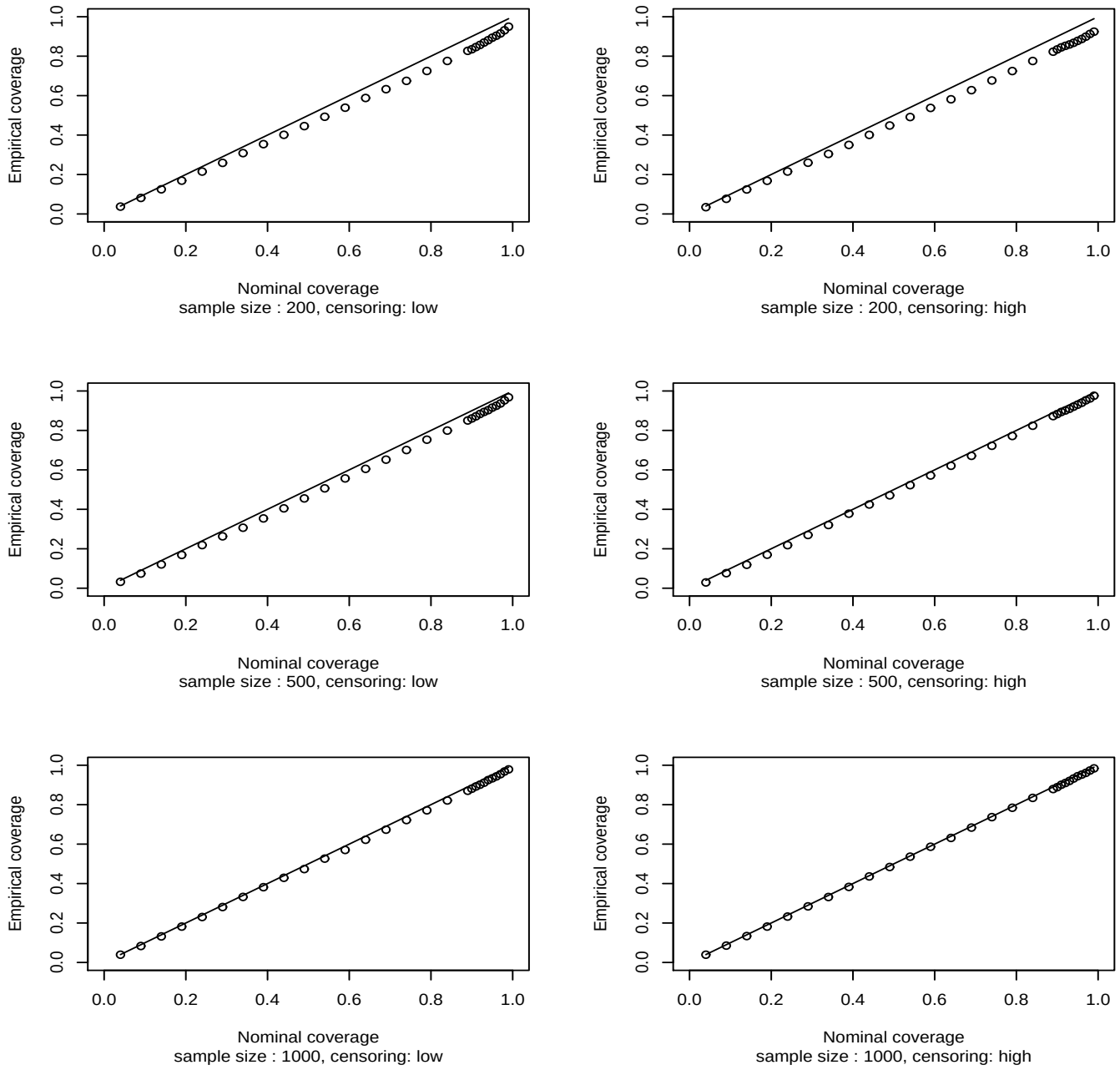

Figure 3.1: Plot of nominal coverage vs empirical coverage of confidence intervals for the population variance $\theta_F = \sigma^2$ with the line nominal=empirical overlayed

Table 3.4: Estimated sample variance from censored data, true value being 0.6674.

	Censoring level		
n	0%	25%	50%
200	0.6667	0.6556	0.6640
500	0.6671	0.6615	0.6655
1000	0.6674	0.6640	0.6659

Table 3.5: Comparing asymptotic variance with empirical variance of the U-statistics for Example 1. Values within parenthesis denote asymptotic variance

	Censoring level	
n	25%	50%
200	1.4632(1.4438)	1.9305(1.8586)
500	1.4487(1.4356)	1.9465(1.8839)
1000	1.4563(1.4317)	1.8921(1.8848)

precision however increases with higher sample sizes. From Table 3.5, we can see that while comparing the asymptotic variance with the empirical variance of the U-statistic, the former is smaller than the later. This is because the variance expression of the U-statistics is a sum of positive terms, the dominating term increases to the asymptotic variance and the nondominating ones disappear in the limit. The P-P plot (Figure 3.1) showing the plot of empirical coverage of confidence intervals of our estimated U-statistic against the nominal coverage (dotted line) is also examined with a plot of the line nominal=empirical (solid line) overlayed. In particular, we are interested in the nature of the plot at the

coverage level of 90% and above. Since the lines remain close to each other in all the situations, we can conclude the validity of the asymptotic normality of our censored U-statistic.

Example 2

For the kernel $h(x_1, x_2) = I(x_1+x_2 \leq 0)$; $x_1, x_2 \in \mathbb{R}^1$, the corresponding U-statistic estimates $\theta(F) = P_F(X_1 + X_2 \leq 0)$. The U-statistic calculates the average number of pairs (X_i, X_j) with $X_1 + X_2 \leq 0$, and can be used as a test for investigating whether the distribution of the observations is centered at zero. This test is asymptotically equivalent to the signed-rank statistic of Wilcoxon. We choose T^* as the event times from a log-normal distribution with parameter 0 and 1 and censoring times (C) from a lognormal distribution with parameters (0.954, 1) and (0,1) to represent about 25% and 50% censoring. Sample sizes of n=200, 500 and 1000 along with iteration size of 5000 was chosen for our simulation. We take $X = \log(\min(T^*, C))$ and consider this to be (X_i, X_j) pair. We expect the mean of our censored U statistic to be near 0.5, which is the value of θ under the assumption that the distribution function is continuous and symmetric about 0. See Table 3.6 for the comparison. Similar to Example 1, we compare the variance of the censored U-statistic with the asymptotic variance. See Table 3.7 for details.

Table 3.6: Estimating the Wilcoxon signed-rank statistic, true value being 0.5.

