Estimating Spearman's Correlation with Bivariate Right-Censored Data.

By

Svetlana K. Eden

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

June 30, 2020

Nashville, Tennessee

Approved:

Ph.D. Committee Chair, Ph.D.Frank E. Harrell Jr.

Ph.D. Scientific Supervisor, Ph.D.Bryan E. Shepherd

Ph.D. Committee Internal Member, Ph.D.Qingxia Chen, Ph.D.

Ph.D. Committee Internal Member, Ph.D.Dandan Liu, Ph.D.

Ph.D. Committee External Member, Ph.D.Peter F. Rebeiro, Ph.D.

ACKNOWLEDGEMENTS

I am very grateful to all members of the Biostatistics Department for their encouragement and support. I want to thank my fellow students, whose optimism always lifted my spirit.

I appreciate all the hard work that our faculty and staff are doing to keep this graduate program welcoming and scientifically relevant. I cannot thank our professors enough - teaching for them is not just a job, but a calling. I feel fortunate to have worked with Dr. Michael Matheny, Dr. Bryan Shepherd, and Dr. Yaomin Xu as a Research Assistant. This work has supported me financially and allowed me to expand my statistical expertise significantly.

I am very grateful to my dissertation committee, Dr. Frank E. Harrell Jr., Dr. Bryan E. Shepherd, Dr. Qingxia Chen, Dr. Dandan Liu, and Dr. Peter F. Rebeiro, for their time and insightful feedback through the process of writing this dissertation. From my scientific advisor, Dr. Bryan E. Shepherd, I have been fortunate to learn about the value of giving one's full attention, the art of providing direction while also encouraging initiative, and the commitment to always looking at the bright side.

I feel fortunate to have wholehearted support from all of my friends, my parents, Liliya and Robert, my sister, Marina, and my sons, Michael and Ittai.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ρ_S	Spearman's rank correlation
$\rho_{S\|\Omega_R}$	Restricted region Spearman's correlation
ρ_S^H	Highest rank Spearman's correlation
ρ_{PSR}	Correlation of probability scale residuals
ρ_{IMI}	Estimator of Spearman's correlation proposed by Schemper et al. (2013)
A1C	Glycated hemoglobin
AIDS	Acquired immunodeficiency syndrome
ART	Antiretroviral therapy
CCASAnet	Caribbean, Central, and South America Network for HIV epidemiology
CD4	Cluster of differentiation 4, type of white blood cells
CDF	Cumulative distribution function
DRS	Diabetic Retinopathy Study
HIV	Human immunodeficiency virus
iid	Independent and identically distributed
IMI	Iterative multiple imputation
KM	Kaplan-Meier
MBI	Body mass index
PDF	Probability density function
PSA	Prostate-specific antigen
PSRs	Probability scale residuals
RMSE	Root mean squared error
SD	Standard deviation

CHAPTER 1

INTRODUCTION

According to the Merriam Webster dictionary, correlation is “a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone”. Some examples of correlated variables include education and income, physical activity and body mass index (BMI), depression and the risk of death. Measuring correlation is essential because it helps us learn and understand physical or socio-economic phenomena. The correlation between variables can also be affected by other factors, e.g., the correlation between physical activity and BMI may vary with age. Therefore, it is also of interest to measure adjusted and conditional correlations. The variables of interest, however, may not always be observed due to an upper detection limit or other events, e.g., lost to follow-up or end of study. Such a mechanism of missing data is called right-censoring. Right-censoring is very common in practice and presents a challenge when estimating correlation. Excluding censored observations may result in a biased estimate. Including censored observations requires special statistical methods.

Statistical methods can be divided into three groups: parametric, semi-parametric, and non-parametric. Ideally, researchers should be able to choose a clinically meaningful population parameter (e.g., the overall correlation, partial, or conditional correlation) independently from the estimation method. However, because of censoring, the choices of the population parameter and the estimation method may affect each other. For instance, with a parametric or semi-parametric estimation approach, it is possible to estimate the overall correlation. But with a purely non-parametric approach, it may no longer be possible. The problem gets even harder when the correlation depends on other variables, and estimation of the adjusted or conditional correlation is desired.

Parametric or semi-parametric methods work well for well-behaved distributions, but the real data often do not fit into known statistical structures. Therefore, non-parametric methods may be preferred. Specifically, Spearman’s rank correlation combines robust performance in the presence of outliers and invariance to monotone transformations with the familiar numeric scale of a well-known, but fully parametric, Pearson’s correlation. Despite the advantages of Spearman’s correlation, up until now, it has been estimated with parametric or semi-parametric methods that make assumptions about the correlation structure, which contradicts the non-parametric nature of Spearman's correlation. Existing estimation methods for adjusted Spearman's correlation also make assumptions about the correlation structure. To fill this gap, I propose several non-parametric methods to estimate the unadjusted rank correlation and semi-parametric methods to estimate partial and conditional rank correlations. All of these methods are based on Spearman's rank correlation; they address the issue of estimation in the presence of censoring without making assumptions about the correlation structure. In the following chapters, I study their performance and apply them to estimate unadjusted, partial, and conditional correlations between times to events. In Chapter 2, I present two non-parametric approaches to compute Spearman's correlation. To address the problem of end-of-study censoring, the first approach, restricted Spearman's correlation, $\rho_{S|\Omega_R}$, focuses only on the study period. The second approach, the highest rank Spearman's correlation, ρ_S^H , aims to approximate the overall Spearman's correlation by accounting for the probability mass of censored observations. Although asymptotically normal and consistent, estimates of $\rho_{S|\Omega_R}$ and ρ_S^H require a non-parametric estimator of the bivariate survival surface, which is difficult to fit in practice. These methods also rely on the bootstrap for computing confidence intervals and are not easily adaptable to estimate the adjusted or conditional correlation. To overcome these limitations, in Chapter 3, I propose a method that does not require estimating a bivariate survival surface and uses only marginal distributions. This method estimates the correlation of probability scale residuals (PSRs), ρ_{PSR}, which has been shown to equal Spearman's correlation when there is no censoring. Because ρ_{PSR} is computed only from marginal distributions, it tends to be less variable than $\rho_{S|\Omega_R}$ and ρ_S^H, although it is biased for Spearman's correlation when a high proportion of observations are censored. Confidence intervals can be constructed based on large sample approximations obtained through M-estimation. Moreover, ρ_{PSR} can be easily extended to partial (adjusted), conditional, and partial-conditional correlations using parametric or semi-parametric models for each time to event.

All estimators are illustrated in a study measuring the correlation between the time to viral failure and the time to regimen change among HIV-positive persons initiating antiretroviral therapy in Latin America. I implemented our methods as open-source statistical software and included them in the R package PResiduals. In Chapter 4, I demonstrate the use of this package by analyzing the correlation of time to retinopathy in treated and untreated eyes for patients with diabetes.

CHAPTER 2

NON-PARAMETRIC ESTIMATION OF SPEARMAN'S RANK CORRELATION WITH BIVARIATE SURVIVAL DATA

2.1 Introduction

In many medical studies researchers are interested in measuring the correlation between two time-to-event variables. For example, in studies of HIV/AIDS, there is interest in studying the correlation between the time from antiretroviral therapy (ART) initiation to viral failure and the time from ART initiation to regimen change. These variables should be highly correlated, and it might be important to know if the observed correlation is weaker than expected. Bivariate survival data may also come from paired subjects. For example, researchers might be interested in assessing the correlation between the time to cardiovascular disease for a patient and that for their parents, or between times to events in twins.

Right censoring is a defining element of bivariate survival data: one or both of the times to event may not be observed. When looking at correlation between times to events that occur in a single subject (e.g. time to viral failure and time to regimen change) the censoring time may be the same time for both outcomes. However, if the times to events are in paired subjects (e.g. times to events in twins) then the censoring times may differ. Our interest is in both scenarios, but we do not consider the setting where an event occurring in one variable causes censoring of the other, i.e. competing risks.

Several different methods have been proposed to measure and test the correlation between two right-censored time-to-event variables. Clayton (1978) introduced a bivariate hazard ratio or a *cross ratio* as a single number summary of correlation in the context of a frailty model. Oakes (1982, 1989) suggested a test for independence based on the cross ratio, showed its relationship with Kendall's tau, and extended its definition to a larger class of models. Fan et al. (2000) used a weighted average of the inverse of the cross ratio and a limited region Kendall's tau. Cuzick (1982) proposed a model of correlation and several test statistics, one of which resembles Spearman's rank correlation for censored data. Dabrowska (1986) derived generalized statistics to test the null hypothesis that the joint survival distribution is equal to the product of the marginals. Under certain assumptions, one of these statistics is related to a censored version of Spearman's correlation, and another corresponds to a log-rank test based on martingale residuals. Shih and Louis (1996) developed two additional statistics based on martingale residuals. Shih and Louis (1995) also suggested a two-stage estimation procedures to evaluate the correlation in bivariate data using copulas (Nelsen (2007)). Other approaches for measuring correlation using copulas for bivariate survival data have been considered by Carriere (2000), Romeo et al. (2006), Zhang (2008), and Schemper et al. (2013).

Spearman's rank correlation is ubiquitous in biomedical research because of its simple interpretation, robustness, and ability to capture non-linear correlations. It is used much more frequently in practice than Kendall's tau, perhaps because it closely approximates Pearson's correlation under normality (Kruskal, 1958), and it is much easier to compute and interpret than a cross ratio. In the absence of censoring, Spearman's correlation is simply the correlation of the ranked data. However, despite related work by Cuzick (1982), Dabrowska (1986), and Oakes (1989), there is no non-parametric estimator of Spearman's correlation for bivariate survival data. Schemper et al. (2013) proposed a semi-parametric iterative multiple imputation (IMI) method to estimate Spearman correlation (denoted as ρ_{IMI} throughout this paper). Their method transforms bivariate survival data into a Gaussian dependency structure using a normal copula, multiply imputes censored observations from this induced bivariate distribution, and approximates Spearman's correlation using Pearson's correlation of the normal deviates. This approach is semi-parametric because it does not require any assumptions about the marginal distributions. However, it uses a Gaussian dependency structure, which may lead to bias due to misspecification.

The goal of this paper is to propose and study non-parametric estimators of Spearman's correlation for right-censored data. Our methods use a non-parametric bivariate survival surface estimator. A challenge with estimating a bivariate survival surface, however, is that it may be non-parametrically estimable only within a certain region, for example due to end-of-study censoring. This motivates us to propose two correlation estimators: one that is defined only within the restricted region and another one that implicitly assigns values outside of this region as having the highest rank.

In Section 2.2 we express Spearman's correlation for time to event data and describe target parameters of interest. In Section 2.3 we address estimation and inference. In Section 2.4 we evaluate the performance of our estimators with several sets of simulations. In Section 2.5, we apply our methods to an HIV study by examining the correlation between times from treatment initiation to viral failure and regimen change. In Section 2.6, we discuss our methods and future directions. We have implemented our methods as part of the `PResiduals` R-package (see Dupont et al. (2018)).

2.2 Population Parameters

2.2.1 Notation and Definitions

We are interested in estimating correlation between two time-to-event variables denoted as (T_X, T_Y) defined on $[0, \infty) \times [0, \infty)$. Variables T_X and T_Y can be observed on a single subject or on a pair of subjects. Each time to event can be censored. We denote time-to-censoring variables as (C_X, C_Y) and assume independence between (T_X, T_Y) and (C_X, C_Y), but C_X and C_Y can be dependent. If T_X and T_Y are observed on a single subject then it is likely that $C_X = C_Y$. If $C_X = C_Y$ with probability one, then we call this *univariate* censoring, otherwise censoring is *bivariate*. In most studies, the follow-up period is bounded. We denote the maximum follow-up times respectively as τ_X and τ_Y, and consider these to be fixed by study design. Censoring due to the end of study has been referred to as type I censoring (Kalbfleisch and Prentice (2011)). With type I censoring, no events will be observed beyond the *restricted region* $\Omega = [0, \tau_X) \times [0, \tau_Y)$, or equivalently, $C_X \leq \tau_X$ and $C_Y \leq \tau_Y$. For our presentation, we distinguish between *strict* type I censoring, where censoring occurs only at τ_X and τ_Y (i.e., $C_X = \tau_X$, $C_Y = \tau_Y$), and *generalized* type I censoring, where censoring may also occur prior to τ_X and τ_Y (i.e., $C_X \leq \tau_X$, $C_Y \leq \tau_Y$). Strict type I censoring is rarely observed in practice, but will be helpful for explaining concepts; generalized type I censoring is quite common in practice, where follow-up time is bounded due to the length of the study, while subjects may start the study at different times or may drop out before the end of study. When the follow-up time is unbounded ($\tau_X = \infty$ and $\tau_Y = \infty$), we refer to the censoring as *unbounded.*

As a result of censoring, we only observe $X = \min(T_X, C_X)$ and $Y = \min(T_Y, C_Y)$ and event indicators $\Delta_X = \mathbb{1}(T_X \leq C_X)$ and $\Delta_Y = \mathbb{1}(T_Y \leq C_Y)$. We denote marginal and joint cumulative distribution functions of T_X and T_Y as $F_X(x) = \Pr(T_X \leq x)$, $F_Y(y) = \Pr(T_Y \leq y)$, $F(x, y) = \Pr(T_X \leq x, T_Y \leq y)$, and marginal and joint survival functions as $S_X(x) = \Pr(T_X > x)$, $S_Y(y) = \Pr(T_Y > y)$, $S(x, y) = \Pr(T_X > x, T_Y > y)$. We define $F_X(x^-) = \lim_{t \uparrow x} F_X(t)$ and $F(x^-, y) = \lim_{t \uparrow x} F(t, y)$; functions $F_Y(y^-)$ and $F(x, y^-)$ are defined similarly.

2.2.2 Spearman's Rank Correlation

As shown by Liu et al. (2018), in the absence of censoring, the population parameter for Spearman's correlation between T_X and T_Y can be defined as:

$$\rho_S = \text{Cor}\left\{ \frac{F_X(T_X) + F_X(T_X^-)}{2}, \; \frac{F_Y(T_Y) + F_Y(T_Y^-)}{2} \right\}. \tag{2.1}$$

When both T_X and T_Y are continuous the above definition translates into a better known expression:

$$\rho_S = \text{Cor}\left\{F_X(T_X), \; F_Y(T_Y)\right\}, \tag{2.2}$$

the grade correlation (Kruskal (1958)). Notice that if F_X and F_Y are estimated with their respective empirical distributions, then $F_X(T_X)$ and $F_Y(T_Y)$ are simply estimated as the ranks of T_X and T_Y respectively, divided by the number of points in the sample, corresponding to the well-known Spearman correlation estimator:

$$\widehat{\rho}_S = \text{Cor}\left\{\text{rank}(T_{X,i}), \; \text{rank}(T_{Y,i})\right\}, \tag{2.3}$$

where $(T_{X,i}, T_{Y,i})$ for $i = 1, ..., n$ are independent and identically distributed (iid) draws of (T_X, T_Y). Liu et al. (2018) have shown that equation (2.1) can be presented as:

$$\begin{aligned} \rho_S / c_\rho &= \text{E}_{T_X, T_Y}\left[\left\{F_X(T_X) + F_X(T_X^-) - 1\right\}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right] \\ &= \int_0^\infty \int_0^\infty \left\{F_X(x) + F_X(x^-) - 1\right\}\left\{F_Y(y) + F_Y(y^-) - 1\right\} F(dx, dy), \end{aligned} \tag{2.4}$$

where $c_\rho = \left[\text{Var}\left\{F_X(T_X) + F_X(T_X^-) - 1\right\}\text{Var}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right]^{-1/2}$; and $c_\rho = 3$ when T_X and T_Y are continuous. The right-hand side of (4.1) is the covariance of probability-scale residuals (PSRs) proposed and studied by Li and Shepherd (2012) and Shepherd et al. (2016) and defined as:

$$\begin{aligned} r_X(t_X, F_X) &= \text{E}\left\{\text{sign}(t_X - T_X)\right\} \\ &= \Pr\left(T_X < t_X\right) - \Pr\left(T_X > t_X\right) = F_X(t_X^-) + F_X(t_X) - 1, \end{aligned} \tag{2.5}$$

where $\text{sign}(t_X - T_X)$ is -1, 0, and 1 for $t_X < T_X$, $t_X = T_X$, and $t_X > T_X$ respectively. We can rewrite definition (4.1) in terms of survival functions:

$$\rho_S / c_\rho = \int_0^\infty \int_0^\infty \left\{1 - S_X(x) - S_X(x^-)\right\}\left\{1 - S_Y(y) - S_Y(y^-)\right\} S(dx, dy). \tag{2.6}$$

Right censoring causes serious challenges for non-parametric estimation of (4.3). First, non-parametric estimation of $S(x,y)$ is challenging due to non-unique solutions of the non-parametric likelihood in the presence of censoring and the fact that even consistent estimators of $S(x,y)$ may have negative mass for some x and y (Dabrowska (1988), Pruitt (1991), and Kalbfleisch and Prentice (2011)). Second, non-parametric estimation of $S_X(x)$, $S_Y(y)$, and $S(x,y)$ beyond the maximum follow-up time(s) is not possible. To overcome the latter challenge, one could focus on estimating Spearman's correlation in a restricted region Ω. Another possible approach is to focus on Spearman's correlation for an altered but estimable joint distribution. Population parameters for these two non-parametric approaches are presented in the next two subsections. Section 2.2.5 contains some examples.

2.2.3 Spearman's Rank Correlation in a Restricted Region

Suppose a researcher is interested in a rank correlation only inside a restricted region, denoted Ω_R. This correlation can be computed as Spearman's correlation defined conditionally on Ω_R, which we denote as $\rho_{S|\Omega_R}$. Typically, $\rho_{S|\Omega_R}$ will be different than the overall rank correlation, ρ_S, and will usually vary based on the choice of Ω_R. With failure time data, others have proposed and advocated the use of estimators in restricted regions including restricted mean survival times (Royston and Parmar (2013)) and limited region Kendall's tau (Fan et al. (2000)).

A natural choice is $\Omega_R = \Omega = [0,\tau_X) \times [0,\tau_Y)$, to estimate the restricted rank correlation over the region for which estimation is possible; to avoid introducing new notation, in this section we will use $\Omega_R = \Omega$. However, investigators can vary Ω_R depending on their research question as long as $\Omega_R \subseteq \Omega$; presumably Ω_R would generally be a rectangle that includes the origin $(0,0)$. The probability of double failure happening in this rectangle is

$$P_R = \Pr\left(x < \tau_X,\ y < \tau_Y\right) = F(\tau_X^-, \tau_Y^-) = 1 - S_X(\tau_X^-) - S_Y(\tau_Y^-) + S(\tau_X^-, \tau_Y^-). \quad (2.7)$$

We consider the conditional distribution over Ω_R. An example of a conditional distribution is illustrated in the middle panel of Figure 2.1. Its probability mass function is

$$S(dx, dy|\Omega_R) = S(dx, dy)/P_R, \quad (2.8)$$

and its marginal survival function on the X-axis is

$$S_X(x|\Omega_R) = 1 - F_X(x|\Omega_R) = \begin{cases} 0 & x \geq \tau_X \\ 1 - \frac{F(x,\tau_Y^-)}{P_R} & x < \tau_X \end{cases}, \tag{2.9}$$

where $F(x,\tau_Y^-) = 1 - S_X(x) - S_Y(\tau_Y^-) + S(x,\tau_Y^-)$. The marginal survival function on the Y-axis, $S_Y(y|\Omega_R)$, is similarly defined. Spearman's correlation in the restricted region is

$$\rho_{S|\Omega_R}/c_{\rho|\Omega_R} = \iint\limits_{\Omega_R} \{1 - S_X(x|\Omega_R) - S_X(x^-|\Omega_R)\}\{1 - S_Y(y|\Omega_R) - S_Y(y^-|\Omega_R)\}\, S(dx,dy|\Omega_R), \tag{2.10}$$

where

$$c_{\rho|\Omega_R} = \Big[\mathrm{Var}\Big\{1 - S_X(x|\Omega_R) - S_X(x^-|\Omega_R\Big\}\,\mathrm{Var}\Big\{1 - S_Y(y|\Omega_R) - S_Y(y^-|\Omega_R)\Big\}\Big]^{-1/2}.$$

Note that the population parameter, $\rho_{S|\Omega_R}$, depends only on τ_X and τ_Y and is invariant to the censoring distribution within Ω_R.

2.2.4 Spearman's Rank Correlation with Highest Ranks

Suppose now that a researcher is interested in the overall rank correlation, but the observations are only available within a restricted region. This is a typical situation in studies with a bounded follow-up time or when measurements have an upper detection limit. In this case, $S_X(x)$, $S_Y(y)$, and $S(x,y)$ are only non-parametrically estimable inside the region $\Omega = [0,\tau_X) \times [0,\tau_Y)$, where $\tau_X < \infty$ and $\tau_Y < \infty$. Since we are interested in a rank correlation, one approach would be to define any observation censored at τ_X as receiving the highest rank value for T_X and any observation censored at τ_Y as receiving the highest rank value for T_Y, which is the same as setting $T_X = \min(T_X,\tau_X)$ and $T_Y = \min(T_Y,\tau_Y)$. Such an approach is sensible because these censored observations at τ_X and τ_Y do have the highest rank values and there is no information to distinguish between these highest rank values without parametric modeling assumptions. Mathematically, this approach replaces $S(dx,dy)$ with a probability mass function $S^H(dx,dy)$ that is $S(dx,dy)$ inside Ω and the left-over

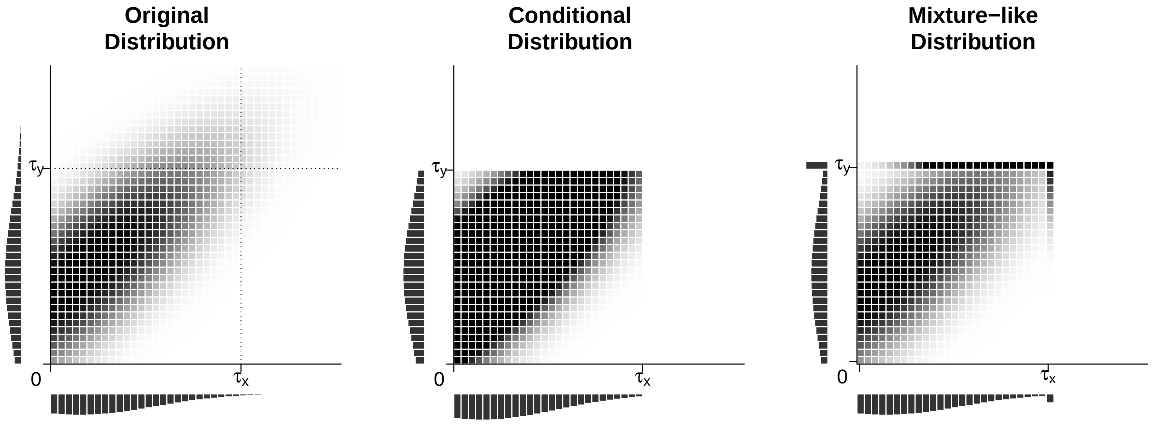

Figure 2.1: Illustration of bivariate distributions underlying the three population parameters. Left: Original distribution over $[0, \infty) \times [0, \infty)$, which has Spearman's correlation ρ_S. Middle: Conditional distribution over $\Omega_R = \Omega$, which has Spearman's correlation $\rho_{S|\Omega_R}$. Right: Mixture-like distribution S^H over region $\Omega \cup [0, \tau_X] \times \tau_Y \cup \tau_X \times [0, \tau_Y]$, which has Spearman's correlation ρ_S^H.

probability mass outside of Ω. That is:

$$S^H(dx, dy) = \begin{cases} S(dx, dy) & x < \tau_X \text{ and } y < \tau_Y, \\ S(\tau_X^-, dy) & x = \tau_X \text{ and } y < \tau_Y, \\ S(dx, \tau_Y^-) & x < \tau_X \text{ and } y = \tau_Y, \\ S(\tau_X^-, \tau_Y^-) & x = \tau_X \text{ and } y = \tau_Y, \\ 0 & x > \tau_X \quad \text{or } y > \tau_Y. \end{cases} \tag{2.11}$$

The new probability mass function $S^H(dx, dy)$ is depicted in the right panel of Figure 2.1. The part of $S^H(dx, dy)$ inside Ω is the same as $S(dx, dy)$ but its part outside of Ω is concentrated on the borders of Ω and at point (τ_X, τ_Y). The corresponding population parameter for the rank correlation of this new distribution is ρ_S^H, which satisfies:

$$\rho_S^H / c_\rho^H = \int_0^{\tau_X} \int_0^{\tau_Y} \left\{1 - S_X^H(x) - S_X^H(x^-)\right\} \left\{1 - S_Y^H(y) - S_Y^H(y^-)\right\} S^H(dx, dy), \tag{2.12}$$

where $S_X^H(x)$ and $S_Y^H(y)$ are the marginal survival functions of $S^H(x, y)$, and

$$c_\rho^H = \left[\text{Var}\left\{1 - S_X^H(T_X) - S_X^H(T_X^-)\right\} \text{Var}\left\{1 - S_Y^H(T_Y) - S_Y^H(T_Y^-)\right\}\right]^{-1/2}, \tag{2.13}$$

In other words, ρ_S^H is Spearman's correlation computed by setting $T_X = \min(T_X, \tau_X)$ and $T_Y = \min(T_Y, \tau_Y)$. Note that the population parameter, ρ_S^H, depends only on τ_X and τ_Y and is invariant to the censoring distribution within Ω_R.