	Censoring level		
n	0%	25%	50%
200	0.5061	0.4999	0.4967
500	0.5081	0.4998	0.4990
1000	0.4998	0.4996	0.4998

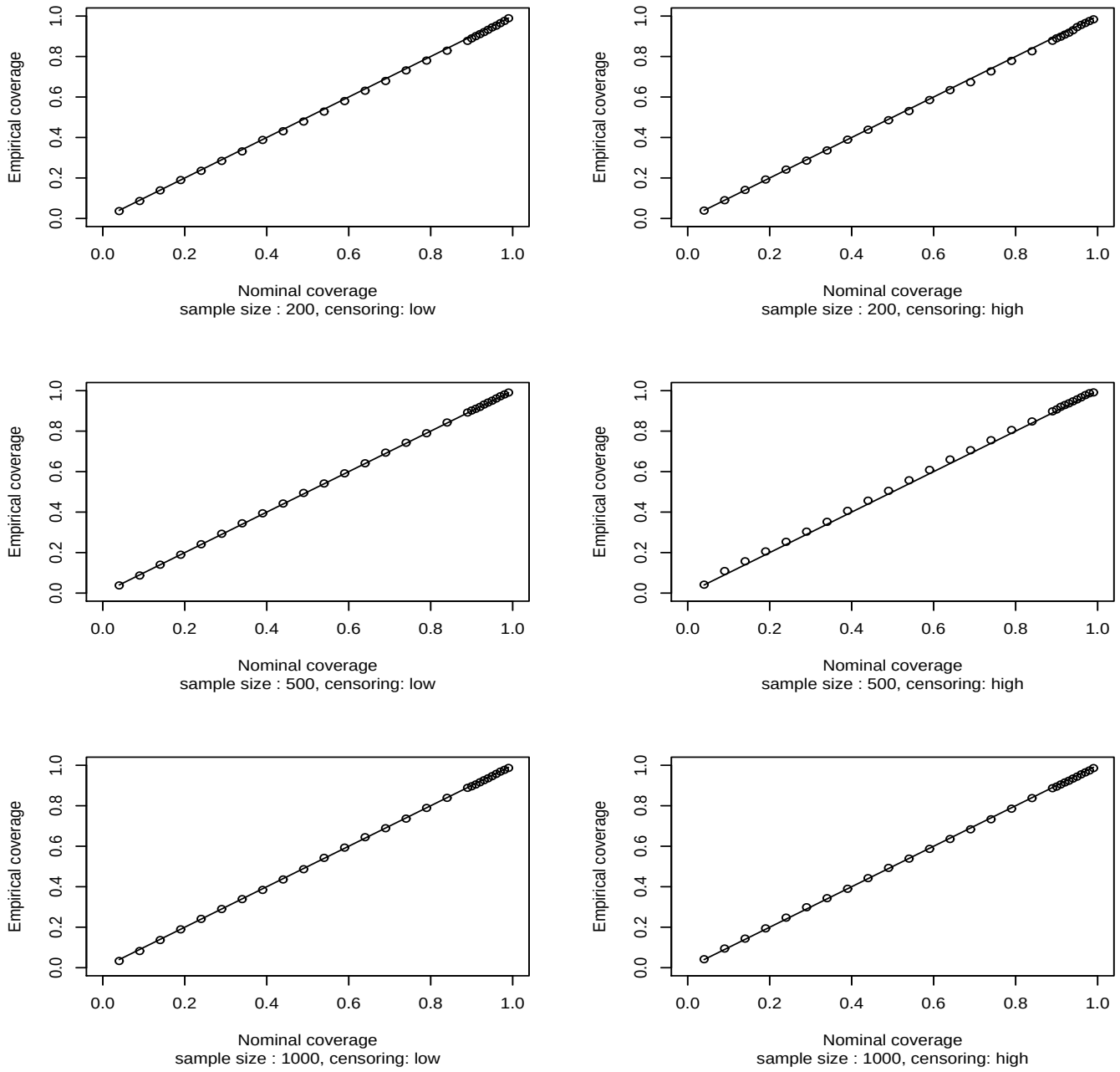

Figure 3.2: Plot of nominal coverage vs empirical coverage of confidence intervals for the population variance $\theta_F = P_F(X_1 + X_2 \leq 0)$ with the line nominal=empirical overlayed

Table 3.7: Comparing asymptotic variance with empirical variance of the U-statistics for Example 2. Values within parenthesis denote asymptotic variance

	Censoring level	
n	25%	50%
200	0.3739(0.3603)	0.5156(0.5006)
500	0.3560(0.3431)	0.4989(0.4826)
1000	0.3557(0.3441)	0.5230(0.5010)

From Table 3.6, it is seen that the censored U-statistic (at both censoring levels) estimates the true value $\theta_F = 0.5$ closely. Similar results to those of Example 1 were observed here while comparing the asymptotic variance of the U-statistic with the empirical variance in Table 3.7. Asymptotic normality in this case is also validated after examining the P-P Plot in Figure 3.2.

3.6 Applications to testing

We can study the dependence structures between time to failure and cause of failure in a competing risk setup using our censored version of the U-statistic. We consider the competing risk network as a multistate continuous time stochastic process $\{Z(t), t \in \mathcal{T}\}$ with a finite state space $\mathcal{S} = \{0, 1, 2\}$ having a tree topology and right-continuous sample paths: $Z(t+) = Z(t)$ where we assume that the states 1 and 2 are absorbing whereas state 0 is transient (the root node). Here $\mathcal{T} = [0, \tau]$ where τ is a large possibly observed time point ($\leq \infty$). Typically, for applications, τ will be taken to be the largest time where some event (failure) took place. Let T^* be the (possibly unobserved) time that an individual leaves stage 0 for a failure and let $d^* = S(T^*)$ be the failure stage. Here we study the properties of the conditional probability

function

$$\Phi_i(t) = P(d^* = i | T^* > t) = \frac{S_i(t)}{S(t)}, \quad i = 1, 2 \tag{3.6.1}$$

where $S_i(t) = P(T^* > t, d^* = i)$ denotes the sub-survival function and $S(t) = P(T^* > t)$ is the usual survival function. Specifically, we consider the problem of testing the null hypothesis

$$H_0 : \Phi_1(t) = \Phi_2(t)$$

Note, H_0 reduces to testing the independence of failure time and failure cause in a competing risk model. Here we propose a test based on the concept of concordance and discordance (using Kendall's τ) to test this hypothesis that extends a test proposed by Dewan *et al.*(2004) for uncensored data. The following notations are necessary for understanding the test statistic. A pair (T_i^*, d_i^*) and (T_j^*, d_j^*) is a concordant pair if $T_i^* > T_j^*$, $d_i^* = 2$, $d_j^* = 1$ or $T_i^* < T_j^*$, $d_i^* = 1$, $d_j^* = 2$ and a discordant pair if $T_i^* > T_j^*$, $d_i^* = 1$, $d_j^* = 2$ or $T_i^* < T_j^*$, $d_i^* = 2$, $d_j^* = 1$. The Kendall's τ test statistic assigns a score $+1$ to a concordant pair and a score -1 to a discordant pair. In other words, the kernel of the statistic is defined as

$$\psi_1(T_i^*, d_i^*; T_j^*, d_j^*) = \begin{cases} 1 & \text{if } T_i^* > T_j^*, d_i^* = 2, d_j^* = 1 \text{ or } T_i^* < T_j^*, d_i^* = 1, d_j^* = 2 \\ -1 & \text{if } T_i^* > T_j^*, d_i^* = 1, d_j^* = 2 \text{ or } T_i^* < T_j^*, d_i^* = 2, d_j^* = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.6.2}$$

The corresponding test statistic given by Dewan *et al.* (2004) is

$$U_2 = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \psi_1(T_i, d_i; T_j, d_j) \tag{3.6.3}$$

Dewan *et al.* (2004) showed that under H_0, $n^{1/2}U_2$ converges in distribution to $N(0, \sigma_f^2)$, where σ_f^2 is consistently estimated by

$$\hat{\sigma}_1^2 = (4/3)\hat{\phi}(1 - \hat{\phi})$$

and

$$\hat{\phi} = n^{-1} \sum_{i=1}^{n} I(d_i^* = 1)$$

In case of right censoring, T^* and d^* are not observed for the censored individuals. Let C be the right-censoring time. Our observed data is $T_i^* = T_i^* \wedge C_i$ and $d_i = d_i^* I(T_i^* \leq C_i), 1 \leq i \leq n$. Let $\delta_i = I(d_i = 0)$. An obvious adaptation of our definition of censored data U-statistic yields the following generalization of the Dewan *et al.* (2004) test statistic

$$U_2 = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \psi_1(T_i, d_i; T_j, d_j) \frac{\delta_i \delta_j}{\hat{K}_c(T_i-)\hat{K}_c(T_j-)}, \tag{3.6.4}$$

where $\hat{K}_c$ is as in Section 3.4. Note that it is indeed computable from the censored data and that it reduces to the Dewan *et al.* (2004) statistic if there is no censored data.

We reject the null hypothesis against the two sided alternative if $|n^{1/2}\hat{U}_2/\sigma| > z_{1-\alpha/2}$ where $\hat{\sigma}$ is obtained as in Section 3.4 and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)th$ percentile point of the standard normal distribution. To assess the test's performance, we conduct a simulation study using sample sizes of $n = 200$, 500 and 1500 and a nominal level of $\alpha = 0.05$. The power is computed as the proportion of rejections out of 2000 Monte-Carlo replications. We consider varying degrees of censoring rate. In case of uncensored data, we use the Dewan *et al.* (2004) formula for the asymptotic standard error.

3.6.1 Simulating event times from a bivariate distribution

We consider a similar simulation scheme (scheme (ii)) as of Section 2.5.1. In general, X and Y are dependent and being lifetimes, they should be non-negative. In this setup, we only observe (T, δ), where $T = \min(X, Y)$ is the failure time and $\delta = 2 - I(X \leq Y)$ is the cause of failure. For the null hypothesis, we generate (X', Y') from a bivariate normal distribution with mean vector (0,0), and variance-covariance matrix $((1, \rho), (\rho, 1))$, where ρ is chosen to be $-0.5, 0$ or 0.5. For the alternative, we generate (X', Y') from a bivariate normal distribution with mean vector $(0, (1-\alpha))$ and variance-covariance matrix $((1, \rho a), (\rho a, a^2))$, where $a = 1(0.1)1.5$.

Finally, we let $X = \exp(X')$ and $Y = \exp(Y')$. The censoring times are generated from log-normal distributions with variance parameter 1 and mean parameters 0.954 and 0 leading to light (8%-17%) and moderately heavy (28%-45%) censoring, respectively. In all cases, we take the nominal level α to be 5%.

3.6.2 Results

We report the empirical sizes for the above simulation scheme in Table 3.8. The empirical sizes of our test is close to the nominal level 0.05. During simulation, the effective sample size for a negative correlation is higher than that of a positive correlation, hence the size values are somewhat better in the case of negative correlation. It is also observed that the empirical sizes are marginally inflated from the nominal value in a few cases for heavy censoring but it stabilizes for higher sample sizes.

Table 3.8: Size of our U-statistics based test using $\alpha = 0.05$ under simulation

		Censoring		
		Light	Moderately	Uncensored
ρ	n		heavy	(Dewan *et.al.* (2004))
0.5	200	0.067	0.07	0.052
	500	0.061	0.067	0.051
	1500	0.059	0.063	0.049
0	200	0.063	0.072	0.053
	500	0.058	0.07	0.052
	1500	0.048	0.055	0.051
-0.5	200	0.055	0.067	0.048
	500	0.045	0.058	0.043
	1500	0.047	0.049	0.052

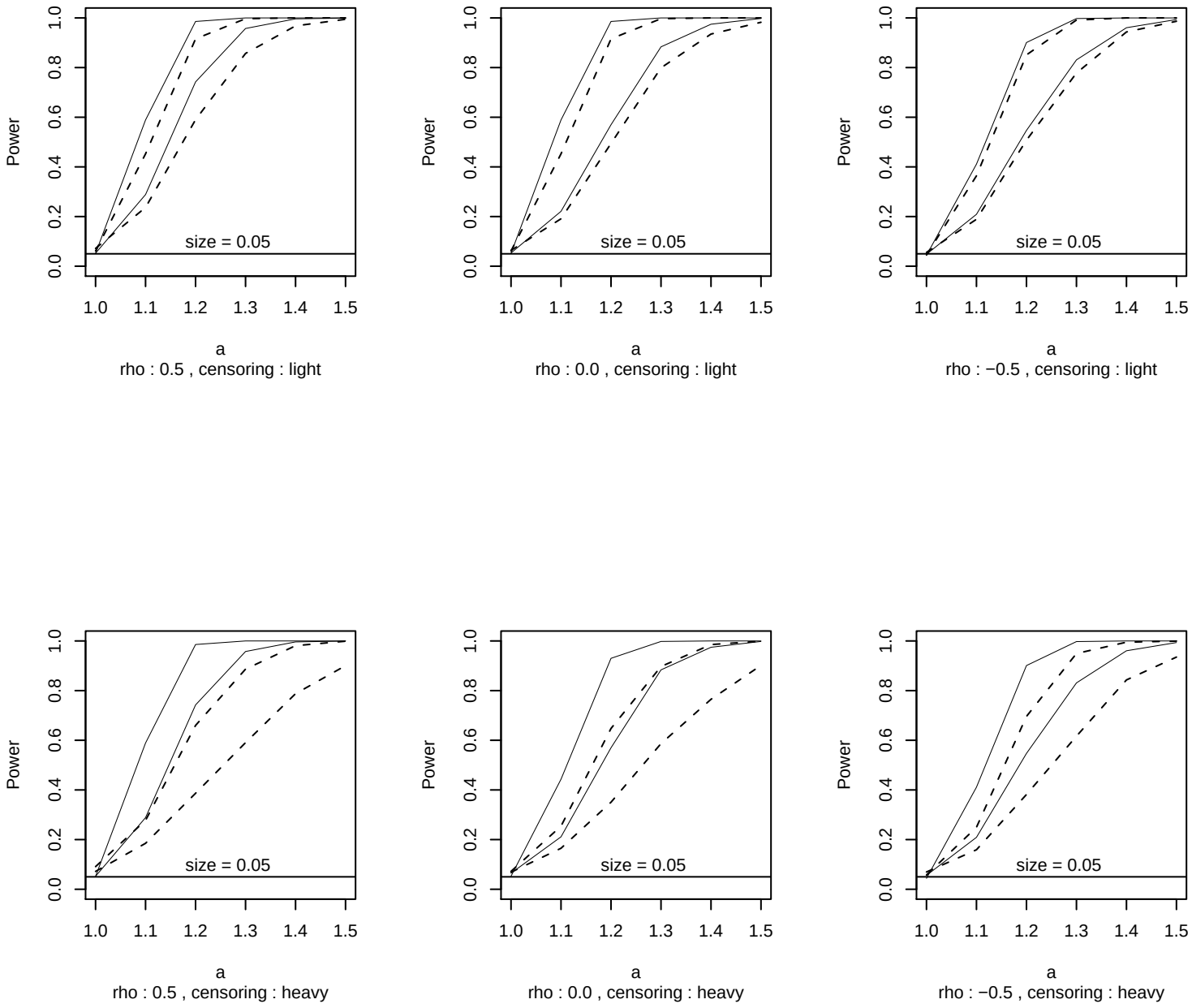

Figure 3.3: Power curves of U-statistic based tests for Simulation scheme (ii). The solid lines correspond to the Dewan *et.al.* (2004) test without censoring and the dashed lines corresponds to our U-statistic based test.

Figure 3.3 displays arrays of plots illustrating the power curves as a function of the alternative parameter a which was defined earlier. $a = 1$ corresponds to the null hypothesis. The set up is identical to the size study above and includes two choices of sample size, namely, 200 and 500. In each figure, we overlay the power curves obtained by our censored U-statistic using the Kendall's τ kernel with the power curves of the test in Dewan *et.al.* (2004) that is calculated using the full data. In the figure, solid lines denote power curves for the U-statistic based test without censoring and dashed lines denote power curve under censoring. It is seen that a considerable increase in power is observed with an increase in sample size. The power curves reveal that for low censoring and larger sample size, our test statistic produces a power profile that is very close to the full data test. However, heavy censoring has some adverse effect on the empirical power which tend to get better with larger sample size.