For practical applications and intuitively, ρ_S^H can be viewed as the rank correlation computed for data with an upper detection limit, where all values above the detection limit are set to a common largest value. Note that although in general, $\rho_S^H \neq \rho_S$ (see examples in Section 2.2.5), unlike $\rho_{S|\Omega_R}$, parameter ρ_S^H is designed to take into account all observations including those outside of Ω_R. When the majority of the probability mass is within the restricted region, ρ_S^H can be viewed as an approximation of ρ_S. When $\tau_X = \infty$ and $\tau_Y = \infty$, then $\rho_S^H = \rho_S$.

2.2.5 A Few Examples

Here, we illustrate how the restricted region can affect ρ_S^H and $\rho_{S|\Omega_R}$. Having a restricted region implies type I censoring. Although the figures in this section are generated with strict type I censoring, ρ_S^H and $\rho_{S|\Omega_R}$ are the same for generalized type I censoring because, as we have mentioned earlier, ρ_S^H and $\rho_{S|\Omega_R}$ depend only on τ_X and τ_Y.

In some settings, ρ_S, $\rho_{S|\Omega_R}$, and ρ_S^H will be quite similar. For example, the values of these parameters for the distribution shown in Figure 2.1 are 0.635, 0.549, and 0.634, respectively. However, these parameters may be very different in some settings.

Figure 2.2 shows $\rho_{S|\Omega_R}$ (left panel) and ρ_S^H (right panel) as a function of ρ_S for different τ_X and τ_Y for Frank's copula family. In this example, ρ_S and ρ_S^H are very similar. In contrast, ρ_S and $\rho_{S|\Omega_R}$ are very different, especially when ρ_S is negative and Ω_R is small (e.g., region defined by 0 to the 0.5 quantiles). This is because in these settings, only a small fraction of the underlying distribution is inside Ω_R, and therefore $\rho_{S|\Omega_R}$ shows a weak negative correlation. Figure 2.3 contains three additional examples; some of these distributions may not be realistic in practice, but are useful for illustrative purposes. The left panel shows an X-like distribution, for which $\rho_S = 0$, $\rho_{S|\Omega_R} = 0.69$, and $\rho_S^H = 0.06$. Here, ρ_S^H is similar to ρ_S because ρ_S^H incorporates the probability mass in the upper left, upper right, and lower right regions. However, $\rho_{S|\Omega_R}$ is very different from ρ_S because Ω_R only contains a positively correlated subset of the distribution. The middle panel of Figure 2.3 shows a distribution with highly correlated values in Ω_R, zero correlation in the upper right region, and no mass in the upper left and lower right regions. For this distribution, $\rho_S = 0.66$, $\rho_{S|\Omega_R} = 0.64$, and $\rho_S^H = 0.99$. Here, $\rho_{S|\Omega_R}$ and ρ_S are similar because only the points in

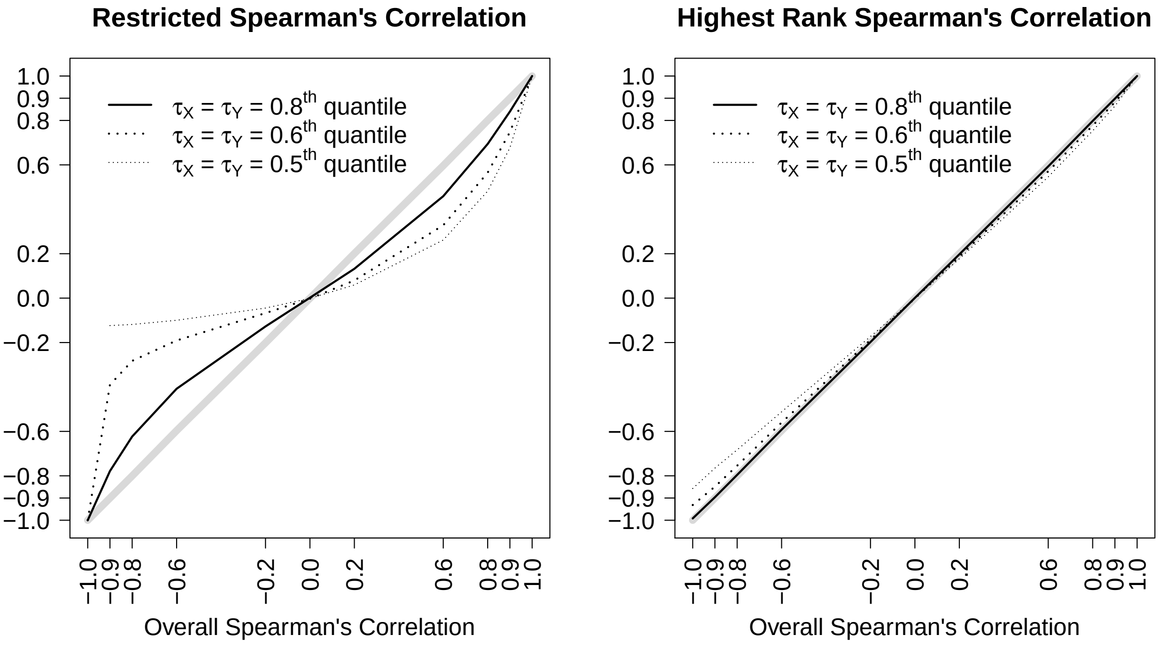

Figure 2.2: Restricted Spearman's correlation, $\rho_{S|\Omega_R}$ (left panel) and highest rank Spearman's correlation, ρ_S^H (right panel) for Frank's copula family for different restricted regions defined by τ_X and τ_Y ($\tau_X = \tau_Y$): 0.5^{th} (50% censored), 0.6^{th} (40% censored), 0.8^{th} quantiles (20% censored). A diagonal grey line is added for reference. Although the plots are generated based on data under strict type I censoring, the population parameters are the same for generalized type I censoring and are invariant to the rate of censoring within the restricted region.

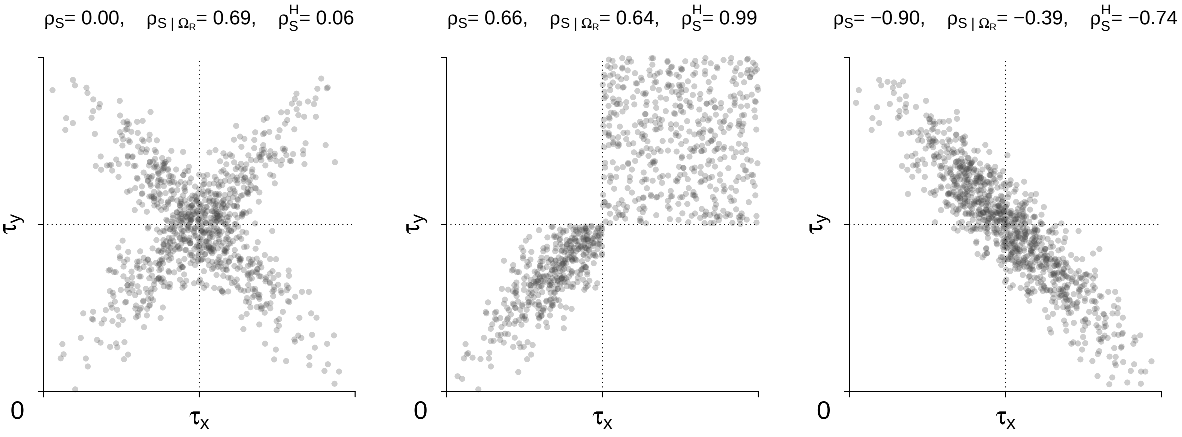

Figure 2.3: Example of bivariate distributions and their population parameters ρ_S with no censoring and ρ_S^H, and $\rho_{S|\Omega_R}$ with strict type I censoring with $\Omega_R = [0, \tau_X) \times [0, \tau_Y)$. The proportions of observed double events in the left, middle, and right panels are 25%, 43%, and 7% respectively. Drawn are 1000 points randomly selected from the underlying distributions. Although the plots are based on strict type I censoring, the population parameters are the same for generalized type I censoring and are invariant to the rate of censoring within Ω_R.

the restricted region are correlated. In contrast, ρ_S^H is quite high because the large probability mass of the upper right region is concentrated on a single highest-rank point when computing ρ_S^H, which pulls its value upwards considerably. The right panel of Figure 2.3 shows a distribution with a highly negative overall correlation, for which $\rho_S = -0.90$, $\rho_{S|\Omega_R} = -0.39$, and $\rho_S^H = -0.74$. Here, ρ_S^H and ρ_S are fairly different because although the probability mass of censored observations outside of Ω_R are taken into account, there is some loss of information with the highest rank assignment. The parameter $\rho_{S|\Omega_R}$ is very different from ρ_S because only a small fraction of the negatively correlated mass is included in Ω_R. Note that parametric or semi-parametric approaches also struggle with many of these settings because they effectively use the information from the restricted region to impute what is occurring outside the restricted region. For example, $\widehat{\rho}_{IMI}$ (see Schemper et al. (2013)) for the middle panel is approximately 0.95.

2.3 Non-parametric Estimation

2.3.1 Estimation of $\rho_{S|\Omega_R}$ under Generalized Type I Censoring

We estimate $\rho_{S|\Omega_R}$ using a plug-in estimator for equation (4.2):

$$\widehat{\rho}_{S|\Omega_R}/\widehat{c}_{\rho|\Omega_R} = \sum_{i:x_i<\tau_X} \sum_{j:y_j<\tau_Y} \Big[\Big\{1 - \widehat{S}_X(x_i|\Omega_R) - \widehat{S}_X(x_i^-|\Omega_R)\Big\} \\ \times \Big\{1 - \widehat{S}_Y(y_j|\Omega_R) - \widehat{S}_Y(y_j^-|\Omega_R)\Big\} \widehat{S}(dx_i, dy_j|\Omega_R)\Big], \tag{2.14}$$

where

$$\widehat{c}_{\rho|\Omega_R} = \left(\widehat{\mathrm{Var}}_X \cdot \widehat{\mathrm{Var}}_Y\right)^{-1/2}, \tag{2.15}$$

$$\widehat{\mathrm{Var}}_X = \sum_{i:x_i<\tau_X} \left(\Big[\Big\{1 - \widehat{S}_X\left(x_i|\Omega_R\right) - \widehat{S}_X\left(x_i^-|\Omega_R\right)\Big\} \\ - \sum_{i:x_i<\tau_X} \Big\{1 - \widehat{S}_X\left(x_i|\Omega_R\right) - \widehat{S}_X\left(x_i^-|\Omega_R\right)\Big\} \widehat{S}_X\left(dx_i|\Omega_R\right)\Big]^2 \widehat{S}_X\left(dx_i|\Omega_R\right)\right),$$

and $\widehat{\mathrm{Var}}_Y$ is computed similarly. The conditional survival curves, $\widehat{S}_X(x|\Omega_R)$, $\widehat{S}_Y(y|\Omega_R)$, and $\widehat{S}(x, y|\Omega_R)$ are estimated using plug-in estimators for (2.7), (2.8), and (2.9).

For $\widehat{S}(dx, dy)$ we use the estimator of Dabrowska (1988). The marginal distributions of Dabrowska's estimator, $\widehat{S}(x)$ and $\widehat{S}(y)$, are Kaplan–Meier estimators. There are other choices for non-parametrically estimating $S(dx, dy)$, including the estimators proposed by Prentice and Cai (1992), van der Laan (1997), Campbell (1981), and Lin and Ying (1993), to name a few. We considered the estimators of Campbell (1981) and Lin and Ying (1993) because of their computational simplicity, but ultimately chose Dabrowska's estimator because it is consistent for $S(dx, dy)$, straightforward to compute, and tended to result in estimates of Spearman's correlation with better performance (see Section 2.4). The confidence interval of $\widehat{\rho}_{S|\Omega_R}$ is estimated using the bootstrap.

The consistency of $\widehat{\rho}_{S|\Omega_R}$ for $\rho_{S|\Omega_R}$ follows from the continuous mapping theorem and the fact that it is a function of a consistent survival surface estimator (see Dabrowska (1988)). When $\rho_{S|\Omega_R} \in (-1, 1)$, estimator $\widehat{\rho}_{S|\Omega_R}$ is also asymptotically normal. Briefly, $\widehat{\rho}_{S|\Omega_R}$ is a function of $\widehat{S}(x, y)$, $\widehat{S}_X(x)$, and $\widehat{S}_Y(y)$, which are Hadamard differentiable estimators that converge to Gaussian processes (see Dabrowska (1989), van der Vaart and Wellner (1996), Gill et al. (1995), and van der Vaart (2000)). It follows from the chain rule (van der Vaart (2000)) that $\widehat{\rho}_{S|\Omega_R}$ is also Hadamard differentiable, and therefore from the functional delta method (van der Vaart and Wellner,

1996) that $\widehat{\rho}_{S|\Omega_R}$ is asymptotically normal. In addition, the bootstrapped estimator obtained by computing $\widehat{\rho}_{S|\Omega_R}$ from re-sampled data converges to the same Gaussian process as $\widehat{\rho}_{S|\Omega_R}$, justifying the use of the bootstrap to construct confidence intervals (van der Vaart and Wellner, 1996).

A few problems can arise in practice when computing $\widehat{\rho}_{S|\Omega_R}$, all due to negative mass at some points in $\widehat{S}(dx, dy)$. First, in some extreme cases (e.g. small sample sizes, strong positive or negative correlation, and heavy censoring), $|\widehat{\rho}_{S|\Omega_R}|$ may exceed 1; if this happens, we correct it by setting it as $\text{sign}(\widehat{\rho}_{S|\Omega_R})$. Second, negative mass can also lead to problems computing $\widehat{c}_{\rho|\Omega_R}$ in (2.15) because $\widehat{\text{Var}}_X$ or $\widehat{\text{Var}}_Y$ may be negative because $\widehat{S}_X\,(dx|\Omega_R)$ or $\widehat{S}_Y\,(dy|\Omega_R)$ is negative at some points. If this happens, we correct the guilty conditional marginal probability mass estimator by assigning negative values to 0, and by normalizing the rest of the probability mass values; specifically, the corrected probability mass is $\widehat{S}^*_X(dx|\Omega_R) = \max\left\{0, \widehat{S}_X(dx|\Omega_R)\right\} / \sum_x \max\left\{0, \widehat{S}_X(dx|\Omega_R)\right\}$. Third, negative mass may lead to a negative estimate of P_R, the probability of both events occurring in Ω_R (see equation (2.7)); when this occurs, $\widehat{\rho}_{S|\Omega_R}$ is not defined.

Although the number of points with negative mass does not decrease as the sample size increases (Pruitt, 1991), the amount of negative mass at each point does go to zero, which therefore reduces the possibility of these problems occurring. Also, the tendency of having negative mass is lower when a lower proportion of observations are singly or doubly censored. To give a sense of the magnitude of these problems, for a sample size of 50 with 70% bivariate censoring, $\widehat{\text{Var}}_X$ or $\widehat{\text{Var}}_Y$ was negative for 1.3% of 1000 simulations, and $\widehat{P}_R$ was less than zero in 2.4% of simulations. With a sample size of 100 and 70% bivariate censoring, these problems occurred in 0% and 0.8% of simulations, respectively.

2.3.2 Estimation of ρ_S^H under Generalized Type I Censoring

Equation (2.12) provides a straightforward way of estimating ρ_S^H, given a nonparametric estimate of the bivariate survival surface, $\widehat{S}(dx, dy)$:

$$\widehat{\rho}_S^H / \widehat{c}_\rho^H = \sum_{i^*} \sum_{j^*} \left\{1 - \widehat{S}_X^H(x_{i^*}) - \widehat{S}_X^H(x_{i^*}^-)\right\} \left\{1 - \widehat{S}_Y^H(y_{j^*}) - \widehat{S}_Y^H(y_{j^*}^-)\right\} \widehat{S}^H(dx_{i^*}, dy_{j^*}), \tag{2.16}$$

where i^* enumerates all the events for X plus τ_X, j^* enumerates all the event for Y plus τ_Y, and $\widehat{S}^H(dx_i, dy_j)$ and $\widehat{c}_\rho^H$ are the plug-in estimators for (2.11) and (2.13), respectively. As before, we compute $\widehat{S}(x, y)$ using Dabrowska's estimator. The confidence interval of $\widehat{\rho}_S^H$ is estimated using the bootstrap. Following arguments similar to those given in Section 2.3.1, the estimator $\widehat{\rho}_S^H$ is consistent and asymptotically normal for $\rho_S^H \in (-1, 1)$. In practice, for some extreme cases similar to those mentioned in Section 2.3.1 for $\widehat{\rho}_{S|\Omega_R}$, $|\widehat{\rho}_S^H|$ may exceed 1; when this occurs we correct it to sign($\widehat{\rho}_S^H$).

2.3.3 Estimation of ρ_S under Unbounded Censoring

The estimator $\widehat{\rho}_S^H$ may also be used to estimate ρ_S with unbounded censoring. Under unbounded censoring, $\tau_X = \tau_Y = \infty$ by definition, and one would naturally estimate $\widehat{\rho}_S$ by plugging $\widehat{S}(x, y), \widehat{S}_X(x)$, and $\widehat{S}_Y(y)$ into (4.3) in a manner similar to that described above. However, if the maximum value of X, for example, is a censored event (i.e., $\Delta_X = 0$) then $\widehat{S}(dx, dy)$ will not sum to 1 resulting in improper integration when using plug-in estimators in (4.3). A workaround is to assign the remaining mass (which is typically very little) to a point just beyond the largest observed event time of X and set τ_X to this point. We then estimate $\widehat{\rho}_S$ with $\widehat{\rho}_S^H$. Although in this setting, τ_X is no longer fixed but unbounded, estimation of $\widehat{\rho}_S$ in this manner seems to perform well (see Section 2.4). Under unbounded censoring, $\widehat{\rho}_S^H$ is consistent for ρ_S (see Appendix 2.A) and asymptotically normal for $\rho_S \in (-1, 1)$ following arguments similar to those earlier in this section.

2.4 Simulations

2.4.1 Simulation Set-up

We performed several simulations to investigate the finite sample performance of our estimators. The random variables T_X and T_Y were simulated using various choices of copula families and parameters. Specifically, following Fan et al. (2000), we simulated dependent random uniform variables, U and V, from Clayton's and Frank's copula families; both copulas are defined by a single parameter, θ. The dependence between U and V was specified by choosing the parameter θ in such a way that the true Spearman's correlation varied among no correlation ($\rho_S = 0$), moderate correlation ($\rho_S = -0.2$ and 0.2 for Frank's family and 0.2 for Clayton's family), and strong correlation ($\rho_S = -0.6$ and 0.6 for Frank's family and 0.6 for Clayton's family). (Clayton's family does not permit negative correlation.) We then set $T_X = -\log(1-U)$ and $T_Y = -\log(1-V)$ such that T_X and T_Y were exponentially distributed with mean 1.

Four types of censoring scenarios were implemented: 1) unbounded univariate,

2) unbounded bivariate, 3) generalized type I univariate, and 4) generalized type I bivariate. Each censoring scenario was implemented for censoring proportions $P_C = \{0.3, 0.7\}$. Bivariate unbounded censoring times C_X and C_Y for each observation were simulated independently from an exponential distribution, $Pr(C_X \leq t) = Pr(C_Y \leq t) = 1 - e^{-\lambda t}$, with $\lambda = P_C/(1-P_C)$. The event times T_X and T_Y were censored if $T_X > C_X$ and $T_Y > C_Y$, respectively. Univariate unbounded censoring was implemented in a similar manner except only one censoring event was generated per (T_X, T_Y) pair. For generalized type I censoring, C_X^* and C_Y^*, were first simulated as designed above with probability P_C and then C_X and C_Y were defined as $\min(C_X^*, \tau_X)$ and $\min(C_Y^*, \tau_Y)$, respectively, with $\tau_X = \tau_Y$ set at the median survival time, $S_X^{-1}(0.5) = S_Y^{-1}(0.5)$. The resulting censoring proportions for generalized type I censoring were therefore higher than P_C: for example, for generalized type I bivariate censoring with $P_C = 0.3$ and 0.7, the outcomes were censored for approximately 56% and 73% of observations, respectively.

We evaluated the performance of $\widehat{\rho}_S^H$ and $\widehat{\rho}_{S|\Omega_R}$ in the presence of unbounded and generalized type I censoring using a sample size of 200 for strong, moderate, and no correlation. For unbounded censoring, the population parameter of ρ_S^H is the same as ρ_S. For generalized type I censoring, the population parameters of ρ_S^H and $\rho_{S|\Omega_R}$ were the same as ρ_S for Clayton's family. For Frank's family with the overall Spearman's correlation of -0.6, -0.2, 0.2, and 0.6, the population parameters of $\rho_{S|\Omega_R}$ were -0.098, -0.042, 0.058, and 0.261; and the population parameters of ρ_S^H were -0.512, -0.173, 0.180, and 0.545 respectively. These population parameters were empirically estimated with a sample size of 10^6.

The bias, root mean squared error (RMSE), type I error rate, and power, computed as the proportion of times that bootstrap confidence intervals (based on 1000 bootstrap samples) did not include zero, were also evaluated for sample sizes of 100 and 200 under unbounded censoring for moderate and no correlation. The performance of $\widehat{\rho}_S^H$ was compared to estimator $\widehat{\rho}_{IMI}$ proposed by Schemper et al. (2013) for these simulation scenarios. We also evaluated the performance of $\widehat{\rho}_S^H$ as an estimator of ρ_S using survival surfaces proposed by Lin and Ying (1993) and Campbell (1981).

Lastly, we compared the performance of $\widehat{\rho}_S^H$ to semi-parametric estimators $\widehat{\rho}_{IMI}$ and $\widehat{\rho}_S^{MLE}$ (maximum likelihood estimator assuming Frank's copula dependency structure). These comparisons were made in the context of a well-behaved dependency structure induced by Frank's copula and in the context of a complex dependency structure, a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall

Spearman's correlation being about -0.0813. This simulation scenario was loosely motivated by data from the Mexican site in our real data analysis presented in Section 2.5; Figure 2.10 in Appendix 2.A shows the scatter plot of the uncensored data. Samples sizes of 200, 500, and 1000 were used and unbounded univariate censoring with $P_C = 0.5$ was applied as described above. The goal of these comparisons was to show better efficiency of $\widehat{\rho}_S^{MLE}$ compared to $\widehat{\rho}_S^H$ under a correctly specified model and to demonstrate greater accuracy of $\widehat{\rho}_S^H$ compared to $\widehat{\rho}_{IMI}$ and $\widehat{\rho}_S^{MLE}$ under model misspecification.

All simulations used 1000 replications. All analyses were performed in statistical language `R` (R Core Team (2017)) and using R-libraries `SurvCorr` (Ploner et al. (2015)), `lcopula` (Belzile and Genest (2017)), and `cubature` (Narasimhan and Johnson (2017)). Example code is posted online in the Supporting Information. Complete simulation and analysis code is posted at http://biostat.mc.vanderbilt.edu/ArchivedAnalyses.

2.4.2 Simulation Results

Figure 2.4 shows the mean point estimates and the 0.025^{th} and 0.975^{th} quantiles of estimators $\widehat{\rho}_S^H$ under unbounded censoring (row 1), the semi-parametric estimator proposed by Schemper et al. (2013), $\widehat{\rho}_{IMI}$, under unbounded censoring (row 2), $\widehat{\rho}_S^H$ under generalized type-I censoring (row 3), and $\widehat{\rho}_{S|\Omega_R}$ under generalized type-I censoring (row 4). The sample size was 200, the censoring was bivariate with varying censoring proportions and $\widehat{\rho}_S^H$ and $\widehat{\rho}_{S|\Omega_R}$ were computed using Dabrowska's survival surface estimator.

For Clayton's and Frank's families under unbounded censoring (row 1), the mean of $\widehat{\rho}_S^H$ was very close to the true population parameter verifying the consistency of $\widehat{\rho}_S^H$ for ρ_S. When data were generated under Frank's copula, the semi-parametric estimator, $\widehat{\rho}_{IMI}$ (row 2), was similarly unbiased for unbounded censoring, and it tended to be less variable than $\widehat{\rho}_S^H$. However, when data were generated using Clayton's copula, $\widehat{\rho}_{IMI}$ was biased for ρ_S (also noted by Schemper et al. (2013)).

Tables 2.2 and 2.3 in Appendix 2.A provide more details and additional comparisons between $\widehat{\rho}_{IMI}$ and $\widehat{\rho}_S^H$ under unbounded censoring for different sample sizes and censoring proportions in terms of bias, RMSE, type I error rate and power. In short, the bias of $\widehat{\rho}_S^H$ for ρ_S was low, even with fairly small numbers of events; both the bias and RMSE decreased as the number of events increased. In general, $\widehat{\rho}_S^H$ compared favorably to $\widehat{\rho}_{IMI}$.