3.6.3 A Real Data Example

We consider the Stanford Heart Transplantation Program data as described in Chapter 2 of this dissertation. We formulate death due to 'transplant rejection' (coded as cause 1) and due to 'other causes' (coded as cause 2) as the two competing risks. We try to test the null hypothesis

$$H_0 : \Phi_1(t) = \Phi_2(t)$$

against two alternatives

$$H_1 : \Phi_1(t) \neq \Phi_2(t)$$

and

$$H_2 : \Phi_1(t) \leq \Phi_2(t)$$

The value of the U-statistic using the kernel given in (3.6.2) is 0.1204 and the value of the standard error obtained by using Theorem 3.1 is 0.049. The value of the Z-statistic is 2.449. It is immediate that H_1 is accepted for the two-sided test for testing H_0 vs H_1 at a

5% level. For a one-sided test for testing H_0 vs H_2 (based on concordance and discordance principle) using the same test statistic, it is expected that the number of discordances will be larger than the number of concordances under H_2. It can be seen from the data set that the failure time due to cause 1 is much larger overall in comparison to failure time due to cause 2. Hence the expected discordances should be larger than the expected concordances. For this data set for testing H_0 vs H_2, we accept H_2 thereby asserting that the probability of death due to 'transplant rejection' is less than the probability of death due to 'other causes'.

3.7 Discussion

In this chapter, we have proposed a censored version of the U-statistic in its true generality, i.e. valid for a general kernel of size m using the technique of inverse probability weighting. This provides substantial generalization over earlier results for a Kaplan-Meier U-statistic of degree two (Bose and Sen, 2002). We also formulate a U-statistic under dependent censoring where the dependence between failure times and censoring times can be explained through a set of observed covariates. A further efficiency correction using the idea of doubly robust (DR) estimators is proposed and two simple examples are illustrated.

A motivating application for this work was a Kendall's τ test recently introduced by Dewan *et al.* (2004) with uncensored data. This test examines the independence hypothesis between time to failure and cause of failure in a competing risk setup. Since right censoring is often present, if not always, with failure time data, its practical utility would be enhanced if a right censored version were available. We obtain this as an easy application to our general U-statistic methodology.

In future, we plan to develop the U-statistics methods in presence of covariates even further. In particular, emphasis will be placed on the computation aspect of the DR approach

which inhibits its practical utility somewhat. See the recent paper by Carpenter *et al.* (2006) for a comparison between reweighting and imputation for a general missing data problem.

3.8 Appendix

Outline of asymptotic normality of $\hat{U}$

Using the Hoeffding's decomposition (Hoeffding, 1948) of $\hat{U}$ we get

$$\sqrt{n}(\tilde{U}-\theta)=\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{H}_1^c(T_i,\delta_i)+o_p(1), \tag{3.8.1}$$

where $\mathcal{H}_1^c=\mathcal{H}_1-\theta$, with $\mathcal{H}_1(t_1,\delta_1)=E(\mathcal{H}(T_1,\delta_1,T_2,\delta_2,\ldots,T_m,\delta_m)|T_1=t_1,\delta_1)$.

Note that $\mathcal{H}_1^c(t_1,\delta_1)=h_1(t_1)\delta_1/K(t_1-)-\theta$, where

$$h_1(t_1)=E(h(t_1,T_2^*,\ldots,T_m^*)|T_1^*=t_1).$$

Moreover (3.8.1) holds uniformly in the kernel $\mathcal{H}$ over a suitable collection which will enable us to replace K by its Kaplan Meier estimator $\hat{K}$ implying

$$\sqrt{n}(\hat{U}-\theta)=\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{h_1(T_i)\delta_i}{\hat{K}_c(T_i-)}-\theta\right\}+o_p(1). \tag{3.8.2}$$

Now express the RHS of (3.8.2) as

$$\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{h_1(T_i)\delta_i}{K_c(T_i-)}-\theta\right\}+\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1}{\hat{K}_c(T_i-)}-\frac{1}{K_c(T_i-)}\right\}h_1(T_i)\delta_i+o_p(1) \tag{3.8.3}$$

Note that $\sum_{i=1}^{n}U(T_i)\delta_i I(T_i\le\tau)=\int_0^\tau U(t)dN(t)$ for any bounded function $U(t)$ where $N(t)=\sum_{i=1}^{n}I(T_i\le t,\delta_i=1)$ is the counting process of failures in the right censored experiment.

Again, $\hat{K}_c(\cdot)=K_c(\cdot)+o_p(1)$ and $n^{-1}N(\cdot)=\bar{n}(\cdot)=P(T_i\le\cdot,\delta_i=1)+o_p(1)$. Using these, the RHS of (3.8.3) reduces to

$$\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{h_1(T_i)\delta_i}{K_c(T_i-)}-\theta\right\}-m\sqrt{n}\int\frac{h_1}{K_c^2(t-)}(\hat{K}_c(t-)-K_c(t-))d\bar{n}+o_p(1)$$

Next note $\sqrt{n}(\hat{K}_c(u-) - K_c(u-)) \approx \sqrt{n}(e^{-\hat{A}^c(u-)} - e^{-A^c(u-)})$, where A^c is the cumulative censoring hazard and $\hat{A}^c$ is its Nelson-Aalen estimator for censoring,

$$\approx -\sqrt{n}K_c(u-)(\hat{A}^c(u-) - A^c(u-)), \text{ by the delta method}$$

$$\approx -\frac{1}{\sqrt{n}}K_c(u-)\sum_{io=1}^{n}\int_0^{u-}\frac{dM_i^c(s)}{y(s)} \text{ (see Andersen } et\ al. \text{ (1993), page 178)}$$

where $M_i^c(t) = N_i^c(t) - \int_0^t \alpha^c(u)Y_i(u)du$ is the martingale of the censoring process defined with respect to the appropriate filtration, α^c is the censoring hazard, $N_i^c(t) = I(T_i \leq t, \delta_i = 0)$ is the counting process of the censored data, $Y_i(t) = I(T_i \geq t)$ be the appropriate 'at-risk' process and $y(t) = EY_i(t)$. Hence,

$$\begin{aligned}
& m\sqrt{n}\int \frac{h_1(u)}{K_c^2(u-)}(\hat{K}_c(u-) - K_c(u-))d\bar{n} \\
\approx \; - \; & \frac{m}{\sqrt{n}}\int_0^{\infty}\frac{h_1(u)}{K_c^2(u-)}K_c(u-)\sum_{i=1}^{n}\left\{\int_0^{u-}\frac{dM_i^c(s)}{y(s)}\right\}d\bar{n}(u) \\
= \; & -\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\infty}\frac{h_1(u)}{K_c(u-)}\left\{\int_0^{u-}\frac{dM_i^c(s)}{y(s)}\right\}d\bar{n}(u) \\
= \; & -\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\infty}\left\{\int_{s+}^{\infty}\frac{h_1(u)}{K_c(u-)}\frac{d\bar{n}(u)}{y(s)}\right\}dM_i^c(s) \text{ (by Fubini's theorem)} \\
= \; & -\frac{m}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\infty}w(s)dM_i^c(s)
\end{aligned}$$

where

$$w(s) = \frac{1}{y(s)}\int_{s+}^{\infty}\frac{h_1(u)}{K_c(u-)}d\bar{n}(u).$$

Combining both pieces we have the following asymptotic linear representation for $\hat{U}$ as:

$$\sqrt{n}(\hat{U} - \theta) = \frac{m}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{h_1(T_i)\delta_i}{K_c(T_i-)} + \int_0^{\infty}w(s)dM_i^c(s) - \theta\right\} + o_p(1).$$