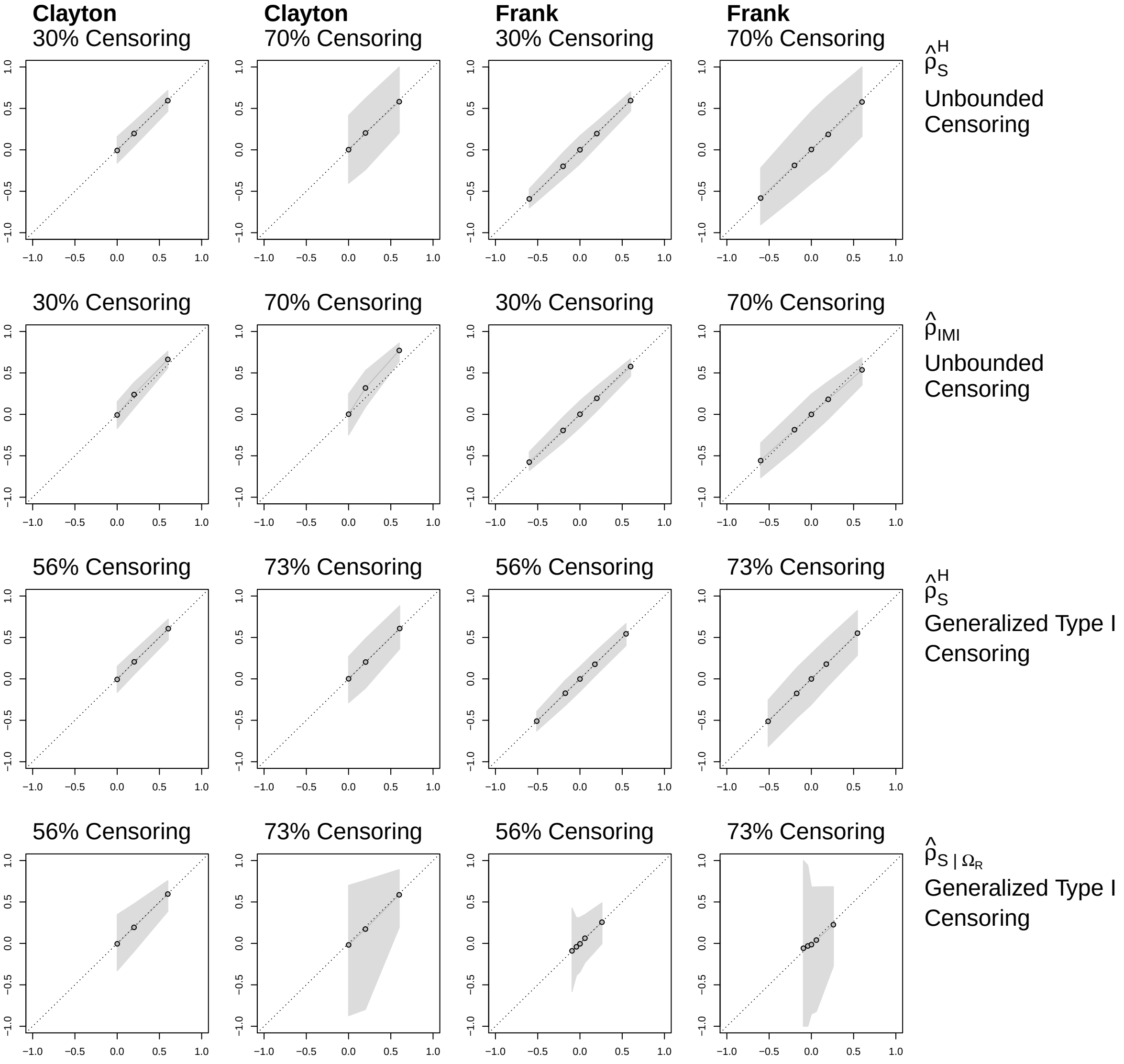

Figure 2.4: Point estimates (X-axis) vs population parameters (Y-axis) under different bivariate censoring scenarios. The top and second rows are $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ as estimators of the overall Spearman's correlation, ρ_S. The third row is $\widehat{\rho}_S^H$ as an estimator of ρ_S^H. The bottom row is $\widehat{\rho}_{S|\Omega_R}$ as an estimator of $\rho_{S|\Omega_R}$. The columns represent Clayton's and Frank's copulas. The population parameters for Clayton's family are 0, 0.2, and 0.6 for all estimates. For Frank's family, the population parameters of ρ_S are -0.6, -0.2, 0.2, and 0.6; the population parameters of ρ_S^H are -0.512, -0.173, 0.180, and 0.545; the population parameters of $\rho_{S|\Omega_R}$ are -0.098, -0.042, 0.058, and 0.261. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles. For generalized type I censoring, the restricted region, Ω_R, was defined by the median survival times.

Under generalized type I censoring (the third and fourth rows of Figure 2.4), the means of $\widehat{\rho}_S^H$ and $\widehat{\rho}_{S|\Omega_R}$ are very close to their true population parameters, suggesting with $n = 200$, these estimators are essentially unbiased for ρ_S^H and $\rho_{S|\Omega_R}$ respectively. The variance of our estimators naturally increased as the probability of censoring increased. The variance of $\widehat{\rho}_{S|\Omega_R}$ was greater than that of $\widehat{\rho}_S^H$ under generalized type I censoring (rows 3 and 4), presumably because $\widehat{\rho}_{S|\Omega_R}$ only uses the events inside $\Omega = \Omega_R$, whereas $\widehat{\rho}_S^H$ assigns probability mass to values outside of Ω. Note that the variance of $\widehat{\rho}_S^H$ under unbounded censoring was slightly larger than that under generalized type I censoring in spite of the lighter censoring. This is probably because under generalized type I censoring, the probability mass, $\widehat{S}^H(dx, dy)$, calculated outside of Ω is concentrated on the same points, making its variance smaller compared to the case of unbounded censoring. Results for univariate censoring were very similar to those for bivariate censoring, except the estimators were slightly less variable (see Figure 2.6 in Appendix 2.A).

With unbounded censoring of 50% and $n = 200$, the correctly specified semi-parametric $\widehat{\rho}_S^{MLE}$ was more efficient than $\widehat{\rho}_S^H$ (relative efficiency in terms of variance ranging from 1.19 (for $\rho_S = 0$) to 1.60 (for $\rho_S = 0.6$), Figure 2.9 in Appendix 2.A); both approaches yielded unbiased estimates of ρ_S. In contrast, when data were generated using a mixture of positively and negatively correlated bivariate distributions, the misspecified semi-parametric estimators $\widehat{\rho}_S^{MLE}$ and $\widehat{\rho}_{IMI}$ were substantially biased and this bias did not decrease with increasing sample size. On the other hand, the non-parametric $\widehat{\rho}_S^H$ was unbiased for ρ_S (see Figure 2.11 in Appendix 2.A). In this more complicated setting, estimates of $\rho_{S|\Omega_R}$ with Ω_R being defined using the median survival times were also unbiased (Figure 2.12 in Appendix 2.A).

The simulations reported above incorporated 1000 bootstrap replications; in general, confidence interval coverage and width were stable and adequate with as few as 200 bootstrap replications (see Figure 2.13 in Appendix 2.A).

We also evaluated the performance of $\widehat{\rho}_S^H$ with survival surfaces of Campbell (1981) for univariate and bivariate censoring and of Lin and Ying (1993) for univariate censoring only (see Figures 2.7 and 2.8 in Appendix 2.A). With univariate censoring, using the estimator of Lin and Ying (1993) resulted in unbiased estimation although with larger variance than that using Dabrowska's estimator. Estimator $\widehat{\rho}_S^H$ computed using the survival surface estimator of Campbell (1981) was visibly biased for the sample size of 200 under heavy censoring.

2.5 Application

We apply our methods to a study of 6691 HIV-positive adults starting ART in Latin America. We are interested in estimating the correlation between two right-censored variables: 1) the time from ART initiation to viral failure and 2) the time from ART initiation to major regimen change. Patients experience viral failure when the amount of virus circulating in their blood (their viral load) is above a certain threshold, which may make them infectious and vulnerable to HIV-related diseases. Viral failure may be caused by many factors including poor adherence or drug resistance; it often triggers changing a patient's ART regimen. However, patients may also change their ART regimens for reasons other than viral failure (e.g., poor tolerability, discovery of a simpler regimen, or patient/provider choice).

Our study uses data from the Caribbean, Central, and South America network for HIV Epidemiology (CCASAnet). The definitions of viral failure and regimen change were the same as those used in a prior CCASAnet study (Cesar et al., 2015). In short, viral failure was defined as a single viral load > 1000 copies/mL or two viral loads > 400 copies/mL after a person's virus had been suppressed or they had been on ART long enough that it should have been suppressed (i.e., 6 months). Regimen change was limited to major changes such that the patient switched drug classes or changed multiple drugs. Each study subject may have had one, both, or neither of these events. Follow-up ended at the last clinic visit; censoring was univariate. Our analysis dataset includes patients from Brazil, Chile, Honduras, Mexico, and Peru. After a median follow-up of 4.1 years (ranging from 1 day to 18.2 years), 1916 persons (28.6%) had a viral failure and 1895 persons (28.3%) changed regimens. Approximately 16.1% of patients had both events over the follow-up period, 12.2% changed regimens but did not have viral failure, 12.5% had viral failure but did not change regimens, and 59.1% were not observed to have either event. The upper left panel of Figure 2.5 shows Kaplan–Meier estimates for the marginal probabilities of viral failure and regimen change as a function of time since ART initiation. Ten years after ART initiation, the estimated probability of not having viral failure was 0.58 and the estimated probability of remaining on the initial regimen was 0.51. The upper right panel shows the estimated joint bivariate probability mass function, $\widehat{S}^H(dx, dy)$, based on Dabrowska's estimator; estimated marginal probability mass functions, $\widehat{S}^H_X(dx)$ and $\widehat{S}^H_Y(dy)$, are also included. Note that because a large proportion of patients experienced only one or neither event, a large amount of mass has been assigned to $\tau_X = 18$ and $\tau_Y = 17$ years.

The estimated highest rank correlation, $\widehat{\rho}^H_S$, between time to viral failure and

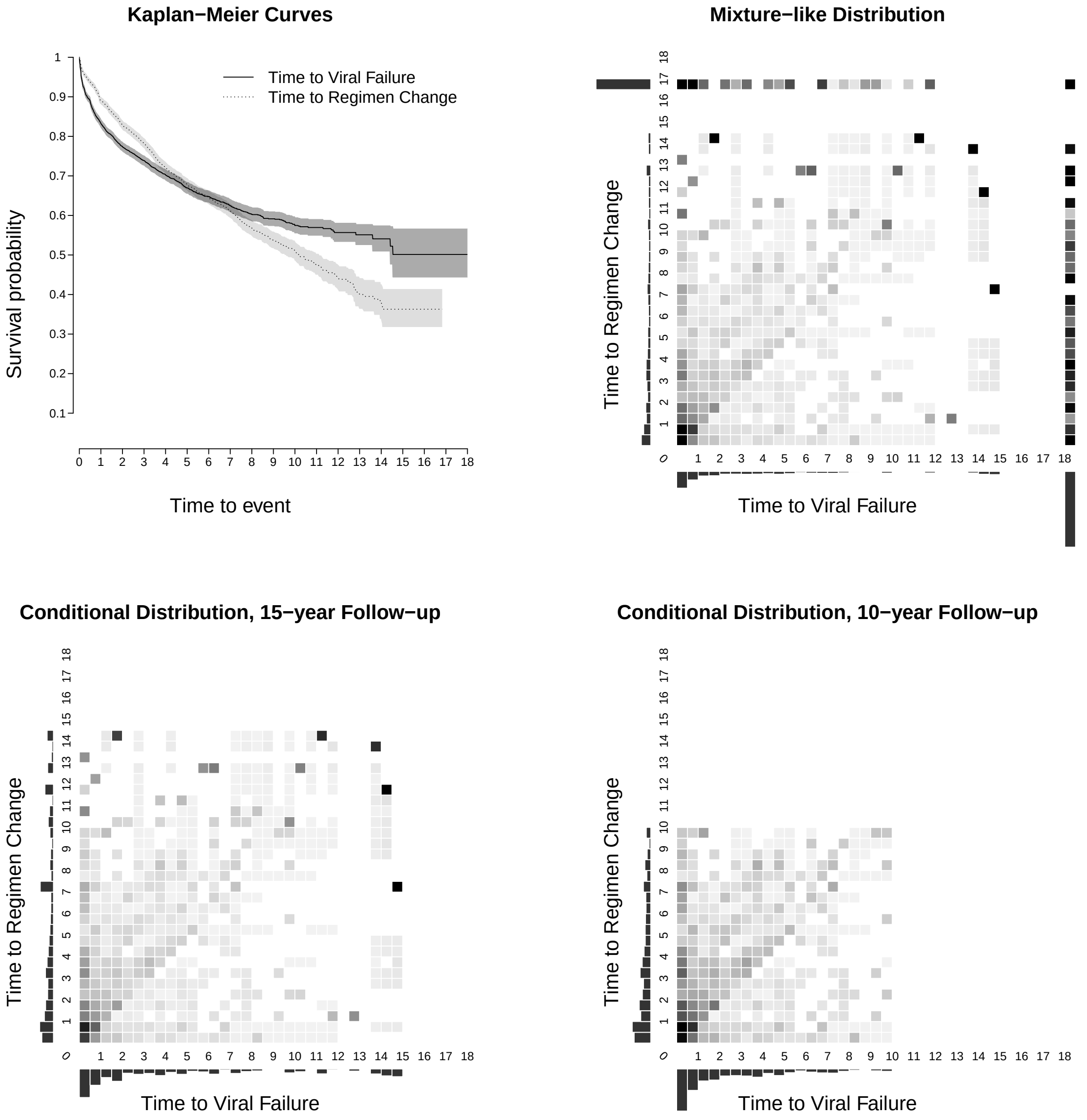

Figure 2.5: Upper left: Kaplan–Meier curves for time to viral failure and time to regimen change, where time is measured in years. Upper right: bivariate probability mass function for the mixture-like distribution, $\widehat{S}^H(dx, dy)$. Lower left: conditional bivariate probability mass function for 15-year follow-up. Lower right: conditional bivariate probability mass function for 10-year follow-up. For probability mass functions, the bars on the left and on the bottom represent histograms of the univariate survival mass for each event. The probability mass function was computed from the Dabrowska's survival surface and then aggregated over half-year bivariate time periods. After aggregation, any negative values were set to 0. Lighter shade represents smaller values.

time to regimen change was 0.35 with a 95% confidence interval (CI) based on 1000 bootstrap replications of 0.27 to 0.44. This result is fairly similar, albeit with a wider confidence interval, to the estimator of Schemper et al., $\widehat{\rho}_{IMI}$, that imputes censored values: 0.37, 95% CI 0.33 to 0.41.

The estimated rank correlation over the restricted region, $\widehat{\rho}_{S|\Omega_R}$, with $\Omega_R = [0, 15) \times [0, 15)$ was substantially higher, 0.65, with a wide 95% CI of 0.30 to 0.97. This can be explained by a careful look at the conditional probability mass over Ω_R (lower left plot of Figure 2.5). Notice that there are several points in the upper right corner of this surface with large probability mass. Large amounts of mass are assigned to these points because a few events occurred at later follow-up times when there were fairly small numbers of patients remaining in the risk set. This is also seen in the Kaplan–Meier estimates, where relatively large drops in the marginal survival curves are noted between 10 to 15 years. These points pushed the probability mass function closer to the diagonal, leading to a larger estimated rank correlation. In addition, since there were only a few points with substantial probability mass, their inclusion/exclusion in various bootstrap samples led to wide variation in confidence intervals. This example serves as a nice illustration of the potential perils of estimating rank correlations over restricted regions that include tail areas with small numbers of events. Perhaps a more reliable rank correlation would be over the restricted region, $\Omega_R = [0, 10) \times [0, 10)$, in which case $\widehat{\rho}_{S|\Omega_R}$ was 0.26 (95% CI [0.17, 0.36]); the lower right plot of Figure 2.5 shows the conditional probability mass over this smaller region.

In addition to showing estimates of overall correlation, Table 2.1 shows estimates based on sex and study site. For the most part, $\widehat{\rho}_S^H$ is fairly close to $\widehat{\rho}_{IMI}$, except for those sites with small sample sizes (i.e., Chile, Mexico, and Honduras). Rank correlations over the restricted region, $\widehat{\rho}_{S|\Omega_R}$ with $\Omega_R = [0, 15) \times [0, 15)$ were generally more variable and typically higher than those over $\Omega_R = [0, 10) \times [0, 10)$.

Subgroup	N	P_{VR}	$P_{\overline{V}R}$	$P_{V\overline{R}}$	$P_{\overline{VR}}$	$\widehat{\rho}_S^H$	$\widehat{\rho}_{S\vert\Omega_R=[0,15)\times[0,15)}$	$\widehat{\rho}_{S\vert\Omega_R=[0,10)\times[0,10)}$	$\widehat{\rho}_{IMI}$
All	6691	16.1	12.2	12.5	59.1	0.35 [0.27, 0.44]	0.65 [0.30, 0.97]	0.26 [0.17, 0.36]	0.37 [0.33, 0.41]
Male	5185	14.8	12.0	12.3	60.9	0.32 [0.22, 0.42]	0.57 [0.14, 1.00]	0.32 [0.18, 0.44]	0.36 [0.30, 0.41]
Female	1506	20.7	13.1	13.3	52.9	0.45 [0.30, 0.57]	0.70 [0.20, 1.00]	0.13 [-0.04, 0.32]	0.40 [0.32, 0.48]
Brazil	2313	22.4	11.8	13.8	51.9	0.45 [0.36, 0.53]	0.42 [0.10, 0.73]	0.17 [-0.01, 0.34]	0.36 [0.30, 0.42]
Chile	1040	19.2	23.0	11.5	46.2	-0.06 [-0.21, 0.30]	0.73 [-0.24, 1.00]	0.19 [0.03, 0.34]	0.26 [0.00, 0.37]
Honduras	138	18.1	17.4	10.1	54.3	-0.18 [-0.53, 0.44]	0.21 [-1.00, 1.00]	0.50 [-0.09, 0.96]	0.29 [0.00, 0.55]
Mexico	975	13.7	16.3	11.9	58.1	0.52 [0.14, 0.74]	-0.39 [-0.54, 0.41]	-0.43 [-1.00, 0.08]	0.25 [0.10, 0.38]
Peru	2225	8.9	5.5	12.1	73.4	0.45 [0.37, 0.53]	0.69 [0.42, 0.93]	0.58 [0.34, 0.80]	0.55 [0.43, 0.66]

Table 2.1: Correlation of time to viral failure and time to regimen change in CCASAnet cohort measured using three estimators: $\widehat{\rho}_S^H$, $\widehat{\rho}_{S|\Omega_R}$, and $\widehat{\rho}_{IMI}$. N is the number of subjects for each subgroup. Columns P_{VR}, $P_{\overline{V}R}$, $P_{V\overline{R}}$, and $P_{\overline{VR}}$ show the percent of subjects having both events (P_{VR}), having regimen change event but censored viral failure ($P_{\overline{V}R}$), having viral failure event but censored regimen change ($P_{V\overline{R}}$), and having both events censored ($P_{\overline{VR}}$). For $\widehat{\rho}_S^H$ and $\widehat{\rho}_{S|\Omega_R}$, the numbers inside brackets are 95% confidence intervals estimated from 1000 bootstrap samples. Column $\widehat{\rho}_{IMI}$ shows the estimates of Schemper et al. (2013) and their confidence intervals.

2.6 Discussion

We have proposed two non-parametric methods of quantifying correlation with bivariate right-censored data. One estimator, $\widehat{\rho}_{S|\Omega_R}$ computes Spearman's correlation within a restricted region. The other estimator, $\widehat{\rho}_S^H$, computes Spearman's correlation for an estimable bivariate distribution, which is analogous to assigning data censored beyond the estimable region to the highest rank values. Under unbounded censoring, $\widehat{\rho}_S^H$ is consistent for the overall Spearman's correlation. Under generalized type I censoring, with the majority of events happening in the restricted region, $\widehat{\rho}_S^H$ can be viewed an approximation of the overall Spearman's correlation. Because our methods assume neither marginal nor joint parametric distributions, they have potential advantages over parametric and semi-parametric methods.

The main limitations of our estimators stem from challenges with non-parametrically estimating the bivariate survival surface. The first challenge is that with generalized type I censoring, the bivariate survival surface, and hence Spearman's rank correlation ρ_S, cannot be identified beyond the region of data support without parametric assumptions. Hence, our inference targets were the estimable parameters, $\rho_{S|\Omega_R}$ and ρ_S^H. Although under generalized type I censoring they do not equal the overall Spearman's correlation, both parameters have sensible interpretations. The alternative, using parametric and semi-parametric models to estimate Spearman's correlation – an inherently non-parametric statistic – has other limitations. Parametric and semi-parametric approaches assume a dependency structure; and as seen in our simulations, semi-parametric estimators such as $\widehat{\rho}_{IMI}$ proposed by Schemper et al. (2013), may be biased for certain dependency structures. In addition, they implicitly assume that the dependency structure outside the region of observation is the same as that seen inside the observation region. Although in the real data example, $\widehat{\rho}_{IMI}$ appeared more stable than our non-parametric estimators, particularly for small sample sizes, with complex real data there is the real possibility of model misspecification. As with most statistics, there may be settings where one might prefer the non-parametric estimators over the parametric estimators, or vice versa.

The second challenge is that even in regions where the bivariate survival curve is identifiable, non-parametric estimators of the survival surface may have negative mass, which leads to potential downstream problems with estimation of $\rho_{S|\Omega_R}$ and ρ_S^H. Researchers have grappled with non-parametric approaches to avoid negative mass, e.g. van der Laan (1996). Thankfully, in our simulations only a minor proportion of our estimates encountered problems due to negative mass, the problems go away as the numbers of events increase, and there are typically workaround solutions that appear to behave reasonably.

In our approach we considered a rectangular restricted region, $[0, \tau_X] \times [0, \tau_Y]$. An anonymous associate editor correctly pointed out that the top right corner of this rectangle is not always identifiable non-parametrically. Alternatively, one could consider defining estimators using the identifiable region $\Omega_R = \{(t_X, t_Y) : T_{X,k} \geq t_X, T_{Y,k} \geq t_Y \text{ for some } k\}$ (see Prentice and Zhao (2019)). Although this idea is intriguing, we decided to keep the definition of Ω_R as $[0, \tau_X] \times [0, \tau_Y]$ due to easier interpretation. Note that Dabrowska's estimator can be computed and is consistent for all points inside $[0, \tau_X] \times [0, \tau_Y]$.

Although not studied here, our approaches can be directly applied to settings where only one of the variables is right-censored. This special case allows non-parametric estimation of the bivariate survival surface without negative mass (Stute, 1993). In future work, we plan to study semi-parametric methods for estimating covariate-adjusted partial and conditional Spearman's correlation for bivariate survival data.

Acknowledgements

This work was partially supported by the US National Institutes of Health, grants R01 AI093234 and U01 AI069923. We thank members of CCASAnet for allowing us to present their data, and Cathy Jenkins for help constructing the analysis dataset.

2.7 Appendix 2.A

2.7.1 Consistency of $\widehat{\rho}_S^H$ for ρ_S under unbounded censoring

In the following proofs, for the sake of brevity, we omit subscript X or Y when dealing with a marginal distribution of a single variable. We use notations a^+ and a^- to denote values infinitesimally above and below a, respectively, and notation $\xrightarrow{p}$ to denote convergence in probability.

Let T and C be time to event and time to censoring and $T \perp C$. Let $X = \min(T, C)$. Let $\Delta = I(T \leq C)$. Let $F(x) = \Pr(T \leq x)$ and $G(x) = \Pr(C \leq x)$ be cumulative distribution functions (CDFs) for T and C, respectively. Let the maximum possible value of T be t_{max} ($t_{max} < \infty$ or $t_{max} = \infty$). Unbounded censoring implies that the maximum possible value of C is at least t_{max}. Let random variable V represent an uncensored event,

$$V = \begin{cases} T, \quad T \leq C & \text{with probability } \int_{[0,t_{max}]} F(dx)\left\{1 - G(x^-)\right\}, \\ 0, \quad T > C & \text{with probability } \int_{[0,t_{max}]} G(dx)\left\{1 - F(x)\right\}. \end{cases} \tag{2.17}$$

Given n independent and identically distributed pairs (T_1, C_1), (T_2, C_2), ..., (T_n, C_n), let variables V_1, V_2, ..., V_n be defined accordingly and let $V_{max,n} = \max(V_1, V_2, ..., V_n)$. Lastly, let $\widehat{\tau}$ be $V^+_{max,n}$ if the last observed time is censored and $\widehat{\tau} = t_{max}$, otherwise.

Lemma 1.

We assume that if T is discrete and $t_{max} < \infty$, then $\Pr(T = t_{max}) > 0$. If $t_{max} = \infty$ then for any t_N there is such $t > t_N$ that $\Pr(T = t) > 0$. If T is continuous, we assume that there exists such $t_0 < t_{max}$ that for any $t \in [t_0, t_{max})$, $\Pr(t \leq T < t_{max}) > 0$. If T is a mixture of continuous and discrete random variables, then one of the conditions mentioned above should hold. Under unbounded censoring,

1. $V_{max,n} \xrightarrow{p} t_{max}$;
2. $\widehat{\tau} \xrightarrow{p} t_{max}$.

Proof:

Because of (2.17), for any $t < t_{max}$

$$\Pr(V < t) = \int_{[0,t_{max}]} G(dx)\left\{1 - F(x)\right\} + \int_{[0,t)} F(dx)\left\{1 - G(x^-)\right\}. \tag{2.18}$$

Because of (2.17), $\int_{[0,t_{max}]} G(dx)\left\{1 - F(x)\right\} = 1 - \int_{[0,t_{max}]} F(dx)\left\{1 - G(x^-)\right\}$, so (2.18) can be rewritten as

$$\Pr(V < t) = 1 - \int_{[t,t_{max}]} F(dx)\left\{1 - G(x^-)\right\}.$$

Note that $1 - G(x^-) > 0$ for all $x \in [0, t_{max}]$ because the maximum support value of C is at least t_{max}. Because of the assumptions, if T is discrete, then $F(dx) > 0$ for some $x \in [t, t_{max}]$, for example t_{max}. If T is continuous, there exists such t_0 that $F(dx) > 0$ for all $x \in [t_0, t_{max})$. Therefore, $\int_{[t,t_{max}]} F(dx)\left\{1 - G(x^-)\right\} > 0$ and thus $\Pr(V < t) = 1 - \int_{[t,t_{max}]} F(dx)\left\{1 - G(x^-)\right\} < 1$. Because $\Pr(V_{max,n} < t) = \left\{\Pr(V < t)\right\}^n$, we have

$$\Pr(V_{max,n} < t) \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{2.19}$$

Because (2.19) holds for any $t < t_{max}$, $V_{max,n} \xrightarrow{p} t_{max}$.

For any $t < t_{max}$,

$$\begin{aligned}\Pr(\widehat{\tau} < t) &= \Pr(V^{+}_{max,n} < t, \text{ and the largest time is censored}) \\ &\leq \Pr(V^{+}_{max,n} < t) \leq \Pr(V_{max,n} < t) \longrightarrow 0 \text{ as } n \longrightarrow \infty.\end{aligned}$$

Because this is true for any $t < t_{max}$, we have $\widehat{\tau} \xrightarrow{p} t_{max}$.

Theorem 1.
Let $\widehat{\tau}_X$ and $\widehat{\tau}_Y$ be defined as $\widehat{\tau}$ for T_X and T_Y respectively, and let $\widehat{\rho}^H_S$ and $\widehat{c}^H_\rho$ be defined as equations (12) and (13) in the main manuscript. Then $\widehat{\rho}^H_S \xrightarrow{P} \rho_S$.

Proof: We recall the expressions for ρ_S, ρ^H_S, $\widehat{\rho}^H_S$, c_ρ, c^H_ρ, and $\widehat{c}^H_\rho$:

$$\rho_S/c_\rho = \int_0^\infty \int_0^\infty \Big\{1 - S_X(x) - S_X(x^-)\Big\} \Big\{1 - S_Y(y) - S_Y(y^-)\Big\} S(dx, dy), \qquad (2.20)$$

$$\rho^H_S/c^H_\rho = \int_0^{\tau_X} \int_0^{\tau_Y} \Big\{1 - S^H_X(x) - S^H_X(x^-)\Big\} \Big\{1 - S^H_Y(y) - S^H_Y(y^-)\Big\} S^H(dx, dy),$$

$$\widehat{\rho}^H_S/\widehat{c}^H_\rho = \sum_{i^*} \sum_{j^*} \Big\{1 - \widehat{S}^H_X(x_{i^*}) - \widehat{S}^H_X(x^-_{i^*})\Big\} \Big\{1 - \widehat{S}^H_Y(y_{j^*}) - \widehat{S}^H_Y(y^-_{j^*})\Big\} \widehat{S}^H(dx_{i^*}, dy_{j^*}), \qquad (2.21)$$

where i^* enumerates all events for X plus τ_X and j^* enumerates all events for Y plus τ_Y and $S_X(x^-) = \lim_{t \uparrow x} S_X(t)$; $S^H_X(x)$ and $S^H_Y(y)$ are the marginal survival functions of $S^H(x, y)$ (defined in (11) in the main manuscript), and

$$\begin{aligned}c_\rho &= \Big[\text{Var}\Big\{1 - S_X(T_X) - S_X(T^-_X)\Big\} \text{Var}\Big\{1 - S_Y(T_Y) - S_Y(T^-_Y)\Big\}\Big]^{-1/2}, \\ c^H_\rho &= \Big[\text{Var}\Big\{1 - S^H_X(T_X) - S^H_X(T^-_X)\Big\} \text{Var}\Big\{1 - S^H_Y(T_Y) - S^H_Y(T^-_Y)\Big\}\Big]^{-1/2}, \\ \widehat{c}^H_\rho &= \Big[\text{Var}\Big\{1 - \widehat{S}^H_X(T_X) - \widehat{S}^H_X(T^-_X)\Big\} \text{Var}\Big\{1 - \widehat{S}^H_Y(T_Y) - \widehat{S}^H_Y(T^-_Y)\Big\}\Big]^{-1/2}.\end{aligned}$$

The righthand side of (2.21) can be written in the following way:

$$
\begin{aligned}
&\sum_{i^*}\sum_{j^*}\left\{1-\widehat{S}_X^H(x_{i^*})-\widehat{S}_X^H(x_{i^*}^-)\right\}\left\{1-\widehat{S}_Y^H(y_{j^*})-\widehat{S}_Y^H(y_{j^*}^-)\right\}\widehat{S}^H(dx_{i^*},dy_{j^*})\\
&=\underbrace{\sum_i\sum_j\left\{1-\widehat{S}_X(x_i)-\widehat{S}_X(x_i^-)\right\}\left\{1-\widehat{S}_Y(y_j)-\widehat{S}_Y(y_j^-)\right\}\widehat{S}(dx_i,dy_j)}_{\widehat{A}}\\
&\quad+\underbrace{\sum_i\left\{1-\widehat{S}_X(x_i)-\widehat{S}_X(x_i^-)\right\}\left\{1-\widehat{S}_Y(\widehat{\tau}_Y^-)\right\}\widehat{S}(dx_i,\widehat{\tau}_Y^-)}_{\widehat{B}}\\
&\quad+\underbrace{\sum_j\left\{1-\widehat{S}_X(\widehat{\tau}_X^-)\right\}\left\{1-\widehat{S}_Y(y_j)-\widehat{S}_Y(y_j^-)\right\}\widehat{S}(\widehat{\tau}_X^-,dy_j)}_{\widehat{C}}\\
&\quad+\underbrace{\left\{1-\widehat{S}_X(\widehat{\tau}_X^-)\right\}\left\{1-\widehat{S}_Y(\widehat{\tau}_Y^-)\right\}\widehat{S}(\widehat{\tau}_X^-,\widehat{\tau}_Y^-)}_{\widehat{D}},
\end{aligned}
\tag{2.22}
$$

where i enumerates all events for X and j enumerates all events for Y. The following two scenarios are possible:

1. Both maximum observed times are event times; $\widehat{S}^H(dx_i,dy_j)=\widehat{S}(dx_i,dy_j)$ and $\widehat{B}=\widehat{C}=\widehat{D}=0$.
2. The maximum observed time X and/or Y is censored; $\widehat{D}\neq 0$ and $\widehat{B}\neq 0$ or $\widehat{C}\neq 0$.

In the first scenario, the righthand side of (2.21) converges in probability to the righthand side of (4.3) because of the consistency of Dabrowska's estimator (Dabrowska, 1988) and continuous mapping theorem (van der Vaart and Wellner, 1996). The same reasoning applies to $\widehat{c}_\rho^H \xrightarrow{P} c_\rho$ and therefore $\widehat{\rho}_S^H \xrightarrow{P} \rho_S$.

In the second scenario, $\widehat{S}^H(x,y)\neq\widehat{S}(x,y)$. For any two values v_X and v_Y from the support of T_X and T_Y, respectively, let

$$
\begin{aligned}
\alpha(v_X,v_Y)&=\int_0^{v_X}\int_0^{v_Y}\left\{1-S_X(x)-S_X(x^-)\right\}\left\{1-S_Y(y)-S_Y(y^-)\right\}S(dx,dy),\\
\widehat{\alpha}(v_X,v_Y)&=\sum_{i:x_i\le v_X}\sum_{j:y_j\le v_Y}\left\{1-\widehat{S}_X(x_i)-\widehat{S}_X(x_i^-)\right\}\left\{1-\widehat{S}_Y(y_j)-\widehat{S}_Y(y_j^-)\right\}\widehat{S}(dx_i,dy_j).
\end{aligned}
$$

Note that for the reasons described above, $\widehat{\alpha}(v_X,v_Y)$ is a consistent estimator of $\alpha(v_X,v_Y)$, which in turn approaches the righthand side of (4.3) as $v_X\to t_{max,X}$ and $v_Y\to t_{max,Y}$. Also, for any realizations $V_{max,n,X}=v_X$ and $V_{max,n,Y}=v_Y$, we have

$\widehat{A} = \widehat{\alpha}(v_X, v_Y)$, which by the continuous mapping theorem makes $\widehat{A}$ a consistent estimator of $\alpha(v_X, v_Y)$ since $V_{max,n,X} \to t_{max,X}$ and $V_{max,n,Y} \to t_{max,Y}$ as $n \to \infty$. Therefore, $\widehat{A}$ is also a consistent estimator for the righthand side of (4.3).

From *Lemma 1*, it also follows that $\widehat{\tau}_X \xrightarrow{P} t^+_{X,max}$ and $\widehat{\tau}_Y \xrightarrow{P} t^+_{Y,max}$, therefore $\widehat{S}_X(\widehat{\tau}_X^-) \xrightarrow{P} 0$, $\widehat{S}_Y(\widehat{\tau}_Y^-) \xrightarrow{P} 0$, and $\widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{P} 0$, which leads to the following:

$$\begin{aligned}
B &= \sum_i \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\}\left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\}\widehat{S}(dx_i, \widehat{\tau}_Y^-) \le \sum_i \widehat{S}(dx_i, \widehat{\tau}_Y^-) = \widehat{\Pr}\{X < \widehat{\tau}_X, Y \ge \widehat{\tau}_Y\} \\
&= S_Y(\widehat{\tau}_Y^-) - S(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{P} 0, \\
C &= \sum_j \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\}\left\{1 - \widehat{S}_Y(y_j) - \widehat{S}_Y(y_j^-)\right\}\widehat{S}(\widehat{\tau}_X^-, dy_j) \le \ \widehat{S}_X(\widehat{\tau}_X^-) - \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{P} 0, \\
D &= \left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\}\left\{1 - \widehat{S}_Y(\widehat{\tau}_Y^-)\right\}\widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \le \ \widehat{S}(\widehat{\tau}_X^-, \widehat{\tau}_Y^-) \xrightarrow{P} 0.
\end{aligned}$$

Because of the above

$$A + (B + C + D) \xrightarrow{P} \int\int \left\{1 - S_X(x) - S_X(x^-)\right\}\left\{1 - S_Y(y) - S_Y(y^-)\right\} S(dx, dy),$$

which proves that the righthand side of (2.21) converges in probability to the righthand side of (4.3).

Now, we prove that $\widehat{c}^H_\rho \xrightarrow{P} c_\rho$. $\widehat{c}^H_\rho$ is a product of square roots of variances of $1 - \widehat{S}^H_X(x) - \widehat{S}^H_X(x^-)$ and $1 - \widehat{S}^H_Y(y) - \widehat{S}^H_Y(y^-)$. Let $Z = 1 - \widehat{S}^H_X(x) - \widehat{S}^H_X(x^-)$. Then $\widehat{\mathrm{E}}(Z) = \sum_{i^*}\left\{1 - \widehat{S}^H_X(x_{i^*}) - \widehat{S}^H_X(x_{i^*}^-)\right\}\widehat{S}^H_X(dx_{i^*}) = 0$ for any properly defined continuous or discrete survival function (see proof of Property 8 in Li and Shepherd, 2012) and

$$\begin{aligned}
\widehat{\mathrm{Var}}(Z) \;=\; \widehat{\mathrm{E}}(Z^2) \;&=\; \sum_{i^*}\left\{1 - \widehat{S}^H_X(x_{i^*}) - \widehat{S}^H_X(x_{i^*}^-)\right\}^2 \widehat{S}^H_X(dx_{i^*}) \\
&=\; \underbrace{\sum_i \left\{1 - \widehat{S}_X(x_i) - \widehat{S}_X(x_i^-)\right\}^2 \widehat{S}_X(dx_i)}_{\widehat{H}} + \underbrace{\left\{1 - \widehat{S}_X(\widehat{\tau}_X^-)\right\}^2 \widehat{S}_X(\widehat{\tau}_X^-)}_{\widehat{J}} \\
&\xrightarrow{P}\; \int \left\{1 - S_X(x) - S_X(x^-)\right\}^2 S_X(dx) \;+\; 0 \\
&=\; \mathrm{Var}(1 - S_X(T_X) - S_X(T_X^-)).
\end{aligned}$$

In the above, $\widehat{J} \xrightarrow{P} 0$ because $\widehat{S}_X(\widehat{\tau}_X^-) \xrightarrow{P} 0$ as $\widehat{\tau}_X \longrightarrow t^+_{X,max}$ (see *Lemma 1*). Similarly to the argument about the consistency of $\widehat{A}$, we have $\widehat{H} \xrightarrow{P} c_\rho$. The proof for the variance of $1 - \widehat{S}^H_Y(y) - \widehat{S}^H_Y(y^-)$ is similar. We have proved that $\widehat{c}^H_\rho \xrightarrow{P} c_\rho$ and that the righthand side of (2.21) is consistent for the righthand side of (4.3), therefore, by the continuous mapping theorem, $\widehat{\rho}^H_S \xrightarrow{P} \rho_S$.

2.7.2 Tables

Table 2.2: Bias [RMSE] of $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ as estimates of the overall Spearman's correlation, ρ_S, under unbounded censoring.

N	Censoring	Percent	Method	Indep	Clayton	Frank	Frank
	Scenario	Censored		$\rho_S = 0$	$\rho_S = 0.2$	$\rho_S = 0.2$	$\rho_S = -0.2$
100	No Censoring		ρ_S^H	-0.001 [0.103]	-0.001 [0.100]	-0.005 [0.099]	0.002 [0.100]
			ρ_{IMI}	0.003 [0.102]	0.002 [0.095]	-0.018 [0.095]	0.008 [0.097]
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.001 [0.112]	-0.001 [0.108]	-0.004 [0.110]	0.003 [0.109]
			ρ_{IMI}	0.001 [0.115]	0.032 [0.118]	-0.002 [0.107]	0.000 [0.106]
		70%	ρ_S^H	0.005 [0.217]	-0.021 [0.216]	-0.006 [0.219]	0.015 [0.215]
			ρ_{IMI}	-0.004 [0.159]	0.091 [0.176]	-0.002 [0.153]	0.017 [0.157]
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	0.003 [0.121]	-0.004 [0.116]	-0.002 [0.116]	0.003 [0.117]
			ρ_{IMI}	0.004 [0.117]	0.036 [0.118]	-0.004 [0.111]	0.005 [0.110]
		$(30\%, 70\%)$	ρ_S^H	0.001 [0.201]	0.001 [0.191]	-0.016 [0.193]	0.012 [0.195]
			ρ_{IMI}	-0.008 [0.142]	0.065 [0.156]	-0.004 [0.138]	0.008 [0.139]
		$(70\%, 70\%)$	ρ_S^H	-0.010 [0.261]	-0.016 [0.271]	-0.036 [0.278]	0.022 [0.260]
			ρ_{IMI}	-0.007 [0.174][1]	0.100 [0.197]	-0.018 [0.169][1]	0.019 [0.179][2]
200	No Censoring		ρ_S^H	0.000 [0.068]	0.000 [0.068]	-0.003 [0.067]	0.001 [0.065]
			ρ_{IMI}	0.001 [0.069]	0.012 [0.069]	-0.012 [0.069]	0.008 [0.067]
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.002 [0.078]	0.000 [0.079]	-0.003 [0.079]	0.002 [0.076]
			ρ_{IMI}	-0.005 [0.083]	0.034 [0.083]	-0.005 [0.076]	0.003 [0.073]
		70%	ρ_S^H	0.003 [0.164]	-0.008 [0.159]	0.001 [0.156]	0.008 [0.158]
			ρ_{IMI}	0.005 [0.120]	0.094 [0.143]	-0.011 [0.110]	0.010 [0.114]
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	-0.003 [0.084]	-0.001 [0.081]	-0.001 [0.082]	-0.002 [0.081]
			ρ_{IMI}	0.001 [0.082]	0.040 [0.087]	-0.006 [0.077]	0.004 [0.080]
		$(30\%, 70\%)$	ρ_S^H	0.003 [0.148]	0.004 [0.139]	-0.004 [0.145]	0.002 [0.148]
			ρ_{IMI}	-0.002 [0.102]	0.081 [0.126]	-0.010 [0.097]	0.014 [0.100]
		$(70\%, 70\%)$	ρ_S^H	0.007 [0.231]	0.004 [0.209]	-0.010 [0.217]	0.009 [0.214]
			ρ_{IMI}	0.003 [0.126]	0.115 [0.167]	-0.012 [0.118]	0.015 [0.124]

[1] In one out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

[2] In four out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

Table 2.3: Power and type I error rate for the overall Spearman's correlation, ρ_S, measured by $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ under unbounded censoring.

N	Censoring Scenario	Percent Censored	Method	Indep $\rho_S = 0$	Clayton $\rho_S = 0.2$	Frank $\rho_S = 0.2$	Frank $\rho_S = -0.2$
100	No Censoring		ρ_S^H	0.056	0.497	0.491	0.497
			ρ_{IMI}	0.055	0.532	0.447	0.488
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.041	0.393	0.390	0.396
			ρ_{IMI}	0.042	0.533	0.397	0.415
		70%	ρ_S^H	0.031	0.137	0.152	0.130
			ρ_{IMI}	0.004	0.201	0.075	0.028
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	0.047	0.361	0.366	0.360
			ρ_{IMI}	0.040	0.516	0.362	0.350
		$(30\%, 70\%)$	ρ_S^H	0.048	0.166	0.162	0.170
			ρ_{IMI}	0.013	0.311	0.132	0.129
		$(70\%, 70\%)$	ρ_S^H	0.032	0.101	0.082	0.090
			ρ_{IMI}	0.001[1]	0.132	0.017[1]	0.010[2]
200	No Censoring		ρ_S^H	0.042	0.802	0.803	0.819
			ρ_{IMI}	0.039	0.858	0.756	0.792
	$C_X \equiv C_Y$,	30%	ρ_S^H	0.034	0.693	0.685	0.704
			ρ_{IMI}	0.054	0.828	0.680	0.691
		70%	ρ_S^H	0.034	0.239	0.237	0.233
			ρ_{IMI}	0.010	0.497	0.149	0.112
	$C_X \perp C_Y$,	$(30\%, 30\%)$	ρ_S^H	0.043	0.666	0.651	0.675
			ρ_{IMI}	0.043	0.837	0.657	0.630
		$(30\%, 70\%)$	ρ_S^H	0.038	0.314	0.299	0.307
			ρ_{IMI}	0.018	0.657	0.291	0.283
		$(70\%, 70\%)$	ρ_S^H	0.038	0.145	0.134	0.138
			ρ_{IMI}	0.001	0.391	0.071	0.037

[1] In one out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

[2] In four out of 1000 cases $\widehat{\rho}_{IMI}$ was not successful in computing the correlation.

2.7.3 Figures

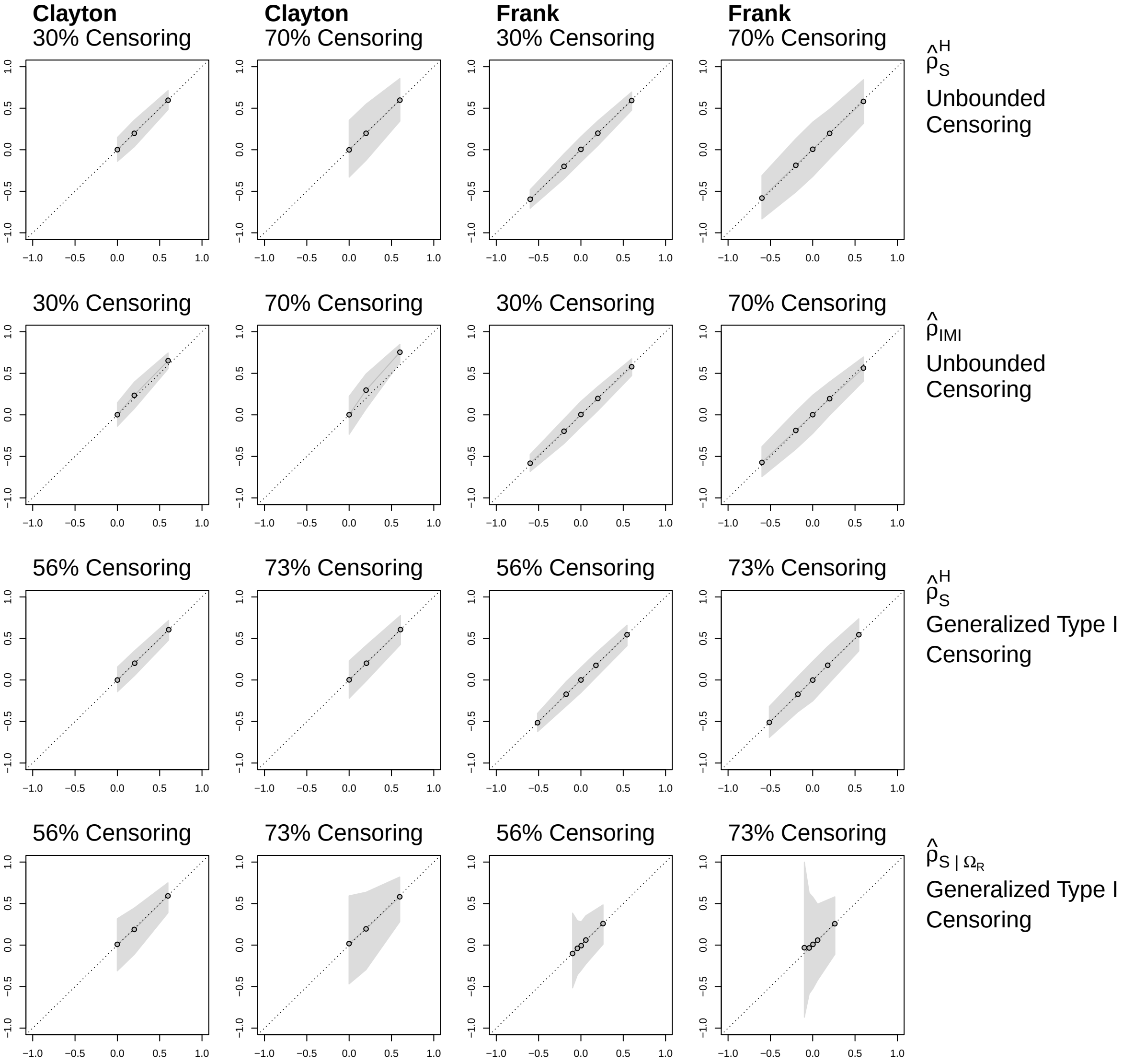

Figure 2.6: Point estimates (x-axis) vs population parameters (y-axis) under different univariate censoring scenarios. The top and second rows are $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ as estimators of the overall Spearman's correlation, ρ_S. The third row is $\widehat{\rho}_S^H$ as an estimator of ρ_S^H. The bottom row is $\widehat{\rho}_{S|\Omega_R}$ as an estimator of $\rho_{S|\Omega_R}$. The columns represent Clayton's and Frank's copulas. The population parameters for Clayton's family are 0, 0.2, and 0.6 for all estimates. For Frank's family, the population parameters of ρ_S are -0.6, -0.2, 0.2, and 0.6; the population parameters of ρ_S^H are -0.512, -0.173, 0.180, and 0.545; the population parameters of $\rho_{S|\Omega_R}$ are -0.098, -0.042, 0.058, and 0.261. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles. For generalized type I censoring, the restricted region, Ω_R, is defined by the median survival times.