Therefore, we have, as $n \to \infty$,

$$\sqrt{n}(\hat{U} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

where

$$\sigma^2 = m^2 Var\left(\frac{h_1(T_1)\delta_1}{K_c(T_1-)} + \int w(s)dM_1^c(s)\right).$$

Estimation of σ^2: Note that

$$\begin{aligned}
\int w(s)dM_i^c(s) &= w(T_i)\delta_i^c - \int \frac{w(s)I(T_i \geq s)}{Y(s)} dN^c(s) \\
&= w(T_i)\delta_i^c - \int \frac{w(s)I(T_i \geq s)}{Y(s)} d\left(\sum_{j=1}^n N_j^c(s)\right) \\
&= w(T_i)\delta_i^c - \sum_{j=1}^n \int \frac{w(T_j)I(T_i \geq T_j)}{Y(T_j)} dN_j^c(s) \\
&= w(T_i)\delta_i^c - \sum_{j=1}^n \frac{w(T_j)I(T_i \geq T_j)\delta_j^c}{Y(T_j)}
\end{aligned}$$

We define estimators of h_1 and w by

$$\hat{h}_1(u) = n^{-(m-1)} \sum_{1 \leq i_2, \ldots, i_m \leq n} h(u, T_{i_2}, \ldots, T_{i_m}) \frac{\delta_{i_2} \ldots \delta_{i_m}}{\hat{K}_c(T_{i_2}-) \ldots \hat{K}_c(T_{i_m}-)}$$

$$\begin{aligned}
\hat{w}(s) &= \frac{1}{Y(s)} \int_{s+}^{\infty} \frac{\hat{h}_1(u)}{\hat{K}_c(u-)} dN(u), \\
&= \frac{1}{Y(s)} \sum_{i=1}^n \frac{\hat{h}_1(T_i)\delta_i}{\hat{K}_c(T_i-)} I[T_i > s]
\end{aligned}$$

Hence we can estimate σ^2 by

$$\hat{\sigma}^2 = \frac{m^2}{n-1} \sum_{i=1}^n (V_i - \bar{V})^2,$$

where

$$V_i = \frac{\hat{h}_1(T_i)\delta_i}{\hat{K}_c(T_i-)} + \hat{w}(T_i)\delta_i^c - \sum_{j=1}^n \frac{\hat{w}(T_j)I(T_i \geq T_j)\delta_j^c}{Y(T_j)}.$$

3.9 References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics*, **6**, 701-726.

[2] Aalen, O.O. (1980). A model for nonparametric regression analysis of counting processes. In: Klonecki, W., Kozek, A., Rosiski, J. (Eds), *Lecture Notes on Mathematical Statistics and Probability*, Vol.2, Springer NY, 1-25.

[3] Aalen, O.O. (1989). A linear regression model for the analysis of lifetimes. *Statistics in Medicine*, **8**, 907-925.

[4] Akritas, M. B. (1986). Empirical processes associated with V-statistics and a class of estimators under random censoring, *Annals of Statistics*, **14**, 619-637.

[5] Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical Models based on Counting Processes*, Springer-Verlag, New York.

[6] Bagai, I., Deshpande, J. V., Kochar, S. (1989). Distribution free tests for stochastic ordering in the competing risk model, *Biometrika*, **76** , 775-781.

[7] Betensky, R. A. and Finkelstein, D. M. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine*, **18**, 3101-3109.

[8] Bickel, P. J. and Lehmann, E. L. (1979). Descriptive statistics for nonparametric models. IV. Spread, Contributions to Statistics (J. H'ajek Memorial Volume), 33-40 (ed. J. Jurecov'a). Academia, Prague.

[9] Bickel, P. J., Klaassen, A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.

[10] Bose, A. and Sen, A. (1999). The strong law of large numbers for Kaplan-Meier U-statistics, *Journal of Theoretical Probability*, **12**, 181-200.

[11] Bose, A. and Sen, A. (2002). Asymptotic distribution of the Kaplan-Meier U-statistics, *Journal of Multivariate Analysis*, **83**, 84-123.

[12] Brown, B. W., Hollander, M and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies, *Reliability and Biometry*, 327-354.

[13] Carpenter, J., Kenward, M. and Vansteelandt, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data, *Journal of the Royal Statistical Society*, Series A, in press.

[14] Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant Survival Data, *Journal of the American Statistical Association*, **72**, 27-36.

[15] Datta, S. (2005). Estimating the mean life time using right censored data, *Statistical Methodology*, **2**, 65-69.

[16] Datta, S. and Satten, G. A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under stage dependent censoring, *Biometrics*, **58**, 792-802.

[17] Dewan, I., Deshpande, J. V. and Kulathinal, S. B. (2004). On testing dependence between time to failure and cause of failure via conditional probabilities, *Scandinavian Journal of Statistics*, **31**, 79-91.

[18] Gijbels, I. and Veraverbeke, N. (1991). Almost sure asymptotic representation for a class of functionals of the Kaplan-Meier estimator, *Annals of Statistics*, **19**, 1457-1470.

[19] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Annals of Statistics*, **19**, 293-325.

[20] Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663-685.

[21] Jannsen, P. (1997). Bootstrapping U-statistics, *South African Statistical Journal*, **31**, 185-216.

[22] Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481, 562-563.

[23] Koul, H., Susarla,V., Van Ryzin, J. (1981). Regression analysis of randomly right-censored data, *Annals of Statistics*, **9**, 1276-1288.

[24] van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal data and Causalty*, Springer, New York.

[25] Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of the competing risk data, *Applied Statistics*, **34**, 201-211.

[26] Lee, A. J. (1990). *U-statistics*, Marcel Dekker, New York.

[27] Martin, E. C. and Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau, *Journal of the American Statistical Association*, **100**, 484-492.

[28] Oakes, D. (1982). A concordance test for independence in the presence of censoring, *Biometrics*, **38**, 451-455.

[29] Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*, Wiley, New York.

[30] Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, *AIDS Epidemiology - Methodological Issues*. Eds: Jewell, N., Dietz, K., Farewell, V. Boston, MA: Birkhäuser, 297-331.

[31] Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, *Proceedings of the Biopharmaceutical Section*, American Statistical Association. Alexandria, VA. 24-33.

[32] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 122-129.

[33] Robins, J.M., Rotnitzky, A.,and van der Laan, M.J. (1999). Discussion of 'On Profile Likelihood' by S.A.Murphy and A.W.van der Vaart, *Journal of the American Statistical Association*, **95**, 477-482.

[34] Rotnitzky, A. and Robins, J. (2003). Inverse probability weighted estimation in survival analysis, *Encyclopedia of Biostatistics*, 2nd Edition, Eds: Amitage, P. and Colton, T., Wiley and Sons, New York.

[35] Satten, G. A. and Datta, S. (2001). The Kaplan-Meier Estimator as an inverse-probability-of-censoring weighted average, *American Statistician*, **55**, 207-210.

[36] Satten, G. A., Datta, S. and Robins, J. M. (2001). An estimator for the survival function when data are subject to dependent censoring, *Statistics and Probability Letters*, **54**, 397-403.

[37] Satten, G. A. and Datta, S. (2004). Marginal Analyses of Multistage Data, *Handbook of Statistics*(N. Balakrishnan and C. R. Rao, Eds.), **23**, 559-574.

[38] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

[39] Stute, W. and Wang, J. L. (1993a). The strong law under random censorship, *Annals of Statistics*, **21**, 1591-1607.

[40] Stute, W. and Wang, J. L. (1993b). Multi-sample U-statistics for censored data, *Scandinavian Journal of Statsitics*, **20**, 369-374.

[41] Stute, W. (1995). The bias of Kaplan-Meier integrals, *Scandinavian Journal of Statistics*, **21**, 475-484.

[42] Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions, *Annals of Mathematical Statistics*, **18**, 309-348.

[43] Wang, W. and Wells, M. T. (2000). Estimation of Kendall's τ under censoring, *Statistica Sinica*, **10**, 1199-1215.

[44] Weier, D. R. and Basu, A. P. (1980). An investigation of Kendall's τ modified for censored data with applications, *Journal of Statistical Planning and Inference*, **4**, 381-390.

Chapter 4

Conclusions and Future Work

This dissertation considered some novel nonparametric approaches to inference in multistage event time data, specifically competing risk. A major contribution is the introduction of a right-censored version of the U-statistic (for a general kernel of size m) which provides substantial generalization over the Kaplan-Meier U-statistic of degree two proposed earlier by Bose and Sen (2002). We also address a unified approach of testing the null hypothesis of independence of time to failure and cause of failure (equivalently the equality of conditional survival functions) which has not received much significance in the nonparametric testing literature despite vast importance in studying the nature of competing risks (Dewan *et al.* 2004). We propose two different tests for studying the above null hypothesis, viz. (a) a family of weighted log-rank type tests based on 'fractional risk sets' and (b) a U-statistic based test with a Kendall's τ kernel incorporating right censoring through a mean preserving reweighing scheme. The former is based on the modification of the risk set through probabilistic arguments while the latter induces reweighting of the observed failure times with the survival function of the censored times through mean preservation. For (a), the test statistic leads to complicated asymptotics with often untractable variance which is estimated through bootstrap resampling. For (b), the asymptotic normality is handled through regular martingale techniques though asymptotics kicks in usually at a high sample size. In the case of dependent censoring, the test statistic might become inconsistent if the probability model of censoring is misspecified. This is taken care of through double-robustness methodology. Using fractional risk sets, we introduce a new Nelson-Aalen type estimator of the cumulative

conditional hazards. It is to be noted here that our testing methodology based on fractional risk sets doesn't confine itself to the realms of competing risks but is extendable to more complicated tree networks. While extending this test to a tree network, we also need to check the legitimacy of the Kaplan-Meier type estimator of the competing risk survival functions at every step.

We provide several extensions to our direction of research. A number of future research problems are proposed here. They will be pursued in near future.

4.1 Extension to weighted Kaplan-Meier type statistics

Since the familiar two-sample t-test cannot be generalized to a two sample censored data problem, a natural approach to extending this test is to standardize an estimator of the difference in means of survival functions of the two arms. This approach was first considered by Pepe and Fleming (1989, 1991) in the context of a two-arm clinical trial. We can extend their test to a bi-competing risk problem using the Satten and Datta (1999) representation of the Kaplan-Meier type estimators of competing risks estimates. Clearly, this test will incorporate censoring through modifications achieved by 'fractional risk sets'. More specifically, in the notation of Chapter 2, our weighted KM test statistic has the form

$$\Delta(\tau) = \int_0^\tau \hat{W}(t) \left\{ \hat{S}_1(t) - \hat{S}_2(t) \right\} dt \tag{4.1.1}$$

with

$$\hat{S}_j(t) = \prod_{t_i \leq t} \left(1 - \frac{\Delta N_j(t_i)}{Y_j^f(t)} \right) \tag{4.1.2}$$

where τ is some study duration, say the largest observed time of some event (say failure), and

$$n^{1/2}(\hat{S}_j(t) - \exp\{-\hat{\Lambda}_j(t)\}) \xrightarrow{P} 0 \tag{4.1.3}$$

for

$$\hat{\Lambda}_j(t) = \int_0^t \frac{\Delta N_j(u)}{Y_j^f(u)} \tag{4.1.4}$$

is the Nelson-Aalen type estimator of the integrated conditional hazard rate for the jth risk. Similarly, we have

$$n^{1/2}(S(t) - \exp\{-\Lambda(t)\}) \xrightarrow{P} 0 \tag{4.1.5}$$

where

$$\Lambda(t) = \int_0^t \alpha(u)du \tag{4.1.6}$$

and $\alpha(t)$ is the hazard function under the null hypothesis of equality of two conditional survival functions. Assume a predictable weight process $\hat{W}(t)$ such that

$$\sup_{u\in[0,t)} \left|\hat{W}(t) - w(t)\right| \xrightarrow{P} 0 \tag{4.1.7}$$

for some nonstochastic function $w(t)$.

Under the null hypothesis, we have

$$n^{-1/2}\Delta(\tau) = n^{-1/2}\int_0^\tau \hat{W}(t)\left\{\left(\hat{S}_1(t) - S(t)\right) - \left(\hat{S}_2(t) - S(t)\right)\right\}dt \tag{4.1.8}$$

$$= \underbrace{n^{-1/2}\int_0^\tau \hat{W}(t)(\hat{S}_1(t) - S(t))dt}_{(I)} - \underbrace{n^{-1/2}\int_0^\tau \hat{W}(t)(\hat{S}_2(t) - S(t))dt}_{(II)} \tag{4.1.9}$$

Now (I) implies

$$\approx n^{-1/2}\int_0^\tau \hat{W}(t)\{\exp(-\hat{\Lambda}_1(t)) - \exp(-\Lambda(t))\}dt \tag{4.1.10}$$

Using triangulations, we obtain

$$\exp(-\hat{\Lambda}_1(t)) - \exp(-\Lambda(t))$$

$$= -S(t)\left\{\int_0^t \frac{dN_1(u)}{Y_1(u)} - \int_0^t \alpha(u)du + \int_0^t \left\{\frac{1}{Y_1^f(u)} - \frac{1}{Y_1(u)}\right\} dN_1(u)\right. \tag{4.1.11}$$

and using the notation of the Appendix in Chapter 2, (4.1.10) reduces to

$$-S(t)\left\{\sum_{i=1}^n \int_0^t \frac{dM_{1,i}(u)}{Y_1(u)} + \int_0^t \left\{\frac{1}{Y_1^f(u)} - \frac{1}{Y_1(u)}\right\} dN_1(u)\right\} \tag{4.1.12}$$

We use similar triangulations for (II). Combining these and following the steps as described in Appendix of Chapter 2, we can obtain the linear representation of this test statistic and hence asymptotic normality is immediate. Once again, the expression would be very cumbersome and one needs to resort to resampling techniques to compute the standard error of this test. Note that this test statistic varies fundamentally from the usual two-sample log-rank type test statistic for censored data which is entirely rank based. Hence the power of this test statistic might not be sensitive to the magnitude of the difference of survival times.

We conducted a small simulation study to assess the finite sample performance of this test based on simulation scheme (ii) as described in Section 2.5.1 of Chapter 2. We choose a sample of size $n = 50$ with two different censoring schemes, light (about 25%) and heavy (about 50%) and use the empirical population variance to compute the test statistic. The results appear in Table 4.1. The value $a = 1$ corresponds to the null hypothesis and the other values of a denote departures from the null. The test is found to behave nicely at the 5% level with steadily increasing power for the alternatives. As expected, there is lower power for the heavy censoring than for the light censoring.

Table 4.1: Size/Power values for Kaplan-Meier based test under simulation scheme (ii)

		a					
ρ	Censoring	1	1.3	1.6	1.9	2.2	2.5
0.5	Light	0.052	0.326	0.880	0.975	0.998	1.000
	Heavy	0.061	0.265	0.684	0.903	0.947	0.982
0	Light	0.051	0.293	0.803	0.947	0.994	0.997
	Heavy	0.051	0.216	0.635	0.896	0.958	0.983
-0.5	Light	0.048	0.285	0.740	0.939	0.996	0.997
	Heavy	0.046	0.237	0.684	0.891	0.983	0.994

Further theoretical details of this statistic along with simulation results will be pursued later.