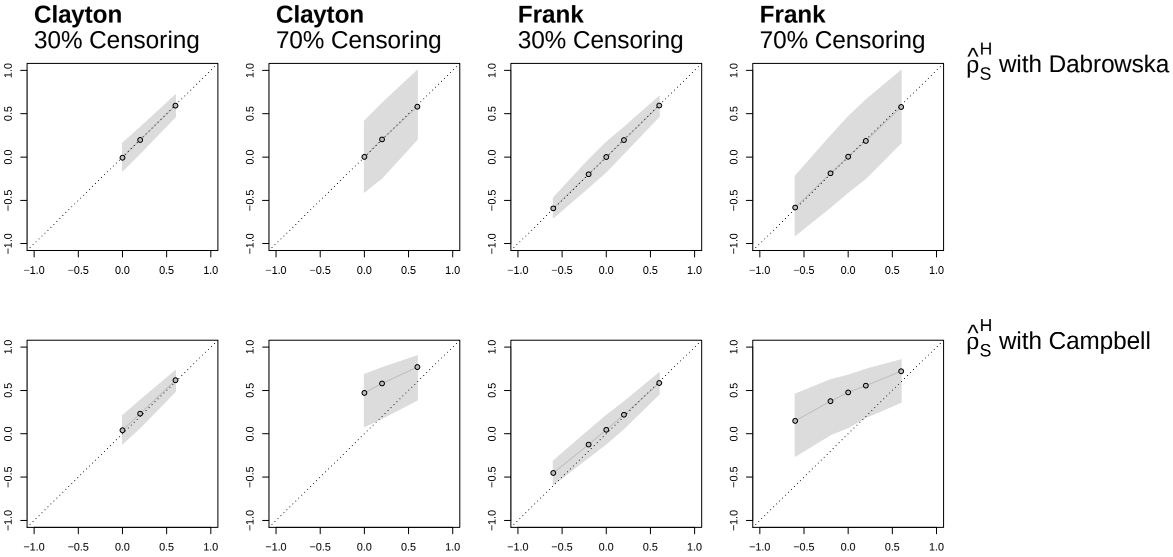

Figure 2.7: Performance of $\widehat{\rho}_S^H$ with survival surface estimators of Dabrowska (1988) (top row) and Campbell (1981) (bottom row) under bivariate unbounded censoring. The columns represent Clayton's and Frank's copulas under moderate and heavy censoring. The x-axis is the true overall Spearman's correlation; the y-axis is an estimate. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles.

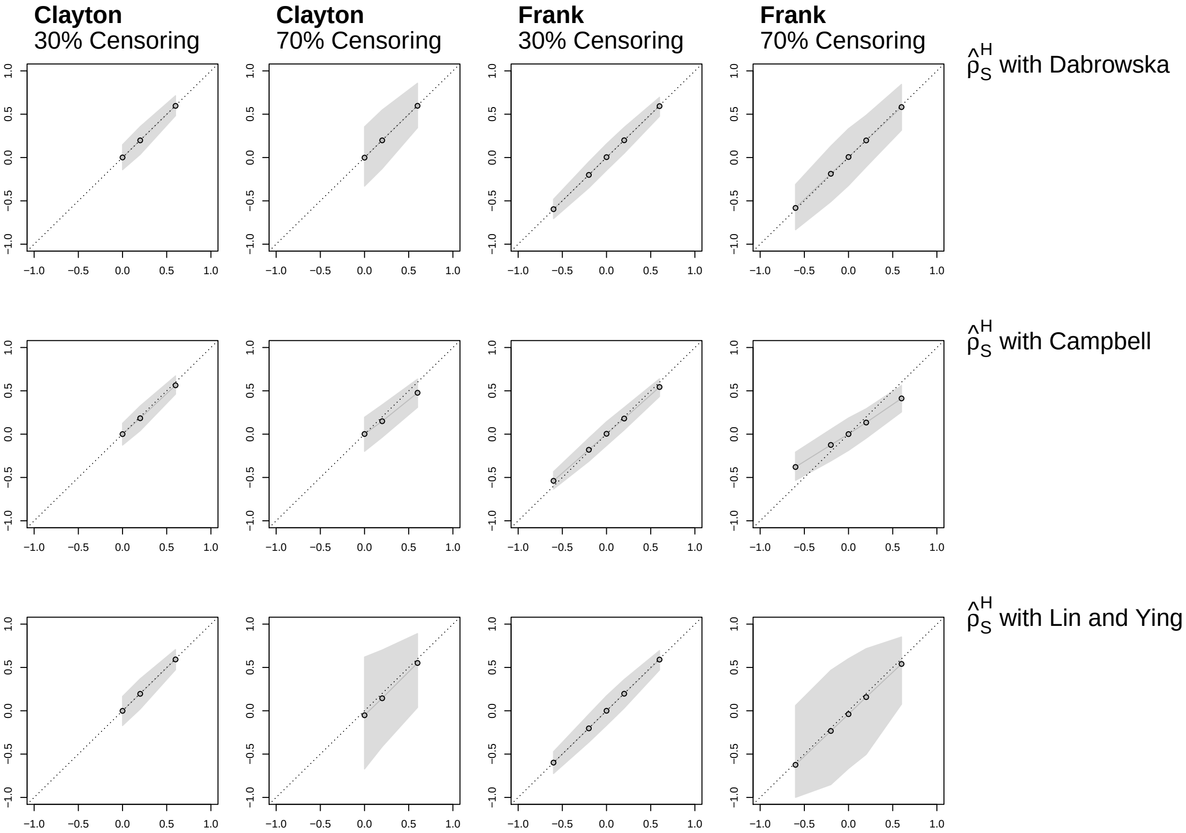

Figure 2.8: Performance of $\widehat{\rho}_S^H$ with survival surface estimators of Dabrowska (1988) (top row), Campbell (1981) (middle row), and Lin and Ying (1993) (bottom row) under univariate unbounded censoring. The columns represent Clayton's and Frank's copulas under moderate and heavy censoring. The x-axis is the true overall Spearman's correlation; the y-axis is an estimate. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025^{th} and 0.975^{th} quantiles.

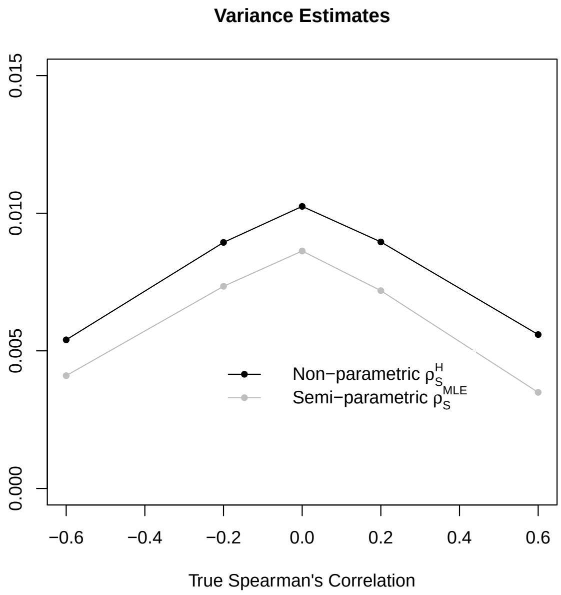

Figure 2.9: Efficiency of $\widehat{\rho}_S^H$ vs $\widehat{\rho}_S^{MLE}$. The black and gray lines are the variances of $\widehat{\rho}_S^H$ vs $\widehat{\rho}_S^{MLE}$ respectively. The data are simulated 1000 times with 200 pairs generated from Frank's copula family; the univariate unbounded censoring at 50% is applied. The relative efficiency $\mathrm{Var}(\widehat{\rho}_S^H)/\mathrm{Var}(\widehat{\rho}_S^{MLE})$ ranged from 1.19 (for $\rho_S = 0$) to 1.60 (for $\rho_S = 0.6$).

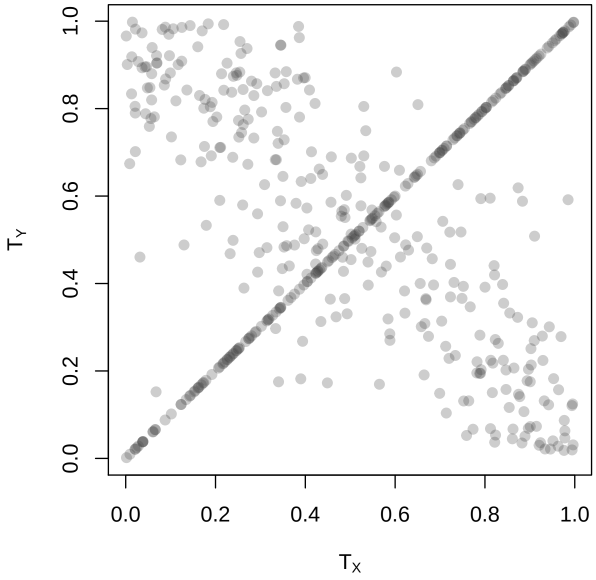

Figure 2.10: Illustration of the mixture distribution composed of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall Spearman's correlation being about -0.0813. T_X and T_Y are uniformly distributed.

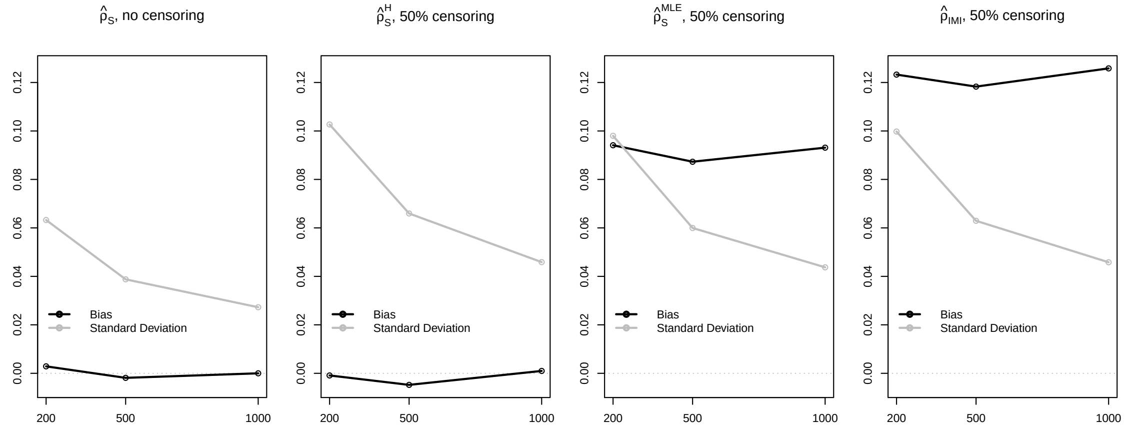

Figure 2.11: Bias and standard deviation as functions of the sample size for $\widehat{\rho}_S$ (first panel), $\widehat{\rho}_S^H$ (second panel), $\widehat{\rho}_S^{MLE}$ (third panel), and $\widehat{\rho}_{IMI}$ (forth panel). Estimator $\widehat{\rho}_S$ is computed as Spearman's rank correlation for uncensored data. Estimators $\widehat{\rho}_S^H$, $\widehat{\rho}_S^{MLE}$, and $\widehat{\rho}_{IMI}$ are computed under 50% random unbounded censoring. The bivariate survival data are simulated as a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall Spearman's correlation being about -0.0813 (see Figure 2.10 for illustration).

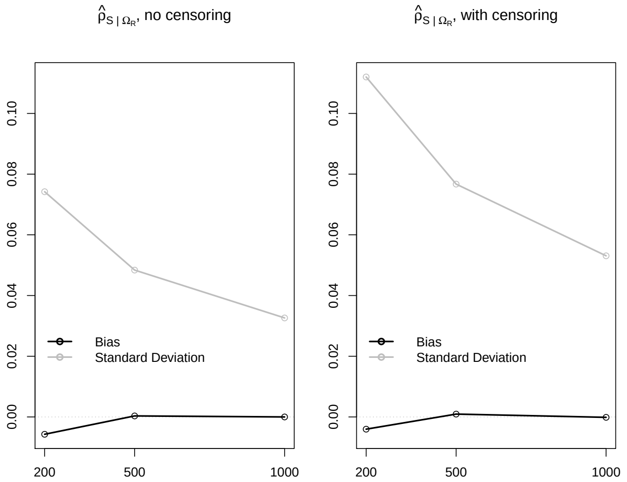

Figure 2.12: Bias and standard deviation as functions of the sample size for estimates of Spearman's correlation within the restricted region with no censoring and therefore using standard methods (left panel) and with censoring and therefore computing $\widehat{\rho}_{S|\Omega_R}$ (right panel) as described in Section 3.1. Estimator $\widehat{\rho}_S$ is computed as Spearman's rank correlation for uncensored pairs with each event time less than the median event time. The restricted region Ω_R is defined by the median follow-up time for both times to event. The bivariate survival data are simulated as a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula family with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the true overall and restricted Spearman's correlations being about -0.081 and 0.85 respectively. Unbounded 50% censoring is applied to the entire sample. The effective proportion of uncensored events for $\widehat{\rho}_{S|\Omega_R}$ is 25% for each time to event (Figure 2.10 illustrates an uncensored sample).

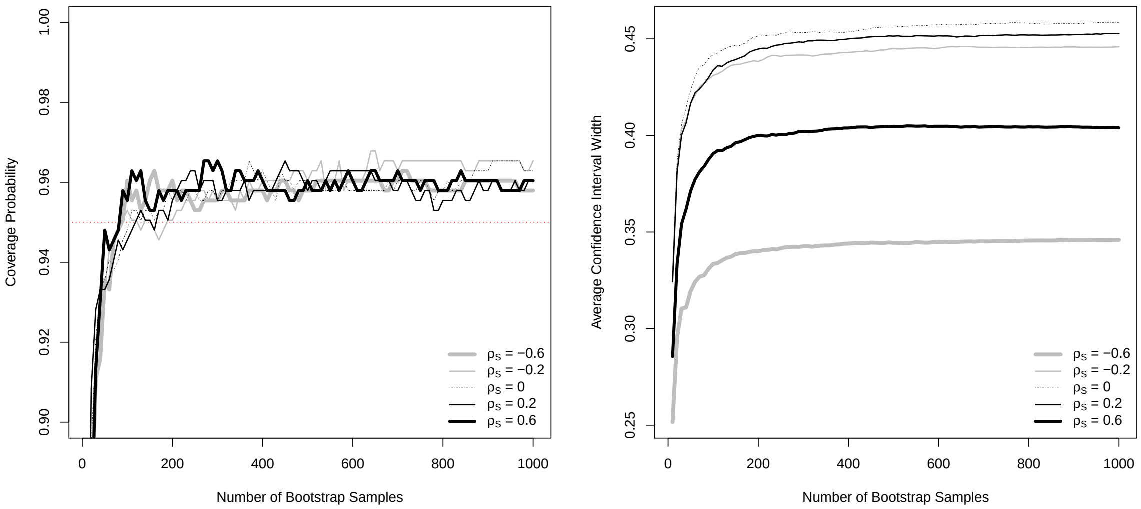

Figure 2.13: Coverage probability (left panel) and average width (right panel) of the bootstrap confidence intervals for $\widehat{\rho}_S^H$ as an estimate of ρ_S under 50% unbounded censoring. The data are simulated from Frank's copula with parameters corresponding to Spearman's correlation of -0.6, -0.2, 0, 0.2, and 0.6. The sample size is 200 and the number of simulations is 1000. The 95% bootstrap confidence bounds are computed as the 0.025^{th} and 0.975^{th} percentiles.

CHAPTER 3

SPEARMAN-LIKE CORRELATION MEASURE ADJUSTING FOR COVARIATES IN BIVARIATE SURVIVAL DATA

3.1 Introduction

Many scientific studies focus on measuring correlation between two variables. Correlation can occur when one variable affects the other or when a third variable affects both variables of interest. In either case, it can be measured using an unadjusted correlation. However, to understand the mechanism behind the relationship between two variables, it is essential to be able to compute the conditional and adjusted correlation. For example, for people living with HIV infection and receiving antiretroviral therapy (ART), an increase in viral load will likely prompt the treating physician to change the person's ART regimen. Therefore, time to viral load and time to regimen change should be correlated. However, because of differences in clinical practices, this correlation may vary depending on the country or region, and other factors such as the patient's age, lab results, or comorbidities may modify this correlation.

Measuring correlation can be challenging in the presence of right-censoring. Right-censoring is a general term for data with values that are not always observed due to either an upper detection limit (e.g., income is often collected with the highest category of "$\$50,000$ *or higher*") or due to the end of study. Sometimes an outcome may be censored due to the occurrence of a competing event that is part of the research question, but we do not consider the setting of competing risks here. We are interested in the correlation between variables in a single subject (e.g., time to viral failure and time to regimen change) or in paired subjects (e.g., education of fathers and income of sons). Specifically, we are interested in Spearman's correlation, a non-parametric rank correlation measure. Unlike Pearson's correlation, it is invariant to variable transformation and can detect associations when variables are non-linearly related. It also approximates well Pearson's correlation for normally distributed variables (Kruskal, 1958). The interpretation of Spearman's correlation is straightforward for continuous and ordinal data (i.e., the correlation of the ranked data) and is desirable in the context of right-censored data.

In the setting of right-censoring, several Spearman-like statistics have been suggested. Test statistics of Cuzick (1982) and Dabrowska (1986) resemble Spearman's correlation under certain assumptions and are effectively unscaled estimates of Spearman's correlation when applied to uncensored data. Semi-parametric approaches for measuring correlation using copulas have been considered by Carriere (2000), Romeo et al. (2006), and Zhang (2008); although some of these authors did not estimate Spearman's correlation, it can be computed from the estimated copulas (Nelsen, 2007). Schemper et al. (2013) proposed a semi-parametric iterative multiple imputation method to estimate Spearman's correlation based on a normal copula. These semi-parametric approaches tend to be stable and efficient when the copula is properly specified but can be misleading in the presence of misspecification. In Chapter 2, we proposed to estimate Spearman's correlation based on non-parametric estimators of the bivariate survival surface; we used the non-parametric and consistent estimator of Dabrowska (1988). Although this approach does not make parametric assumptions, its reliance on non-parametric estimators of the bivariate survival surface, which is notoriously difficult to estimate (Kalbfleisch and Prentice, 2011), can lead to poor efficiency as well as instability when the sample size is small and there is heavy censoring.

Additionally, it is desirable to measure the rank correlation between bivariate right-censored data while adjusting for covariates. Several bivariate survival models have been suggested to estimate adjusted correlation. To mention a few, Clayton and Cuzick (1985) proposed a method of estimating a *cross ratio*, another association measure used in bivariate survival (Clayton, 1978), in the context of a frailty model that may include covariates. Shih and Louis (1995) estimated Kendall's tau by first fitting separate Cox models conditional on covariates for both of the time-to-event variables, and then using maximum likelihood to estimate association assuming different parametric dependency structures defined with copulas. Prentice and Hsu (1997) developed a method that used Cox models to estimate marginal distributions for each variable conditional on covariates and assumed a semi-parametric pairwise dependency structure. To our knowledge, there is no estimator of Spearman's partial correlation adjusted for covariates for bivariate survival data.

In this manuscript, we derive unadjusted, partial, and conditional estimators of Spearman's correlation for bivariate survival data. Our estimators are extensions of the approach of Liu et al. (2018), who showed that Spearman's rank correlation for uncensored continuous or discrete variables is equivalent to the correlation between probability-scale residuals. Probability-scale residuals are well defined with continuous, ordinal, and right-censored data (Shepherd et al., 2016), and can be computed with unadjusted or adjusted estimates of the marginal survival distributions. This is advantageous because it avoids computing estimates of the bivariate survival surface, and it provides a straightforward extension for covariate-adjustment. In Section 3.2, we review the definition of PSRs, define our unadjusted Spearman's correlation estimator using PSRs, describe our approach to estimating its variance, discuss its population parameter, and show how other Spearman-like test statistics are related to our method. In Section 3.3, we focus on estimation and inference of partial, and conditional Spearman's correlation with right-censored data using PSRs. In Section 3.4, we use simulations to estimate the performance of our statistics and compare them to other approaches. In Section 3.5, we apply our method to an HIV study examining the association between times from treatment initiation to viral failure and regimen change. Finally, in Section 3.6, we discuss our approach and future directions.

3.2 Unadjusted Correlation of PSRs

3.2.1 Notation and Definitions

Let T_X and T_Y be time to event variables for a single subject or a pair of subjects. Time to events T_X and T_Y can be censored at times C_X and C_Y, respectively. We assume independence between (T_X, T_Y) and (C_X, C_Y), but C_X and C_Y can be dependent. Without loss of generality we can assume that (T_X, T_Y) and (C_X, C_Y) are defined on $[0, \infty) \times [0, \infty)$. If T_X and T_Y are observed on a single subject then it is likely that $C_X = C_Y$. When $C_X = C_Y$ with probability 1, we call this *univariate* censoring, otherwise censoring is *bivariate.* Often, studies are restricted by the maximum follow-up time, which we denote as τ_X and τ_Y for T_X and T_Y, respectively. When $\tau_X = \tau_Y = \infty$, we call it *unbounded* censoring. When $\tau_X < \infty$ or $\tau_Y < \infty$, we refer to it as *type I* censoring. Type I censoring can be *strict* or *generalized.* Strict type I censoring implies that all subjects start the study at the same calendar time, and there is no censoring other than at the end of the study. Generalized type I censoring allows other patterns of study entry and censoring before the end of the study as long as the resulting censoring mechanism is uninformative.

If the relationship between T_X and T_Y are confounded by a set of covariates $\boldsymbol{Z}$ we assume independence between T_X and C_X conditional on $\boldsymbol{Z}$ and between T_Y and C_Y conditional on $\boldsymbol{Z}$. As a result of censoring, we only observe $X = \min(T_X, C_X)$ and $Y = \min(T_Y, C_Y)$ and event indicators $\Delta_X = \mathbb{1}(T_X \leq C_X)$ and $\Delta_Y = \mathbb{1}(T_Y \leq C_Y)$. We denote marginal and joint cumulative distribution functions of T_X and T_Y as

$F_X(x) = \Pr(T_X \le x)$, $F_Y(y) = \Pr(T_Y \le y)$, $F(x,y) = \Pr(T_X \le x, T_Y \le y)$, and marginal and joint cumulative distribution functions of C_X and C_Y as $G_X(x) = \Pr(C_X \le x)$, $G_Y(y) = \Pr(C_Y \le y)$, $G(x,y) = \Pr(C_X \le x, C_Y \le y)$. We define $F_X(x^-) = \lim_{t\uparrow x} F_X(t)$ and $F(x^-,y) = \lim_{t\uparrow x} F(t,y)$; functions $F_Y(y^-)$ and $F(x,y^-)$ are defined similarly.

As mentioned by Liu et al. (2018), in the absence of censoring, the population parameter for Spearman's correlation between T_X and T_Y can be defined as

$$\rho_S = \mathrm{Cor}\left\{ \frac{F_X(T_X) + F_X(T_X^-)}{2},\ \frac{F_Y(T_Y) + F_Y(T_Y^-)}{2} \right\}.$$

When both T_X and T_Y are continuous the above definition translates into a better known expression, $\rho_S = \mathrm{Cor}\{F_X(T_X),\ F_Y(T_Y)\}$, the grade correlation (Kruskal, 1958), which according to Liu et al. (2018) can be presented as

$$\rho_S/c_\rho = \mathrm{Cov}\left[\left\{F_X(T_X) + F_X(T_X^-) - 1\right\}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right], \tag{3.1}$$

where $c_\rho = \left[\mathrm{Var}\left\{F_X(T_X) + F_X(T_X^-) - 1\right\}\mathrm{Var}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right]^{-1/2}$, and $c_\rho = 3$ when T_X and T_Y are continuous. The right-hand side of (4.1) is the covariance of probability-scale residuals (PSRs) proposed and studied by Li and Shepherd (2012) and Shepherd et al. (2016) and defined as

$$\begin{aligned} r(t_X, F_X) &= \mathrm{E}\{\mathrm{sign}(t_X, T_X)\} \\ &= \Pr(T_X < t_X) - \Pr(T_X > t_X) = F_X(t_X^-) + F_X(t_X) - 1, \end{aligned}$$

where $\mathrm{sign}(t_X, T_X)$ is -1, 0, and 1 for $t_X < T_X$, $t_X = T_X$, and $t_X > T_X$, respectively. Shepherd et al. (2016) extended this definition to right-censored time-to-event data. When the time to event T_X is unknown because of censoring, they suggested to use the expectation of PSRs, $r(x, F_X, \Delta_X = 0) = E\{r(T_X, F_X)|T_X > x\}$. This led to the following definition: $r(x, F_X, \delta_X) = F_X(x) - \delta_X(1 - F_X(x-))$, where (x, δ_X) is a realization of (X, Δ_X). PSRs can also be presented as a random variable,

$$r(X, F_X, \Delta_X) = F_X(X) - \Delta_X(1 - F_X(X-)). \tag{3.2}$$

3.2.2 Correlation of Probability-Scale Residuals

Because in the absence of censoring the correlation of PSRs equals Spearman's correlation, it is natural to consider it as a measure of association in the presence of censoring,

$$\rho_{PSR} = \text{ Cor}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)\right\} = \frac{\text{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)\right\}}{\sqrt{\text{Var}\left\{r(X, F_X, \Delta_X)\right\}\ \text{Var}\left\{r(Y, F_Y, \Delta_Y)\right\}}}. \tag{3.3}$$

Shepherd et al. (2016) proved that PSRs have zero expectation when F_X are F_Y are properly specified and $T_X \perp C_X$ and $T_Y \perp C_Y$; therefore definition (3.3) can be rewritten as

$$\rho_{PSR}/c_\rho = \text{E}_{X,Y,\Delta_X,\Delta_Y}\left\{r(X, F_X, \Delta_X)\ r(Y, F_Y, \Delta_Y)\right\}, \tag{3.4}$$

where $c_\rho = \left[\text{E}_{X,\Delta_X}\left\{r^2(X, F_X, \Delta_X)\right\}\ \text{E}_{Y,\Delta_Y}\left\{r^2(Y, F_Y, \Delta_Y)\right\}\right]^{-1/2}$. The estimation of (3.4) is straight forward. Let $(x_i, \delta_{X,i}, y_i, \delta_{Y,i})$ for $i = 1, ..., n$ be independent and identically distributed (iid) draws from $(X, \Delta_X, Y, \Delta_Y)$. Then

$$\widehat{\rho}_{PSR}/\widehat{c}_\rho \ = \ \frac{1}{n}\sum_n \left\{r(x_i, \widehat{F}_X, \delta_{X,i})r(y_i, \widehat{F}_Y, \delta_{Y,i})\right\}, \tag{3.5}$$

where $\widehat{c}_\rho = \left[\left\{\frac{1}{n}\sum_n r^2(x_i, \widehat{F}_X, \delta_{X,i})\right\}\left\{\frac{1}{n}\sum_n r^2(y_i, \widehat{F}_Y, \delta_{Y,i})\right\}\right]^{-1/2}$, and $\widehat{F}_X$ and $\widehat{F}_Y$ are Kaplan-Meier estimates of F_X and F_Y, respectively. The variance of the estimator can be computed using a large sample approximation approach with M-estimation and the delta-method (see Stefanski and Boos (2002)). M-estimation can be used under very general assumptions and requires computing estimating equations and their derivatives for each unknown parameter, including Kaplan-Meier (KM) estimates of $F_X(x_i)$ and $F_Y(y_i)$ for all i. We use the results of Stute (1995), who developed estimating equations for the Kaplan-Meier estimator. Appendix 3.C provides some details.