4.2 Generalization of the test in presence of covariates

It is often noticed that competing risk survival data appear with covariates, some of which may be potential confounders. In such cases, it is perhaps more appropriate to test the independence of failure time and cause (or equality of failure time distributions amongst groups) conditional on the covariates. For example, the data in Lagakos (1978) related to lung cancer clinical trial conducted by the Eastern Cooperative Oncology Group has three covariates: Z_1= performance status (ambulatory = 0, non-ambulatory = 1), Z_2 = treatment (A=0, B=1) and Z_3 = age in years. Thus, the covariates can be both categorical or quantitative. We will be interested in testing the null hypothesis

$$H_0 : S(t|X = m, \mathbf{Z}) = S(t|X = n, \mathbf{Z}),\ \forall\ 0 \leq t \leq \tau \text{ and } 1 \leq m \neq n \leq J, \tag{4.2.1}$$

where, as before $X \in \{1, \ldots, J\}$ denotes the subpopulation membership. In general, the cause X will be unknown for censored individuals; however the covariate information Z will be assumed to be known for everyone. If X were known, we could proceed semi-parametrically. We can model the marginal hazard (in presence of covariates) by the Cox proportional hazards model (Cox, 1972)

$$\lambda_i(t|\mathbf{Z_i}, X_i) = \lambda_0(t)e^{\beta^T \mathbf{Z_i} + \gamma . X_i + \nu^T Z_i X_i} \tag{4.2.2}$$

where $\lambda_0(t)$ denotes the baseline hazard. In this framework, the null hypothesis reduces to $H_{01} : \gamma = \mathbf{0};\ \nu = \mathbf{0}$ (considering ν to be of dimension k) which can be tested from the standard Cox methodology of partial likelihood. Consider the vector $\boldsymbol{\theta} = (\beta^T, \gamma, \boldsymbol{\nu}^T)^T$ and let $\hat{\theta}$ be its partial likelihood estimate. Then under the null, a Wald-type Chi-squared test can be proposed as

$$T = \hat{\theta}_{\mathbf{2}} \hat{\mathbf{\Sigma}}_{\mathbf{2}}^{-\mathbf{1}} \hat{\theta}_{\mathbf{2}} \tag{4.2.3}$$

where $\hat{\theta}_{\mathbf{2}} = (\gamma, \nu^T)^T$, and $\hat{\mathbf{\Sigma}}_{\mathbf{2}}$ is its estimated asymptotic variance-covariance matrix. It is expected that under the null hypothesis, T will have an asymptotically χ^2 distribution with $J-1$ degrees of freedom. It is to be noted however that this test procedure cannot be used if X is unknown for the censored individuals.

We would use the weighted estimating equation approach (Satten and Datta, 2004) to construct the appropriate estimating equations that correspond to Cox's partial likelihood if indeed the Z were available. The partial likelihood score equation for estimating the parameters θ can be written as (see, e.g. Andersen *et al.*, 1993)

$$S(\theta) = \sum_t \int \left\{ \boldsymbol{W}_i - \frac{A^*(t)}{B^*(t)} \right\} dN^i(t),$$

where $\boldsymbol{W}_i = (Z_i, X_i, Z_i, X_i)^T$, $N^i(t) = I(T_i \leq t,\, \delta_i = 0)$,

$$A^*(t) = \sum_{i=1}^{n} \boldsymbol{W}_i e^{\theta^T \boldsymbol{W}_i} I(T_i \geq t)$$

and

$$B^*(t) = \sum_{i=1}^{n} e^{\theta^T \boldsymbol{W}_i} I(T_i \geq t)$$

Since part of $\boldsymbol{W}_i$ is unavailable for the censored individuals, we modify A^* and B^* to

$$\hat{A}(t) = \sum_{i=1}^{n} \boldsymbol{W}_i e^{\theta^T \boldsymbol{W}_i} \frac{I(T_i \geq t, \delta > 0)}{\hat{K}_c(T_i-)}$$

and

$$\hat{B}(t) = \sum_{i=1}^{n} e^{\theta^T \boldsymbol{W}_i} \frac{I(T_i \geq t, \delta_i > 0)}{\hat{K}_c(T_i-)}$$

which are computable from the observed data, where $\hat{K}_c(u) = \exp\{-\hat{\Lambda}_c(u|\boldsymbol{W})\}$. Following earlier work by Satten and Datta (2004) we will advocate using Aalen's linear model for the censoring hazard $\lambda_c(.|\boldsymbol{W})$ which would yield the values for $\hat{K}_c$. More details on using Aalen's linear hazard model for estimating $\lambda_c(.|\boldsymbol{W})$ appears in Section 3.4.2. We will derive a martingale representation for the resulting approximate score function in terms of the martingales corresponding to both failure and censoring event counting processes following approaches as in Satten, Datta and Robins (2001) and Satten and Datta (2004).

4.3 Construction of tests for current status data

A great deal of recent interest focuses on nonparametric estimation based on current status data, a more severe form of interval censored data. We will consider the same testing problem with current status data. This type of incomplete data occurs when individuals are not monitored constantly. Current status data represents the status of individuals who are inspected only at a single random inspection time, i.e. a single snapshot per individual. In the context of current status data, the data for the ith individual is the pair $\{C_i, s(C_i)\}$, where C_i is the inspection time and $s(C_i)$ represents one of the possibilities: the person is still alive at time C_i or the person is dead due to cause j, say. Such data sets have been analyzed very recently by Jewell *et al.* (2003), Ding and Wang (2004), Datta and Sundaram (2006), etc.

Important research problems in this direction are (a) construction of a version of the log-rank test for testing equality of survival curves for various failure types that can be computed from current status data and (b) establishing asymptotic distributions of the test and conduct power study. In order to achieve this, we will replace the various risk processes $Y_j(t)$ and the counting processes $N_j(t)$ by their nonparametrically estimated conditional expectations given the current status data $\{C_i, s(C_i), 1 \leq i \leq n\}$. The following construction is adapted from Datta and Sundaram (2006).

Denote by U_j the (unobserved) transition time of an individual failing due to cause j ($= \infty$ if this individual fails due to a cause other than j). Let $N_j^*(t)$ denote the usual counting process counting the number of type j failures in $[0, t]$ with the complete data. By the law of large numbers, we have

$$n^{-1} N_j^*(t) \xrightarrow{P} P\{U_j \leq t\}$$

Consider the event that an individual has failed due to type j by time C denoted by $(U_j \leq C)$. Now, $I(U_j \leq C)$ is computable from the available current status information since $I(U_j \leq$

$C) = I\{s(C) = j\}$. Then, for any $t \geq 0$,

$$E(I(U_j \leq C)|C = t) = P\{U_j \leq t\}.$$

Now, we can estimate $nP\{U_{jj'} \leq t\}$ by a nonparametric kernel regression estimator

$$\hat{N}_j(t) = \sum_{i=1}^{n} \frac{I(U_{j,i} \leq C_i)K_h(C_i - t)}{\hat{g}_C(t)} \tag{4.3.1}$$

with

$$\hat{g}_C(t) = n^{-1} \sum_{i=1}^{n} K_h(C_i - t) \tag{4.3.2}$$

where K is a probability density (kernel) and $h = h(n) \to 0$ be a sequence of bandwidths and $K_h = h^{-1}K(./h)$. Since $P\{U_j \leq t\}$ is monotonic in t, $n^{-1}\hat{N}_j(.)$ can be constructed by isotonic regression of $I(U_j \leq t)$ on C, based on the pairs $(C_i, I(U_{j,i} \leq C_i))$ using PAV (pooled adjacent violators) algorithm (Barlow *et al.* 1972). Monotonocity is maintained through a combination of isotonic regression followed by kernel smoothing using a log-concave density (Mukerjee, 1988). Let $\hat{N}_j^P(.)$ be the estimate obtained by standard PAV extended to the set of non negative reals in the usual way

$$\hat{N}_j^P(t) = \begin{cases} N_j^P(C_{(1)}), & \text{if} \quad t \leq C_{(1)} \\ N_j^P(C_{(i)}), & \text{if} \quad C_{(i)} \leq t \leq C_{(i+1)}; 1 \leq i \leq n, \\ N_j^P(C_{(n)}), & \text{if} \quad t \geq C_{(n)} \end{cases} \tag{4.3.3}$$