3.2.3 Marginal Spearman-like Statistics in the Literature

In this section, we briefly review some Spearman-like statistics in the literature and highlight their relationship to $\widehat{\rho}_{PSR}$. The idea of testing association in bivariate survival data using marginal distributions was suggested previously. Cuzick (1982) studied a situation where the two underlying times T_X and T_Y are assumed to be connected to a common latent variable: $T_X = aZ + e_X$, $T_Y = bZ + e_Y$, where the parameters a and b are constrained to be $b = a\lambda$ and both e_X and e_Y follow a logistic distribution. To test the null hypothesis $a = 0$, Cuzick suggested statistic $\sum_i (s_{X,i} \cdot s_{Y,i})$, where $s_{\cdot,i} = 1 - 2\widehat{F}_{\cdot,i}$ for observed and $1 - \widehat{F}_{\cdot,i}$ for censored observations, which is "a statistic equivalent to Spearman's rank correlation coefficient" when there is no censoring. In comparison, for continuous time, $r(\cdot, \widehat{F}_{\cdot}, 1) = 2\widehat{F}_{\cdot,i} - 1$ (for observed) and $r(\cdot, \widehat{F}_{\cdot}, 0) = \widehat{F}_{\cdot,i}$ (for censored times-to-event); therefore, the statistic of Cuzick is the unscaled covariance of PSRs for uncensored and continuous time. Dabrowska (1986) defined a more general version of $\sum_i (s_{X,i} \cdot s_{Y,i})$ for testing the null of independence of F_X and F_Y when the underlying times are continuous, where $s_{\cdot,i} = \delta_{\cdot,i} - (1 + \delta_{\cdot,i})\widehat{F}_{\cdot,i}$ is "the censored-data version of the Spearman test". Since $s_{\cdot,i} = \delta_{\cdot,i}(1 - \widehat{F}_{\cdot,i}) - \widehat{F}_{\cdot,i} = -r_{\cdot,i}(\cdot, \widehat{F}_{\cdot,i}, \delta_{\cdot,i})$, Dabrowska's statistic is the unscaled covariance of PSRs for continuous time. To test independence for bivariate current status data, Ding and Wang (2004) suggested to use statistic $(1/n) \sum_i \left[\left\{ \delta_{X,i} - \widehat{F}_X(c_{X,i}) \right\} \left\{ \delta_{Y,i} - \widehat{F}_Y(c_{Y,i}) \right\} \right]$, where $c_{X,i}$ and $c_{Y,i}$ are the times of collecting the status data. Their statistic is a covariance of PSRs for current status data, as defined by Shepherd et al. (2016). All three statistics above are scaled or unscaled covariances of PSRs under certain conditions. The advantage of using the correlation instead of the covariance is that it has a convenient range from -1 to 1, and therefore can be used not only as a test but as a Spearman-like correlation measure.

3.2.4 Population Parameters of PSRs' Correlation and Spearman's Correlation

It is important to recognize that in the presence of censoring, ρ_{PSR} does not equal Spearman's correlation, ρ_S. Unfortunately, the population parameter, ρ_{PSR}, depends on the censoring distribution. In addition, it is not possible to derive a general expression of ρ_{PSR} in terms of ρ_S. Some details are in Appendix 3.B. However, as will be illustrated in Section 3.4 via simulations, the mean squared error of $\widehat{\rho}_{PSR}$ for ρ_S is often smaller than that of other unbiased non-parametric estimators of ρ_S. The difference between ρ_{PSR} and ρ_S is often quite small, particularly when the probability of censoring is low. And if $\rho_S = 0$, $\rho_{PSR} = 0$ regardless of the censoring distribution (see Appendix 3.A).

To illustrate the difference between ρ_{PSR} and ρ_S, we derived algebraic expressions of ρ_{PSR} for four specific cases. We considered two correlation structures, perfect positive and negative Spearman's correlation ($\rho_S \in \{-1, 1\}$), and two censoring scenarios, strict type I censoring and unbounded censoring, where $C_X \perp C_Y$. The proportion of censored observations, γ, was varied from 0 to 1/2. Details are in Appendix 3.B. The population parameter ρ_{PSR} for each of the four cases (Spearman's correlations -1 and 1 for type I and unbounded censoring), is illustrated in Figure 3.1.

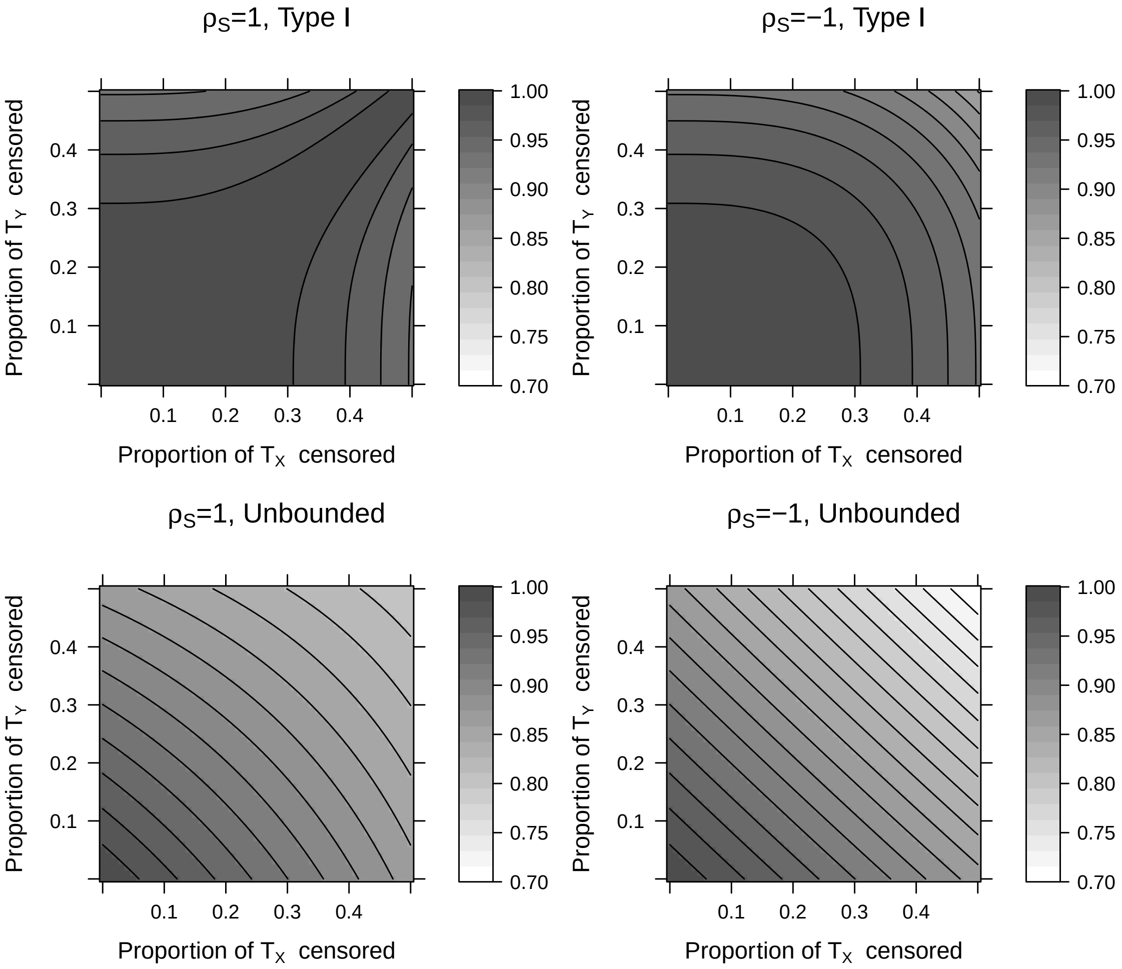

Figure 3.1: Contour plots for the absolute value of ρ_{PSR} as a function of the proportion censored when $\rho_S = 1$. The left and right columns represent scenarios of perfect positive and perfect negative correlations, respectively. The top and bottom rows are strict type I and unbounded bivariate censoring, respectively. Each contour represents a change of 0.016 absolute correlation value.

The Figure shows that heavier censoring results in lower absolute values of estimated correlation except for strict type I censoring with equal proportions of censoring for both events. The estimated absolute correlation values are also lower when censoring is unbounded, and the correlation is negative. For example, for perfect positive correlation and 50% strict type I censoring, $\rho_{PSR} = \rho_S = 1$. For perfect positive correlation and 50% unbounded censoring, $\rho_{PSR} = 0.80$. For perfect negative correlation when both events are 50% censored, $\rho_{PSR} = -0.86$ for strict type I censoring and $\rho_{PSR} = -0.70$ for unbounded censoring. The performance of ρ_{PSR} is also studied through simulations of other dependency structures and censoring scenarios in Section 3.4.

3.3 Partial, Conditional, and Partial-conditional Correlation of PSRs

3.3.1 Population Parameters

Correlation between two variables is often confounded by another variable or a set of variables, $\boldsymbol{Z}$. For instance, the correlation between the time to viral failure and time to regimen change could be confounded by study site, age, sex, and CD4 count at ART initiation. For uncensored data, Liu et al. (2018) showed that adjusted or *partial* Spearman's correlation could be computed as the correlation between PSRs from models adjusting for confounders. For right-censored data, partial Spearman's correlation can be defined in a similar way:

$$\rho_{PSR\cdot\boldsymbol{Z}}/c_{\rho\cdot\boldsymbol{Z}} = \ \operatorname{Cov}\Big\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)\Big\} \tag{3.6}$$

where $c_{\rho\cdot\boldsymbol{Z}} = \Big[\operatorname{Var}\Big\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X)\Big\}\operatorname{Var}\Big\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)\Big\}\Big]^{-1/2}$, where $F_{X|\boldsymbol{Z}}$ is the distribution of T_X conditional on $\boldsymbol{Z}$ and $F_{Y|\boldsymbol{Z}}$ is similarly defined.

In addition to studying adjusted correlation, we might be interested in studying how the correlation is modified by another variable or a set of variables. For example, we might ask whether the correlation between time to viral failure and time to regimen change varies with the CD4 count at ART initiation. This question can be answered by estimating a *conditional* correlation. Following Liu et al. (2018), we define an extension of Spearman's conditional correlation for right-censored data as

$$\rho_{PSR|\boldsymbol{Z}}/c_{\rho|\boldsymbol{Z}} = \ \operatorname{Cov}\Big\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}\Big\}, \tag{3.7}$$

where $c_{\rho|\boldsymbol{Z}} = \Big[\operatorname{Var}\Big\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X)|\boldsymbol{Z}\Big\}\operatorname{Var}\Big\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}\Big\}\Big]^{-1/2}$. Note that unlike partial correlation (3.6), which is a single number, the conditional correlation defined in (3.7) is a set of numbers corresponding to different categories of $\boldsymbol{Z}$ or as a continuous function of $\boldsymbol{Z}$.

Finally, we may be interested in studying how association adjusted for variables $\boldsymbol{Z}_1$ is modified by variables $\boldsymbol{Z}_2$. This can be addressed by combining the previously described approaches and computing a *partial-conditional* correlation. For example, we might be interested in estimating correlation for different levels of CD4 count at ART initiation ($\boldsymbol{Z}_2$) adjusted for study site, age, and sex ($\boldsymbol{Z}_1$). Similar to Liu et al. (2018), we define the extension of partial-conditional Spearman's correlation for right-censored data as

$$\rho_{PSR \cdot \boldsymbol{Z}_1|\boldsymbol{Z}_2}/c_{\rho \cdot \boldsymbol{Z}_1|\boldsymbol{Z}_2} = \operatorname{Cov}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}_2\right\}, \tag{3.8}$$

where $c_{\rho \cdot \boldsymbol{Z}_1|\boldsymbol{Z}_2} = \left[\operatorname{Var}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X|\boldsymbol{Z}_2)\right\} \operatorname{Var}\left\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y|\boldsymbol{Z}_2)\right\}\right]^{-1/2}$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$.

Similar to the unadjusted parameters described in Section 3.2, these covariate-adjusted Spearman-like parameters depend on the censoring distribution and therefore are not equal to Spearman's correlation in the presence of censoring. However, they are bounded between -1 and 1, they are equal to 0 when T_X and T_Y are independent conditional on $\boldsymbol{Z}$, and they do not require estimation of the joint distribution of T_X and T_Y conditional on $\boldsymbol{Z}$.

3.3.2 Estimation

Estimation of the parameters (3.6), (3.7), and (3.8) can be performed using the steps suggested by Liu et al. (2018). Under correctly specified models and independent censoring $T_X \perp T_Y|Z$, Shepherd et al. (2016) showed that $E\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X)|\boldsymbol{Z}\right\} = E\left\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}\right\} = 0$, so the covariance and variance estimates can be approximated as expectations of the product and squared PSRs. Therefore, a plug-in estimator for the partial correlation is

$$\widehat{\rho}_{PSRs\boldsymbol{Z}}/\widehat{c}_{\rho\boldsymbol{Z}} = \frac{1}{n}\sum_n r(x_i, \widehat{F}_{X|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{X,i}) r(y_i, \widehat{F}_{Y|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{Y,i}),$$

where $\widehat{c}_{\rho \cdot \boldsymbol{Z}} = \left[\left\{\frac{1}{n}\sum_n r^2(x_i, \widehat{F}_{X|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{X,i})\right\}\left\{\frac{1}{n}\sum_n r^2(y_i, \widehat{F}_{Y|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{Y,i})\right\}\right]^{-1/2}$, and $F_{X|\boldsymbol{Z}}$ and $F_{Y|\boldsymbol{Z}}$ are fitted distributions.

Although one could use parametric survival models (e.g., exponential, Weilbull, or log-normal regressions) to estimate $F_{X|\boldsymbol{Z}}$ and $F_{Y|\boldsymbol{Z}}$, this choice seems contrary to the non-parametric nature of Spearman's rank correlation. If we wanted to preserve its non-parametric nature, we would fit a non-parametric model for each outcome. Still, these models are hard to estimate with multivariable or continuous $\boldsymbol{Z}$. We can compromise and use a rank-based semi-parametric model, for example, Cox proportional hazards regression. Other choices could be the larger class of semi-parametric linear transformation models proposed by Zeng and Lin (2007).

Estimating conditional and partial-conditional correlations is a little more complicated. For conditional correlation, after obtaining PSRs, $r(x_i, \widehat{F}_{X|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{X,i})$ and $r(y_i, \widehat{F}_{Y|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{Y,i})$, we fit linear models of $H = r(x_i, \widehat{F}_{X|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{X,i})r(y_i, \widehat{F}_{Y|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{Y,i})$, $U = r^2(x_i, \widehat{F}_{X|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{X,i})$, and $W = r^2(y_i, \widehat{F}_{Y|\boldsymbol{Z}=\boldsymbol{z}_i}, \delta_{Y,i})$ conditional on $\boldsymbol{Z}$. When $\boldsymbol{Z}$ is continuous, we use flexible modeling techniques (e.g., parametric models with $\boldsymbol{Z}$ expanded using restricted cubic splines). Next, we obtain predicted values $\widehat{H}$, $\widehat{U}$, and $\widehat{W}$ conditional on $\boldsymbol{Z}$, and then calculate $\widehat{\rho}_{PSR|\boldsymbol{Z}}$ as $\frac{\widehat{H}}{\sqrt{\widehat{U}\widehat{W}}}$ conditional on $\boldsymbol{Z}$. The ratio $\left|\frac{\widehat{H}}{\sqrt{\widehat{U}\widehat{W}}}\right|$ can exceed 1 for some $\boldsymbol{Z}$, in which case we assign $\widehat{\rho}_{PSR|\boldsymbol{Z}}$ as $\text{sign}\left(\frac{\widehat{H}}{\sqrt{\widehat{U}\widehat{W}}}\right)$. Estimating partial–conditional correlation is performed similarly, except estimating $\widehat{H}$, $\widehat{U}$, and $\widehat{W}$ conditional on $\boldsymbol{Z}_2$.

We estimate the variance of partial, conditional, and partial-conditional correlations using the bootstrap or M-estimation. If PSRs are estimated using a parametric regression, then score equations can be conveniently obtained from standard statistical software. If PSRs are estimated using Cox proportional hazards regression (Cox, 1972), the estimating equations for β-coefficients are often easily extractable. However, to obtain estimating equations for the baseline hazard, we use a full likelihood approach suggested by Breslow (1972). Variance estimation for conditional and partial-conditional correlations also requires estimating equations from the linear models that provide $\widehat{H}$, $\widehat{U}$, and $\widehat{W}$. Although M-estimation is straightforward, it can be tedious, and we do not provide all the details here. The process is similar to that described by Liu et al. (2018). Appendix 3.C contains derivatives needed for computing the standard error of our estimators when fitting either Cox models or popular parametric survival models. We have developed code in **R** that performs estimation and inference using parametric and Cox survival models, which is available as part of package `PResiduals` (Dupont et al., 2018).

3.3.3 Choice of Covariates

As an aside, the choice of adjustment variables, $\boldsymbol{Z}$, deserves some careful consideration. Note that $F_{X|\boldsymbol{Z}}$ and $F_{Y|\boldsymbol{Z}}$ are the distributions of T_X and T_Y conditional on the same set of covariates $\boldsymbol{Z}$. When times to event belong to different paired subjects, then $\boldsymbol{Z}$ will contain the union of covariates relevant for both subjects. For example, if one were interested in the prostate-specific antigen (PSA)-adjusted association between times to prostate cancer in father-son pairs, $\boldsymbol{Z}$ would need to include both PSA for the father and PSA for the son. Excluding the son's PSA from the father's model of time-to-prostate cancer (or vice versa), would implicitly assume these variables are independent conditional on the covariates remaining in the model. In contrast, when times to event belong to the same subject (e.g., times from viral failure and regimen change), it is more natural for $\boldsymbol{Z}$ to be the same for $F_{X|\boldsymbol{Z}}$ and $F_{Y|\boldsymbol{Z}}$. Throughout this paper, we assume that T_X and T_Y are adjusted for the same covariates.

3.4 Simulations

To investigate the finite sample performance of our estimators, we applied them to simulated data with different sample sizes, dependency structures, and censoring scenarios. The dependency structures were simulated using copulas (see Nelsen (2007)). For random variables T_X and T_Y with marginal CDFs $F_X(x)$ and $F_Y(y)$, a copula is the joint CDF of random variables $U = F_X(T_X)$ and $V = F_Y(T_Y)$, $C_{U,V}(u,v) = \Pr(U \leq u, V \leq v)$. Following Fan et al. (2000), we employed two commonly used copulas, *Frank*'s and *Clayton*'s. Clayton's family produces only positive association; the magnitude of the association is defined by parameter θ. Frank's family also has one parameter θ and can generate positive and negative correlations. Our general approach for unadjusted, partial, and conditional correlation was to choose parameter θ that corresponds to a target value of Spearman's correlation, ρ_S. We provide more details on simulating dependency structures below.

All simulations used 1000 replications and were performed in statistical language **R** (R Core Team, 2017) using libraries `survival` (Therneau, 2015), `lcopula` (Belzile and Genest, 2017), and `cubature` (Narasimhan and Johnson, 2017).

3.4.1 Unadjusted Correlation

3.4.1.1 Simulation Set-up

The data $(X_i, \Delta_{X,i}, Y_i, \Delta_{Y,i})$ were simulated in the following manner:

$$
\begin{aligned}
&(U_i, V_i) \sim C_{U,V}(u, v, \theta); && (3.9)\\
&T_{X,i} = F_X^{-1}(U_i), \quad \text{where } F_X(x) = 1 - e^{-x}; && (3.10)\\
&T_{Y,i} = F_Y^{-1}(V_i), \quad \text{where } F_Y(y) = 1 - e^{-y}; && (3.11)\\
&C_{X,i} \sim \text{Exponential}(\text{rate} = \lambda_X), \quad C_{Y,i} \sim \text{Exponential}(\text{rate} = \lambda_Y); && (3.12)\\
&X_i = \min(T_{X,i}, C_{X,i}); \quad Y_i = \min(T_{Y,i}, C_{Y,i});\\
&\Delta_{X,i} = 1(T_{X,i} \leq C_{X,i}); \quad \Delta_{Y,i} = 1(T_{Y,i} \leq C_{Y,i});
\end{aligned}
$$

where (3.9) was one of the following:

1. $\rho_S = 0$ implemented using $C_{U,V}(u, v) = uv$;
2. $\rho_S = 0.2$ implemented using Clayton's copula $C_{U,V}(u, v,\ \theta = 0.311)$;
3. $\rho_S = 0.2$ implemented using Frank's copula $C_{U,V}(u, v,\ \theta = 1.224)$;
4. $\rho_S = -0.2$ implemented using Frank's copula $C_{U,V}(u, v,\ \theta = -1.224)$.

We studied samples sizes of 100 and 200 and simulated two types of unbounded censoring, univariate ($C_X \equiv C_Y$) and bivariate ($C_X \perp C_Y$). The desired censoring proportion, P, was achieved by choosing parameters $\lambda = P/(1 - P)$. For unbounded univariate censoring, we used censoring proportions (P_X, P_Y) of $(0.3, 0.3)$ and $(0.7, 0.7)$. For unbounded bivariate censoring, the censoring proportions were $(0.3, 0.3)$, $(0.3, 0.7)$, and $(0.7, 0.7)$. For strict type I censoring, median survival time was used as the follow-up period. To simulate data under generalized type I censoring, we first simulated data under unbounded censoring and then censored all observations after the median survival time. As a result, the censoring proportions for generalized type I censoring were a little higher, $(0.56, 0.56)$ and $(0.73, 0.73)$.

Type I error rate and power of $\widehat{\rho}_{PSR}$ were compared to previously suggested methods:

1. Spearman's correlation estimator, $\widehat{\rho}_S^H$, (see Chapter 2) with bootstrap confidence intervals obtained with 1000 bootstrap samples;
2. Spearman-like statistic, S_n, proposed by Dabrowska (1986);
3. Log-rank statistic, T_n, proposed by Dabrowska (1986);

4. Log-rank statistic, U_n, as a special case of one of the martingale-based statistics proposed Shih and Louis (1996);

5. Weighted log-rank statistic, V_n, with optimal weights by Shih and Louis (1996).

Bias and root mean squared error (RMSE) of $\widehat{\rho}_{PSR}$ and $\widehat{\rho}_S^H$ were reported with respect to ρ_S. For $\widehat{\rho}_{PSR}$, bias and RMSE were also reported with respect to ρ_{PSR}. The 95%-confidence intervals of $\widehat{\rho}_{PSR}$ were computed using M-estimation with estimating equations proposed by Stute (1995).

3.4.1.2 Results

Figure 3.2 shows the type I error rate and power for $\widehat{\rho}_{PSR}$ compared to the other test statistics for $n = 200$. The type I error rate for $\widehat{\rho}_{PSR}$ tended to be at the nominal 0.05 level except at high levels of bivariate censoring. The power of $\widehat{\rho}_{PSR}$ tended to be competitive with that of the other test statistics. The power of our method tended to be similar or higher across the studied dependency structures compared to the previously suggested methods.

The bias and RMSE of $\widehat{\rho}_{PSR}$ and $\widehat{\rho}_S^H$ are reported in Figure 3.3. The Figure also displays the population parameters of ρ_{PSR} and ρ_S as horizontal dashed lines. For the studied simulation scenarios, $\widehat{\rho}_{PSR}$ was biased towards zero for ρ_S with larger bias observed for heavier censoring. Although in the presence of censoring $\widehat{\rho}_{PSR}$ is biased for ρ_S, the RMSE of $\widehat{\rho}_S^H$ was generally lower than that of $\widehat{\rho}_S^H$, as the variance of $\widehat{\rho}_{PSR}$ was typically much smaller than that of $\widehat{\rho}_S^H$.

Figures 3.8 and 3.9 in Appendix 3.A show simulation results for the sample size of 100: the type I error rate of $\widehat{\rho}_{PSR}$ is more elevated; the bias is very similar, and the RMSE is larger. Tables 3.2 and 3.3 in Appendix 3.A provide numeric values for type I error rate, power, bias, and RMSE for the sample sizes of 100 and 200.

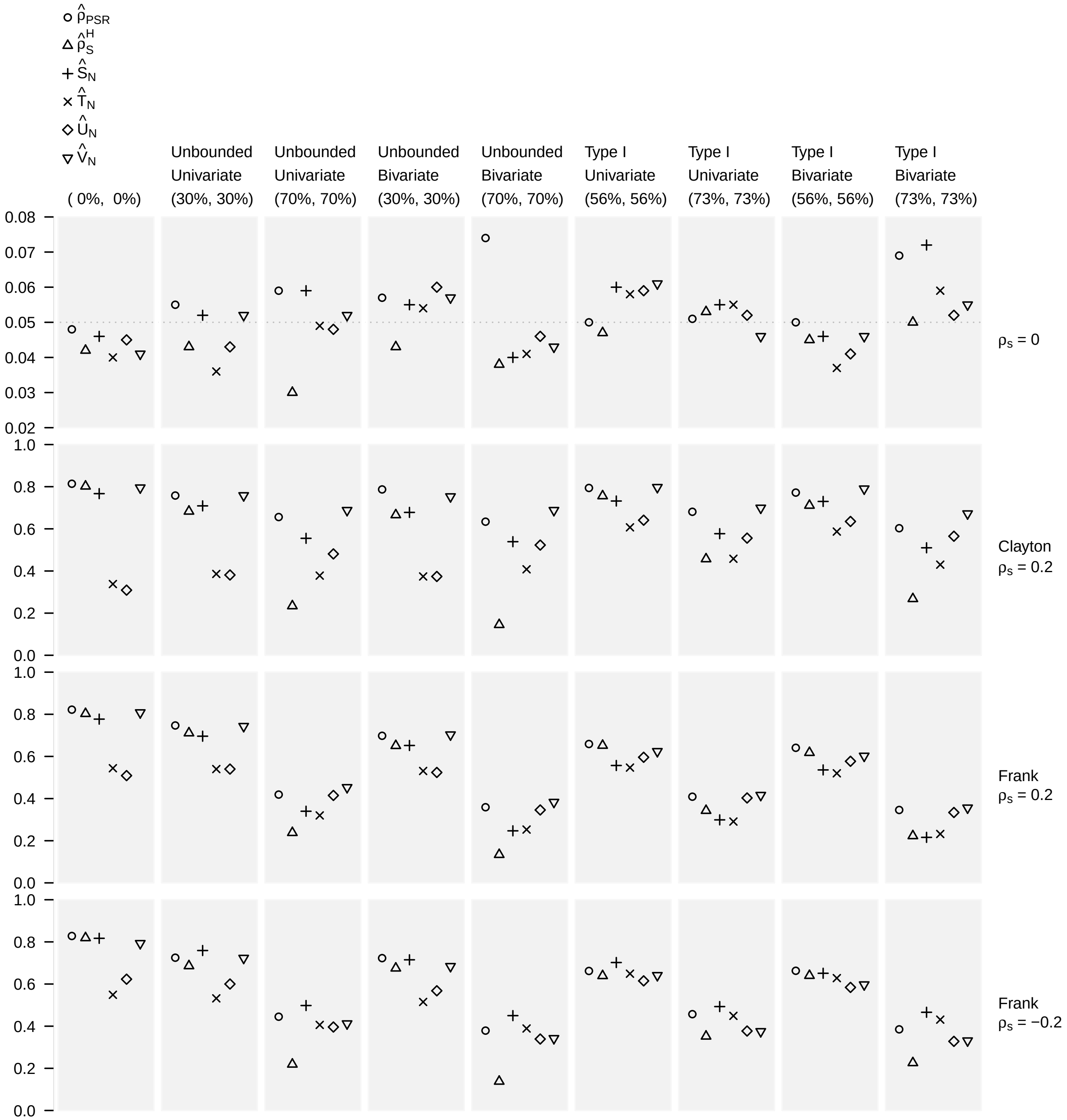

Figure 3.2: Type I error rate and power for unadjusted correlation and sample size of 200. The following methods are presented: 1) $\widehat{\rho}_{PSR}$ with Stute's estimating equations; 2) $\widehat{\rho}_S^H$ (see Chapter 2); 3) $\widehat{S}_N$ (Dabrowska, 1986); 4) $\widehat{T}_N$ (Dabrowska, 1986); 5) $\widehat{U}_N$ (Shih and Louis, 1996); 6) $\widehat{V}_N$ (Shih and Louis, 1996).

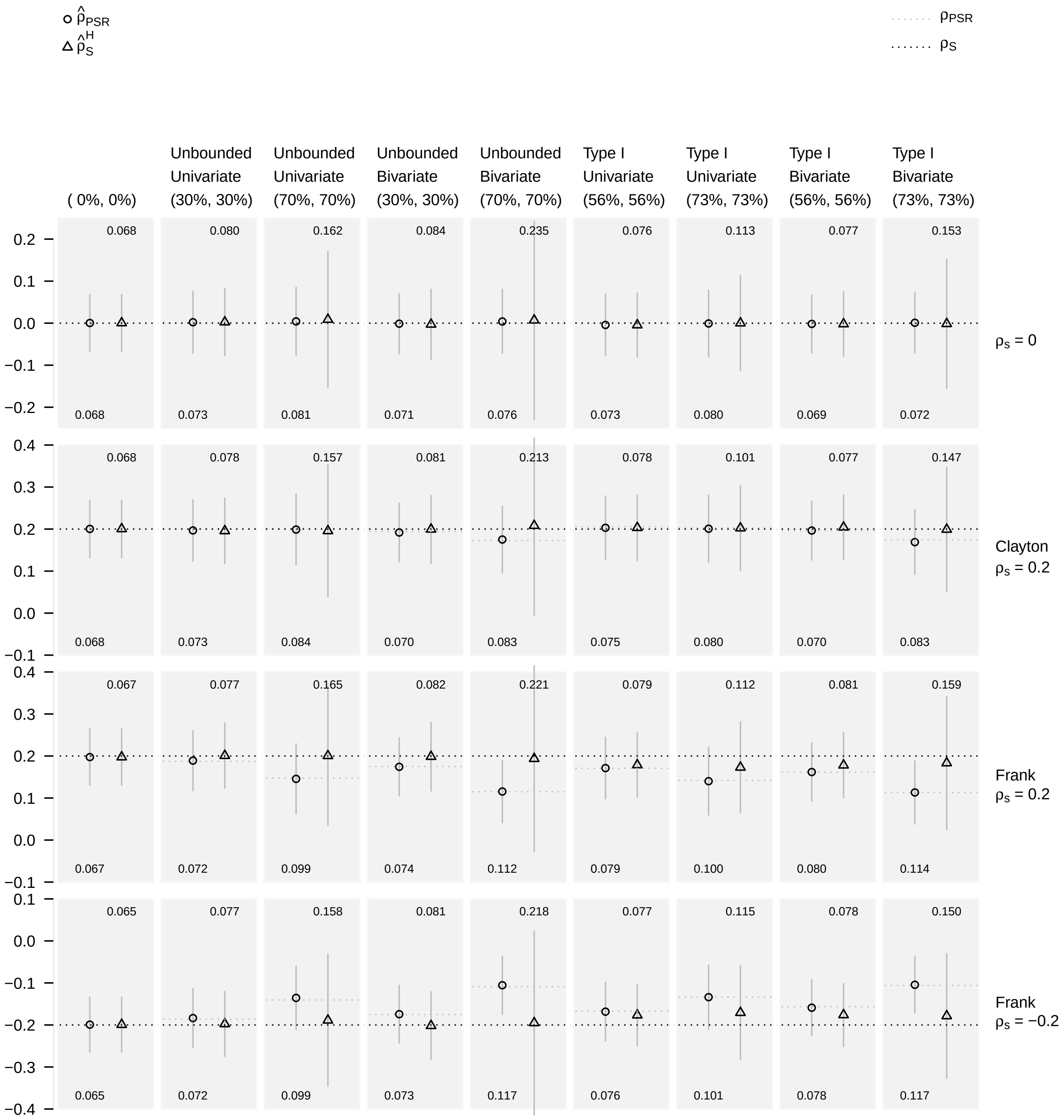

Figure 3.3: Point estimate ±SD for unadjusted correlation and sample size of 200. The following methods are presented: 1) $\widehat{\rho}_{PSR}$ with Stute's estimating equations; 2) $\widehat{\rho}_S^H$ (see Chapter 2). The numbers represent the corresponding RMSEs.

3.4.2 Partial Correlation

3.4.2.1 Simulation Set-up

To simulate partial correlation, we followed steps similar to Section 3.4.1.1, but instead of (3.10) and (3.11) we used

$$T_{X,i} = F_X^{-1}(U_i, \boldsymbol{Z}_i), \quad \text{where } F_X(x) = 1 - e^{-(e^{-\boldsymbol{Z}_i \boldsymbol{\beta}_{\boldsymbol{X}}})x};$$
$$T_{Y,i} = F_Y^{-1}(V_i, \boldsymbol{Z}_i), \quad \text{where } F_Y(y) = 1 - e^{-(e^{-\boldsymbol{Z}_i \boldsymbol{\beta}_{\boldsymbol{Y}}})y};$$

where $\boldsymbol{Z} = (Z_0, Z_1, Z_2)$, $Z_0 \equiv 1$, Z_1 was normally distributed with mean 0 and variance 1, and Z_2 was binary with $\Pr(Z_2 = 1) = \frac{1}{2}$, $\boldsymbol{\beta}_X = (1,\ 1,\ 0.5)$ and $\boldsymbol{\beta}_Y = (0, -1,\ 2)$. Times to censoring, C_X and C_Y, were simulated from exponential distributions with $\lambda = P/(1-P)$, where the censoring proportion P was chosen so that the average censoring proportion in both variates was either 0.3 or 0.7. For type I censoring, these censoring proportions were approximately 0.56 or 0.73. We estimated the partial correlation, as outlined in Section 3.3.2. We fit the following models for $F_X(\cdot|\boldsymbol{Z})$ and $F_X(\cdot|\boldsymbol{Z})$:

1. Log-normal survival model (misspecified model);

2. Exponential survival model (true model);

3. Cox proportional hazards model (true model) with variance estimated using partial likelihood score equations, and ignoring uncertainty in baseline hazard;

4. Cox proportional hazards model (true model) with variance estimated using full likelihood score equations.

We evaluated the performance of our method for the sample sizes of 100 and 200 under correctly and incorrectly specified models. The type I error rate, power, bias, and RMSE were reported with respect to $\rho_{S \cdot \boldsymbol{Z}}$.

3.4.2.2 Results

Figure 3.4 shows the type I error rate and power for the sample size of 200. The performance of partial $\widehat{\rho}_{PSR \cdot \boldsymbol{Z}}$ for all models was quite similar. In general, the type I error rate was near the nominal 0.05 level except with high rates of censoring. The power for all four models was practically the same, with the Cox proportional hazards models being slightly lower. Bias and RMSE were similar across all four models, even the misspecified log-normal model. As expected, $\widehat{\rho}_{PSR \cdot \boldsymbol{Z}}$ was further from $\rho_{S \cdot \boldsymbol{Z}}$ as the proportion censored increased. Results were somewhat similar for $n = 100$ (see

Figures 3.10 and 3.11, Appendix 3.A).

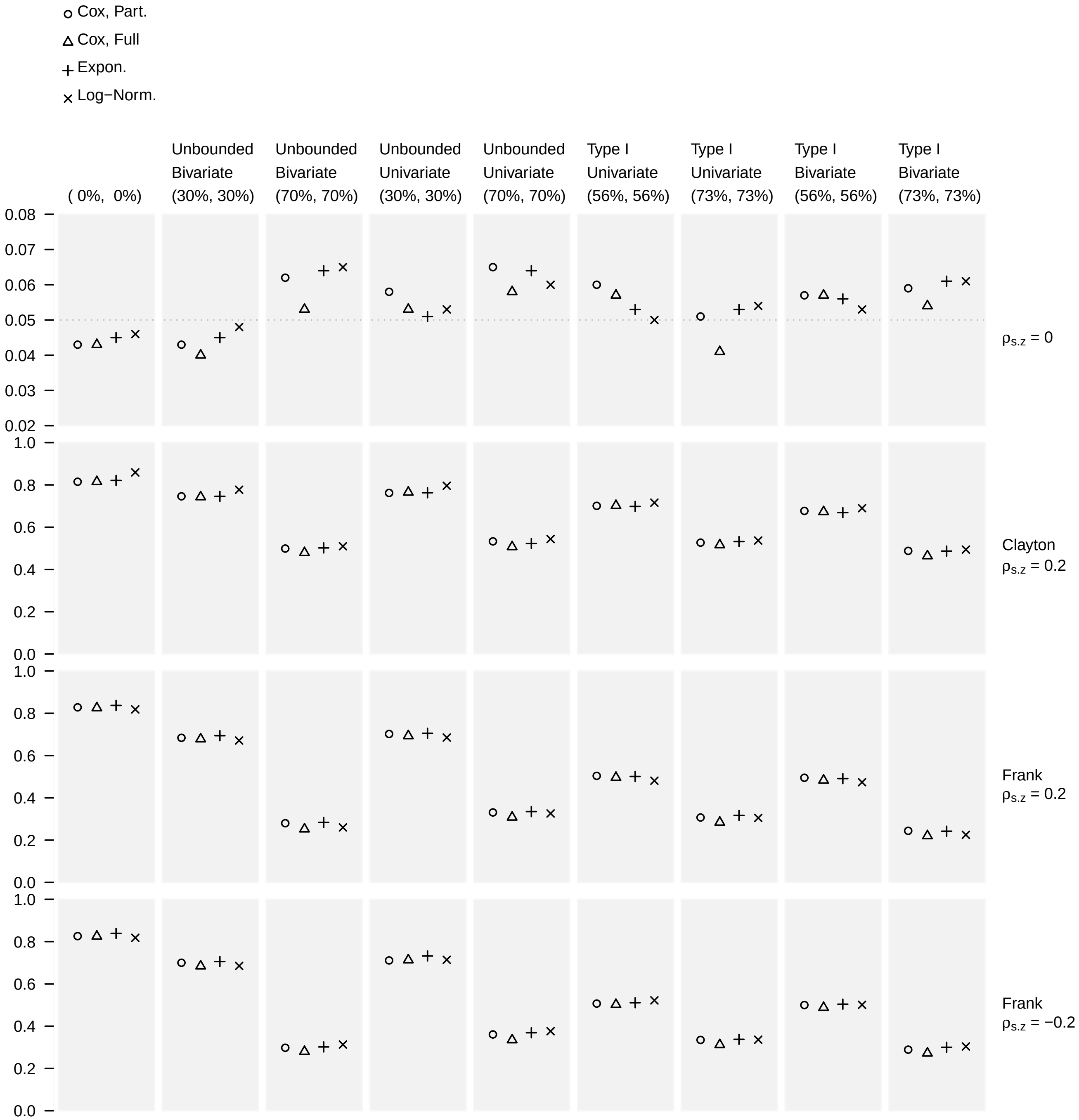

Figure 3.4: Type I error rate and power for partial correlation and sample size of 200.

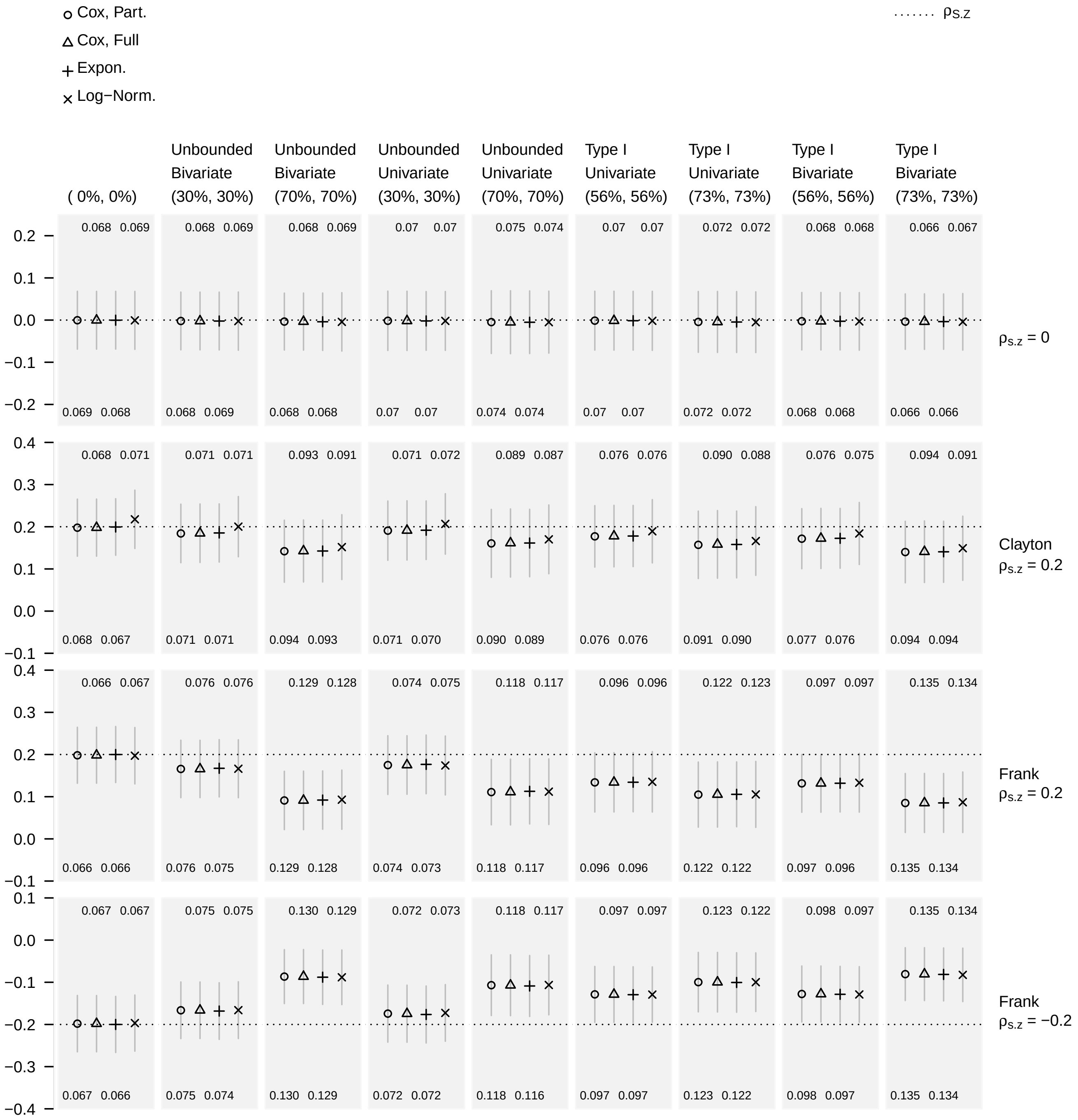

Figure 3.5: Point estimate $\pm SD$ for partial correlation and sample size of 200.

3.4.3 Conditional Correlation

3.4.3.1 Simulation Set-up

To simulate conditional correlation, we followed the steps similar to Section 3.4.2, but instead of a vector of covariates (Z_0, Z_1, Z_2), we simulated a single covariate Z, a uniformly distributed random variable with support in $[0, 3]$. For conditional correlation structure, we used Clayton's copula with the following $\rho_S(Z)$:

1. Constant correlation, $\rho_S(Z) = 0.2$;

2. Linear increasing correlation, $\rho_S(Z) = 0.1331437Z$;

3. Quadratic correlation (bell-shaped), $\rho_S(Z) = 0.001 + 0.48Z - 0.16Z^2$.

The parameters for univariate censoring were chosen in such a way that the proportions of censored events were approximately (0.3, 0.3). We studied the sample size of 500 simulated 1000 times. A Cox proportional hazards model was used to estimate PSRs. All cases were analyzed using correctly specified models. For bell-shaped conditional correlation, the linear regression models were fit with restricted cubic splines with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles. The performance of $\widehat{\rho}_{PSR}(Z)$ was evaluated visually by plotting the bias and coverage probability of $\widehat{\rho}_{PSR}(Z)$ for $\rho_S(Z)$ and $\rho_{PSR}(Z)$.

3.4.3.2 Results

Figure 3.6 shows the population parameters of $\rho_S(Z)$ (in black) and $\rho_{PSR}(Z)$ (in gray), bias, and coverage probability of $\widehat{\rho}_{PSR}(Z)$ for $\rho_S(Z)$ (in black) and $\rho_{PSR}(Z)$ (in gray) as functions of Z under $(0.3, 0.3)$ unbounded univariate censoring. Survival probabilities were modeled using Cox proportional hazards regression (true model) with variance estimated using full likelihood score equations. Although ρ_{PSR} is not the same as ρ_S, the bias of $\widehat{\rho}_{PSR}(Z)$ for $\rho_S(Z)$ was reasonable, and the coverage was mostly above 90%. Figure 3.12 in Appendix 3.A shows that our method performs very well in the absence of censoring, although a small bias is still observed.

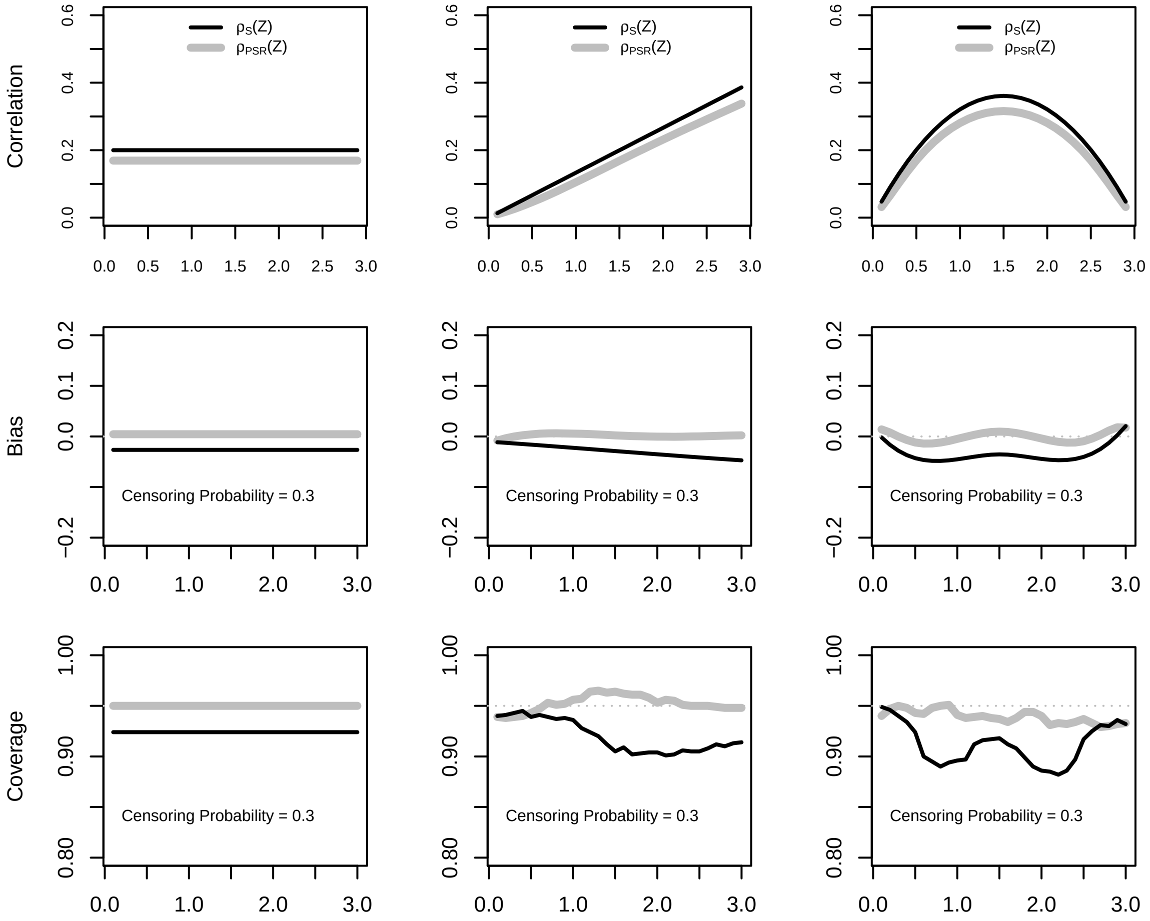

Figure 3.6: Top row: the population parameters for $\rho_S(Z)$ (in black) and $\rho_{PSR}(Z)$ (in gray) as functions of Z. Middle row: bias of $\widehat{\rho}_{PSR}(Z)$ for $\rho_S(Z)$ (in black) and $\widehat{\rho}_{PSR}(Z)$ for $\rho_{PSR}(Z)$ (in gray) as functions of Z. Bottom row: coverage probability of $\widehat{\rho}_{PSR}(Z)$ for $\rho_S(Z)$ (in black) and $\widehat{\rho}_{PSR}(Z)$ for $\rho_{PSR}(Z)$ (in gray) as functions of Z. Survival probabilities were modeled using Cox proportional hazards regression (true model) with variance estimated using full likelihood score equations. For bell-shaped conditional correlation, the linear models were fit with restricted cubic splines with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles. The data were simulated 1000 times with a sample size of 500. Unbounded univariate censoring of $(0.3, 0.3)$ was applied.