Let $K > 0$ be a differentiable, log-concave density and $h = h(n) > 0 \downarrow 0$ be a bandwidth sequence. Our final estimator of $N_j(.)$ becomes

$$\hat{N}_j(t) = n^{-1} \sum_{i=1}^{n} \frac{\hat{N}_j^P(C_i)K_h(C_i - t)}{\hat{g}_C(t)} \tag{4.3.4}$$

where K_h and $\hat{g}_C$ (using the same K) are described before. Consistency of this class of regression estimators was established in Mukerjee (1988). Construction of the at-risk process is similar

$$\hat{Y}_j(t) = \sum_{i=1}^{n} \frac{I(s_i(C_i) = j)K_h(C_i - t)}{\hat{g}_C(t)}. \tag{4.3.5}$$

Once again, we will combine this with the PAV estimators. Similar to Chapter 2, we can now introduce a family of log-rank tests based on the above choices of $\hat{N}$ and $\hat{Y}$. This construction of a test statistic for current status data will once again involve an extremely complicated structure for the variance-covariance matrix of the test process. Due to the presence of smoothing, a smoothed version of resampling based approximation needs to be introduced for its estimation. The exact relationship between the two will emerge from a careful asymptotic analysis of the bootstrapped test process which will be part of the future research.

4.4 Linking gene expression profiles to survival times

The recent development of DNA microarray technology allows simultaneous measurements of the expression levels of thousands of genes and enhanced the impetus to explore the genetic basis of patients' clinical outcomes. Newer statistical methods are being developed for relating gene expression profiles to censored survival data such as time to cancer recurrence or death. From the statistical point of view, one challenge is that the time to cancer recurrence or death is often right censored because during the course of follow ups, some patients may still be cancer-free or alive and therefore techniques from survival analysis will be needed. Perhaps, a greater challenge for the microarray gene expression data sets is that the sample size of tissues or cell lines are usually very small compared to the number of genes from expression arrays.

Two approaches happen to be common in this context. In order to increase the sample size, one can first cluster tumor samples into several clusters based on gene expression patterns across many genes and use Kaplan-Meier curves or log-rank tests to test whether there is a difference in survival times among different tumor groups. But here the phenotype information is completely ignored during clustering resulting in loss of efficiency. Otherwise, one

can cluster genes first based on their expression on different samples and use the sample averages in a Cox model for survival outcome. Both methods are clustering dependent. Hastie *et al.* (2001) proposed a tree harvesting method where a stepwise selection procedure is used to select genes (or cluster of genes) that are related to the phenotypes using Cox proportional hazards. Nguyen and Rocke (2002a) used the 'partial least square' (PLS) technique for dimension reductions in the framework of Cox models by using residuals. They also used the partial least squares technique for tumor and multi-class cancer classifications using gene expression data (Nguyen and Rocke, 2002b; 2002c). Their method is limited to linear functions of gene expression levels. In addition, use of residuals for parameter estimation in the Cox model is not well established in survival analysis literature since there are many different ways of defining residuals and hence the justification of their PLS algorithm is questionable.

We plan to consider a comparative study of various regression models to predict survival using the gene expression values as covariates along with various regression procedures to handle the 'large p, small n' problems. Recently, Gui and Li (2005) considered extension of LASSO to Cox's regression model. Datta, Le-Rademacher and Datta (2005) considered AFT modeling using PLS and LASSO and compared the two methods under a variety of simulation settings.For details regarding LASSO, see Efron *et al.* (2004). Li and Li (2004) applied sliced inverse regression (SIR), a dimension reduction technique to the analysis of censored microarray survival data. So far, no one has carried out a comparison of two different regression models. Specifically, we plan to compare the predictive values of the Cox's model with LASSO versus AFT model with LASSO. In a simulation setting, the performances of these models will be studied both under the correct models as well as incorrect models of data generation. We also plan to undertake a comparative analysis of some publicly available microarray-survival data such as the Michigan Cancer Study data. (Beer *et al.* 2002) using data based cross validation approach.

4.5 References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, New York, Springer-Verlag.

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference under Order restrictions*, New York, Wiley.

[3] Beer, D.G. *et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, **8**, 816-824.

[4] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics*, **2**, 437-453.

[5] Cox, D.R. (1972). Regression models and life-tables (with discussion), *J. Roy. Stat. Soc.* Ser. B., **74**, 187-220.

[6] Datta, S. and Sundaram, R. (2006). Nonparametric Estimation of Stage Occupation Probabilities in a Multistage Model with Current Status Data, *Biometrics*, to appear.

[7] Datta, S., Le-Rademacher, J. and Datta, S. (2005). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO, *Biometrics*, revision submitted.

[8] Dewan, I., Deshpande, J.V., Kulathinal, S.B. (2004). On Testing Dependence between Time to Failure and Cause of Failure via Conditional Probabilities, *Scand. J. Statist.*, **31**, 79-91.

[9] Ding, A.A. and Wang, W. (2004). Testing independence for bivariate current status data, *Journal of the American Statistical Association*, **99**, 145-155.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression, *Annals of Statistics*, **32**, 407-451.

[11] Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, *Bioinformatics*, **21**, 3001-3008.

[12] Hastie, T., Tibshirani, R., Botstein, D. and Brown, P. (2001). Supervised harvesting of expression trees, *Genome Biology* 2, research0003.1.

[13] Jewell, N.P., van der Laan, M. and Henneman, T. (2003). Nonparametric estimation from current status data with competing risks, *Biometrika*, **90**, 183-197.

[14] Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data, *Bioinformatics*, **20**, 3406-3412.

[15] Mukerjee, H. (1988). Monotone nonparametric regression, *Annals of Statistics*, **16**, 741-750.

[16] Nguyen, D.V., Rocke, D.M. (2002a). Partial least squares proportional hazard regression for application to DNA microarray survival data, *Bioinformatics*, **18**, 1625-1632.

[17] Nguyen, D.V., Rocke, D.M. (2002b). Tumor classification by partial least squares using gene expression data, *Bioinformatics*, **18**, 39-50.

[18] Nguyen, D.V., Rocke, D.M. (2002c). Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, **18**, 1216-1226.

[19] Pepe, M. and Fleming, T. (1989). Weighted Kaplan-Meier statistics - A class of tests for censored survival data, *Biometrics*, **45**, 497-507.

[20] Pepe, M. and Fleming, T. (1991). Weighted Kaplan-Meier Statistics: Large Sample and Optimality considerations, *Journal of the Royal Statistical Society B*, **53**, 341-352.

[21] Satten, G.A., Datta, S. and Robins, J.M. (2001). An estimator for the survival funciton when data are subject to dependent censoring, *Statistics and Probability Letters*, **54**, 397-403.

[22] Satten, G.A. and Datta, S. (2004). Marginal Analyses of Multistage Data, In *Handbook of Statistics* (N. Balakrishnan and C.R. Rao, Eds.), Elsevier-North Holland, **23**, 559-574.