3.5 Application

We use our method to compute the correlation between the time from ART initiation to viral failure and the time from ART initiation to major regimen change among HIV positive persons living in Latin America and belonging to the Caribbean, Central, and South America Network for HIV epidemiology (CCASAnet), see McGowan et al. (2007). This dataset was also used in Chapter 2, and additional details can be found in there. In short, viral failure and regimen change tend to be correlated as failure to suppress the virus often triggers a regimen change. However, not all regimen changes are due to viral failure, and not all viral failures lead to a regimen change.

Variable definitions were as defined elsewhere (Cesar et al. (2015); Chapter 2). Censoring was univariate. Adults (18-year old or older when starting ART) and children (under 18) were analyzed separately. The analysis datasets included 374 children from Brazil and Peru, and 6691 adults from Brazil, Chile, Honduras, Mexico, and Peru. For adults and children, the median follow-up times were 4.1 years (ranging from 1 day to 18.2 years) and 7.4 years (ranging from 1 day to 19.1 years), respectively. For adults, about 28.6% had viral failure, 28.3% had regimen change, 58.6% had neither, and only 16.2% had both events. For children, 55.6% had viral failure, 43.9% had regimen change, 36.2% had neither, and 36.7% had both events.

Table 3.1 presents rank correlation estimates for various subgroups. It shows that $\widehat{\rho}_{PSR}$ was positive for all studied subgroups. The unadjusted correlations for adults and children were very similar, 0.32 (95% confidence interval [CI] 0.29, 0.35) and 0.32 (95% CI 0.22, 0.42), respectively. For adults, the correlations across sites were a little more variable with lower correlations in Chile (0.22) and Mexico (0.21) and higher correlations in Peru (0.40) and Brazil (0.33). Estimates based on $\widehat{\rho}_S^H$, the fully non-parametric approach that requires estimation of the bivariate survival surface (see Chapter 2), and $\widehat{\rho}_{IMI}$, the semi-parametric multiple imputation approach of Schemper et al. (2013), are included in Tables 3.1 for comparison. Estimates $\widehat{\rho}_{PSR}$ tended to be fairly similar to $\widehat{\rho}_S^H$ and $\widehat{\rho}_{IMI}$ except in a few cases. However, confidence intervals of $\widehat{\rho}_S^H$ largely overlapped with or contained the confidence intervals of $\widehat{\rho}_{PSR}$.

Table 3.1 also presents partial correlations computed using Cox proportional hazards models. Each model was adjusted for five covariates: gender, age at ART initiation, CD4 count at ART initiation (square-root transformed), viral load at ART initiation (log-transformed), and study site. Both CD4 count and viral load were included using restricted cubic splines with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles. The Table shows that after covariate adjustment, the correlation was generally similar to the unadjusted correlation, suggesting that the positive rank correlation between times to viral failure and regimen change were likely not due to the confounding by these covariates.

Figure 3.7 shows the rank correlation of time to viral failure and time to regimen change conditional on CD4 and age. Both conditional correlations were adjusted for sex, CD4 count at ART initiation, viral load at ART initiation, study site, and age at the time of first ART. The modeling techniques were the same as for partial correlations. To allow for greater flexibility of the correlation's functional form, the linear regression models of PSRs (see Section 3.3) included the variable of interest using restricted cubic spline with 3 knots (at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles) for children and 5 knots (at 0.05^{th}, 0.275^{th}, 0.5^{th}, 0.725^{th}, 0.95^{th} percentiles) for adults. We chose a smaller number of knots for children because of the smaller sample size. The figure shows that the partial-conditional correlation as a function of CD4 looks similar in children and adults. However, for children, the results could be less robust because of the smaller sample size. Note that for adults, there was a drop in the correlation between CD4 counts of 200 and 350 cell/mm^2, which are both clinical thresholds, reflecting potential medical decisions to change regimens based on these thresholds. The correlation conditional on age is about the same for children at the age right below 18 and adults at age 18 and remains more or less the same (around 0.3) up until the age of 60, where it starts declining towards zero.

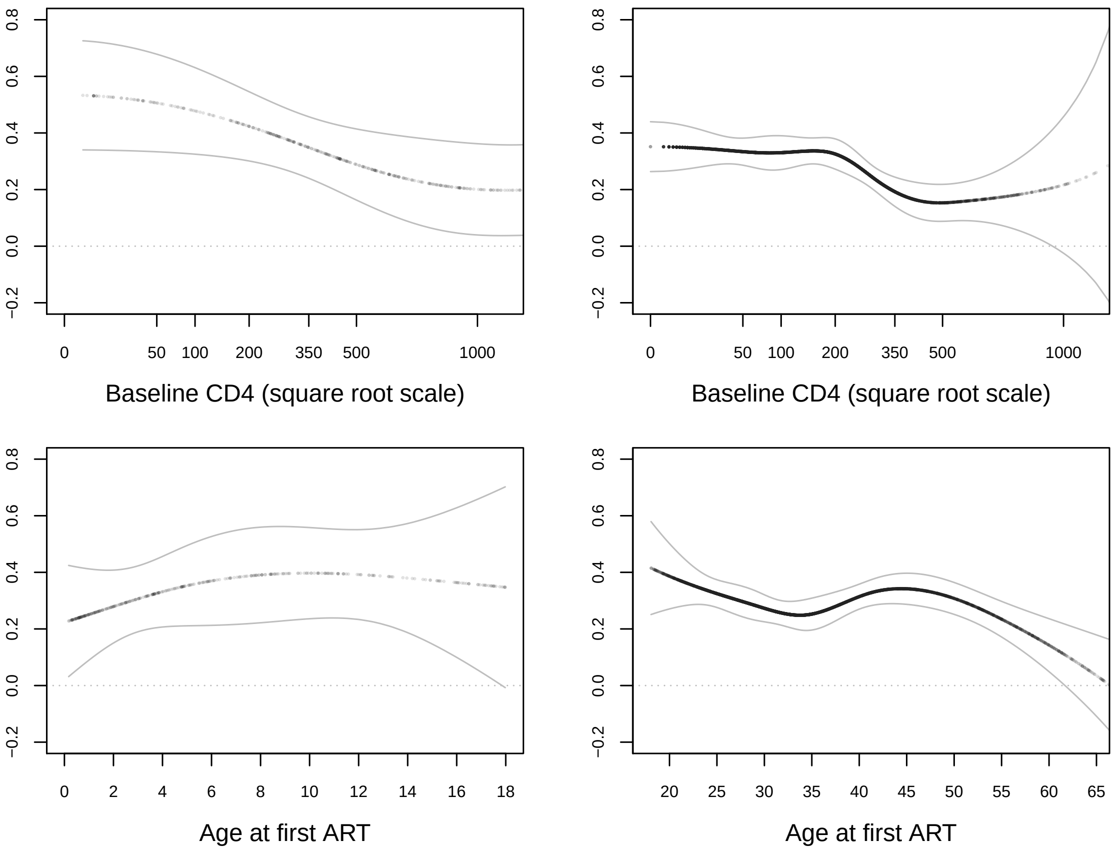

Figure 3.7: Partial-conditional correlation of PSRs between time to viral failure and time to regimen change computed as a function of CD4 count at ART initiation and age at the time of first ART. The correlations were adjusted for sex, CD4 count at ART initiation, viral load at ART initiation, study site, and age at the time of first ART. The left column shows results for children, the right column for adults. Cox proportional hazards regression was used to model survival probabilities. The variance was estimated using full likelihood score equations.

Table 3.1: Correlation between time to viral failure and time to regimen change measured using $\widehat{\rho}_{PSR}$, $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$, $\widehat{\rho}_S^H$ (see Chapter 2), and $\widehat{\rho}_{IMI}$ (Schemper et al., 2013). The confidence intervals for $\widehat{\rho}_{PSR}$ are computed using Stute's estimating equations, for $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$ using Cox full likelihood score equations, and for $\widehat{\rho}_S^H$ using bootstrap with 1000 bootstrap samples.

Group	**N**	$\widehat{\rho}_{PSR}$	$\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$	$\widehat{\rho}_S^H$	$\widehat{\rho}_{IMI}$
Adults	6691	0.32 (0.29, 0.35)	0.29 (0.26, 0.32)	0.35 (0.26, 0.43)	0.37 (0.32, 0.41)
Male adults	5185	0.31 (0.28, 0.34)	0.29 (0.25, 0.32)	0.32 (0.22, 0.42)	0.36 (0.30, 0.42)
Female adults	1506	0.35 (0.30, 0.40)	0.31 (0.25, 0.36)	0.45 (0.30, 0.58)	0.39 (0.31, 0.47)
Adults, Brazil	2313	0.33 (0.29, 0.37)	0.30 (0.26, 0.35)	0.45 (0.36, 0.53)	0.36 (0.29, 0.43)
Adults, Chile	1040	0.22 (0.16, 0.28)	0.21 (0.15, 0.28)	-0.06 (-0.22, 0.29)	0.26 (0.16, 0.36)
Adults, Honduras	138	0.28 (0.08, 0.47)	0.27 (0.07, 0.47)	-0.18 (-0.51, 0.48)	0.29 (-0.02, 0.55)
Adults, Mexico	975	0.21 (0.14, 0.28)	0.22 (0.15, 0.29)	0.52 (0.16, 0.75)	0.24 (0.10, 0.37)
Adults, Peru	2225	0.40 (0.35, 0.45)	0.39 (0.34, 0.44)	0.45 (0.37, 0.53)	0.55 (0.43, 0.65)
Children	374	0.32 (0.22, 0.42)	0.32 (0.22, 0.42)	0.36 (0.20, 0.53)	0.32 (0.20, 0.44)
Male children	191	0.26 (0.12, 0.40)	0.27 (0.12, 0.42)	0.33 (0.11, 0.52)	0.26 (0.06, 0.43)
Female children	183	0.38 (0.24, 0.52)	0.41 (0.28, 0.55)	0.34 (0.10, 0.63)	0.40 (0.22, 0.55)
Children, Brazil	301	0.31 (0.19, 0.43)	0.31 (0.19, 0.43)	0.36 (0.20, 0.53)	0.31 (0.17, 0.44)
Children, Peru	73	0.36 (0.15, 0.57)	0.40 (0.20, 0.61)	0.42 (0.15, 0.65)	0.41 (0.09, 0.65)

3.6 Discussion

We proposed a method of measuring correlation between right-censored variables by estimating the correlation of probability scale residuals. In the absence of censoring, our method equals Spearman's correlation; in the presence of censoring it approximates Spearman's. Our method has several advantages. It is based on ranks and not affected by extreme values or by monotonic transformations. Together with M-estimation, it provides an easy method to compute unadjusted correlation and confidence intervals for bivariate right-censored data. For moderate censoring and sample size of 200 or more, its power and type I error rate are comparable to previously suggested linear rank tests, while also providing an interpretable measure of association. The unadjusted correlation of PSRs is purely non-parametric because it does not assume the form of the marginal distributions and can be estimated using non-parametric Kaplan-Meier estimates for marginal cumulative distribution functions. Our method does not assume the dependency structure and does not require estimating the joint bivariate survival distribution. The correlation of PSRs can be easily extended to conditional, partial, or partial-conditional correlation. Although parametric assumptions have to be made when computing the partial and partial-conditional correlations, using Cox regression or other semi-parametric survival models maintain to some extent the non-parametric nature of Spearman's correlation; and partial correlations seemed to be quite robust to the choice of model. Our method can be used with continuous and discrete data.

The main limitation of our approach is that in the presence of censoring, the population parameter, ρ_{PSR}, depends on the censoring distribution. As seen from our simulations, the bias of $\widehat{\rho}_{PSR}$ for ρ_S increases with heavier censoring. However, with heavy censoring, all methods for estimating the bivariate correlation have limitations. Semi-parametric approaches rely on parametric assumptions to extrapolate. The other non-parametric approach proposed in Chapter 2 can be highly variable because it requires estimation of the bivariate survival distribution, and it will also be biased for ρ_S with type I censoring. Indeed, despite the bias of $\widehat{\rho}_{PSR}$, its mean squared error tended to be smaller than that of $\widehat{\rho}_S^H$ in our simulations. Minor challenges include computational complexity of estimating equations for the Kaplan-Meier curve (see Stute (1995)). Estimating partial and partial-conditional correlations with Cox regression and full likelihood can present some challenges because the dimensionality of M-estimation matrices depends on the sample size. Our simulations show, however, that assuming a fixed baseline hazards and using Cox partial likelihood results in very similar performance.

3.7 Acknowledgments

This work was partially supported by the US National Institutes of Health, grants R01 AI093234, and U01 AI069923. We thank the members of CCASAnet for allowing us to present their data and Cathy Jenkins for her help with constructing the analysis dataset. We also thank Dr. Shih for sharing her code related to the paper of Shih and Louis (1996).

3.8 Appendix 3.A

3.8.1 Figures

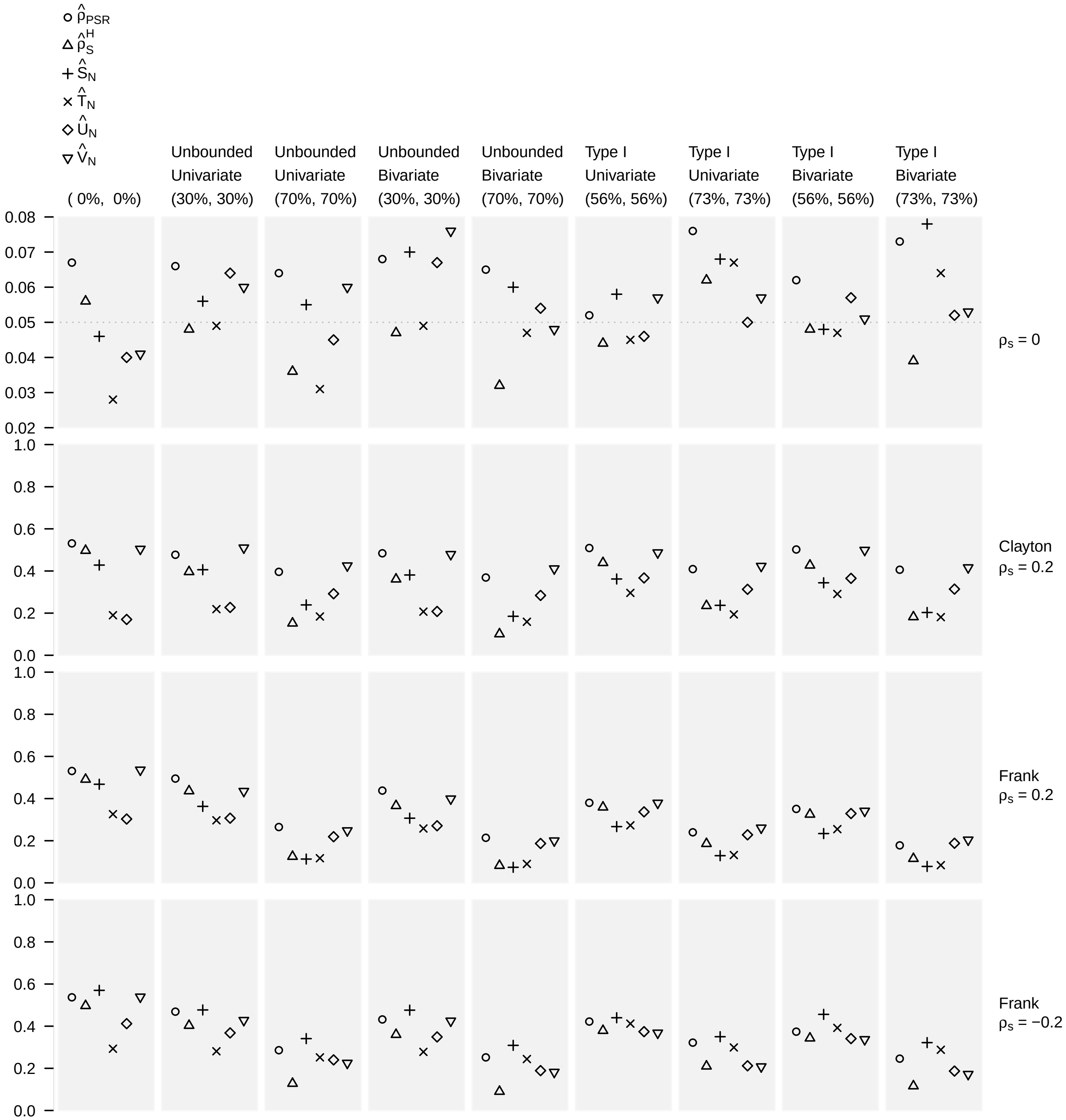

Figure 3.8: Type I error rate and power for unadjusted correlation and sample size of 200. The following methods are presented: 1) $\widehat{\rho}_{PSR}$ with Stute's estimating equations; 2) $\widehat{\rho}_S^H$ (see Chapter 2); 3) $\widehat{S}_N$ (Dabrowska, 1986); 4) $\widehat{T}_N$ (Dabrowska, 1986); 5) $\widehat{U}_N$ (Shih and Louis, 1996); 6) $\widehat{V}_N$ (Shih and Louis, 1996).

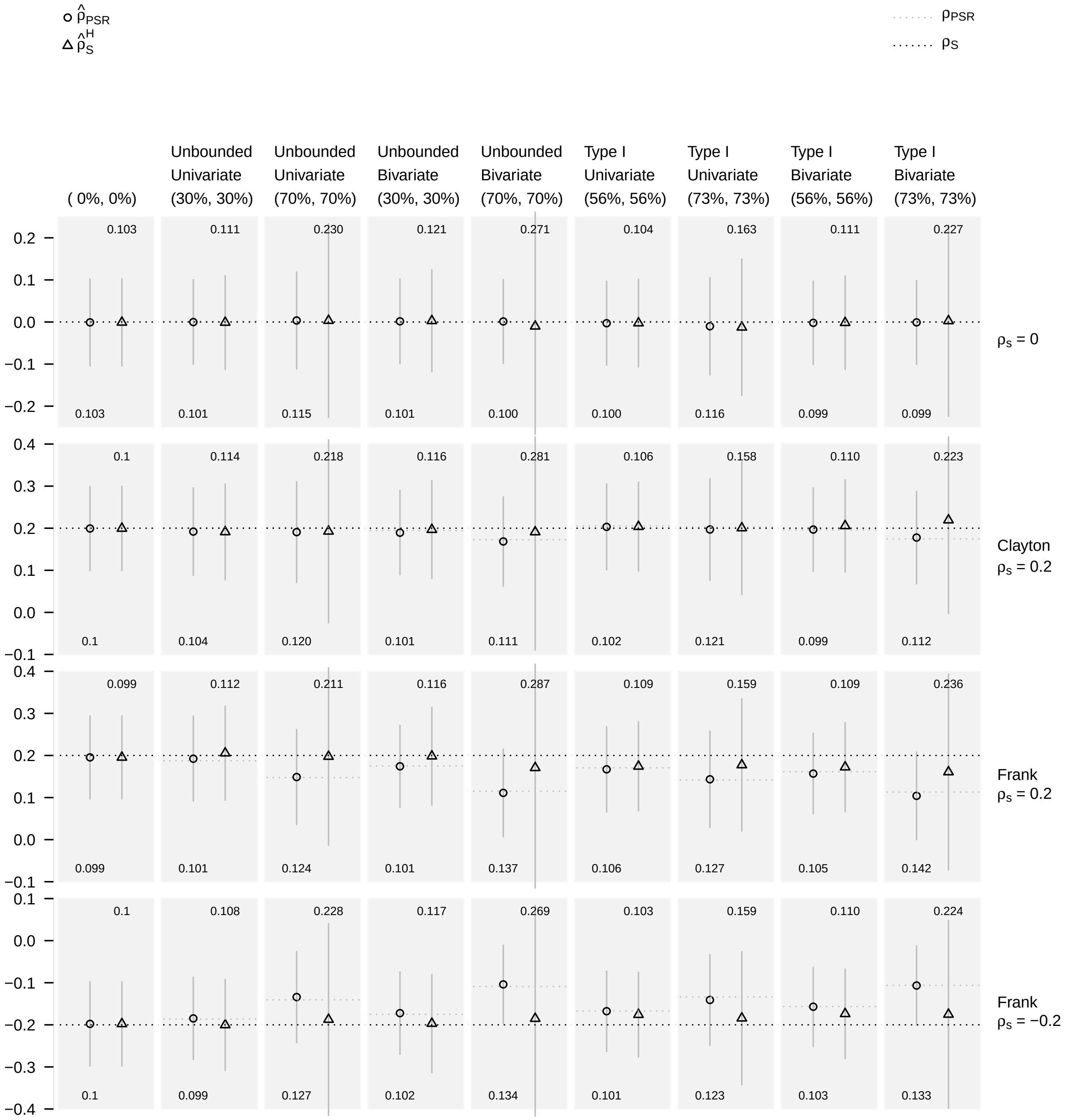

Figure 3.9: Point estimate ±SD for unadjusted correlation and sample size of 100. The following methods are presented: 1) $\widehat{\rho}_{PSR}$ with Stute's estimating equations; 2) $\widehat{\rho}_S^H$ (see Chapter 2). The numbers represent the corresponding RMSEs.

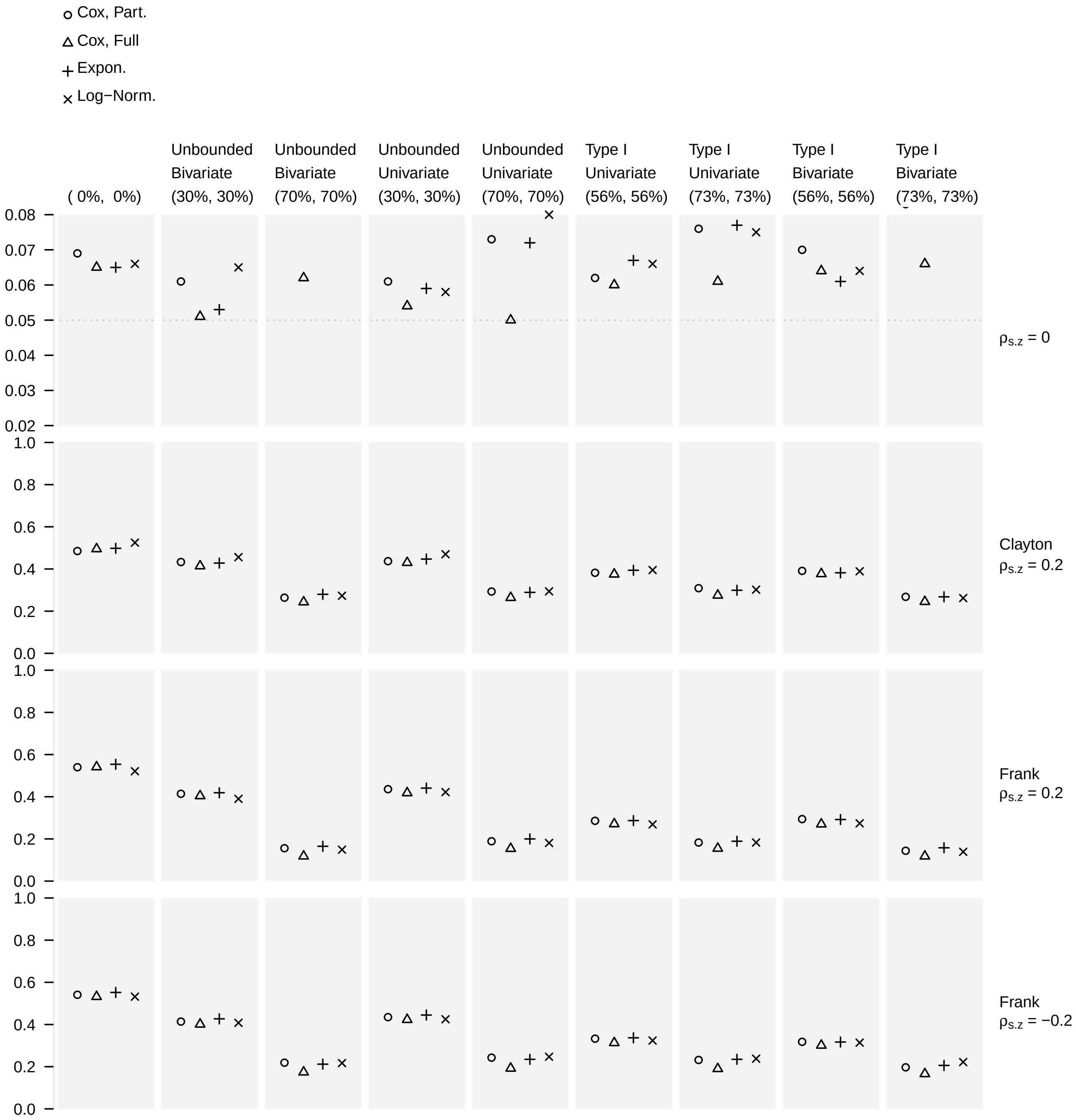

Figure 3.10: Type I error rate and power for partial correlation and sample size of 100.

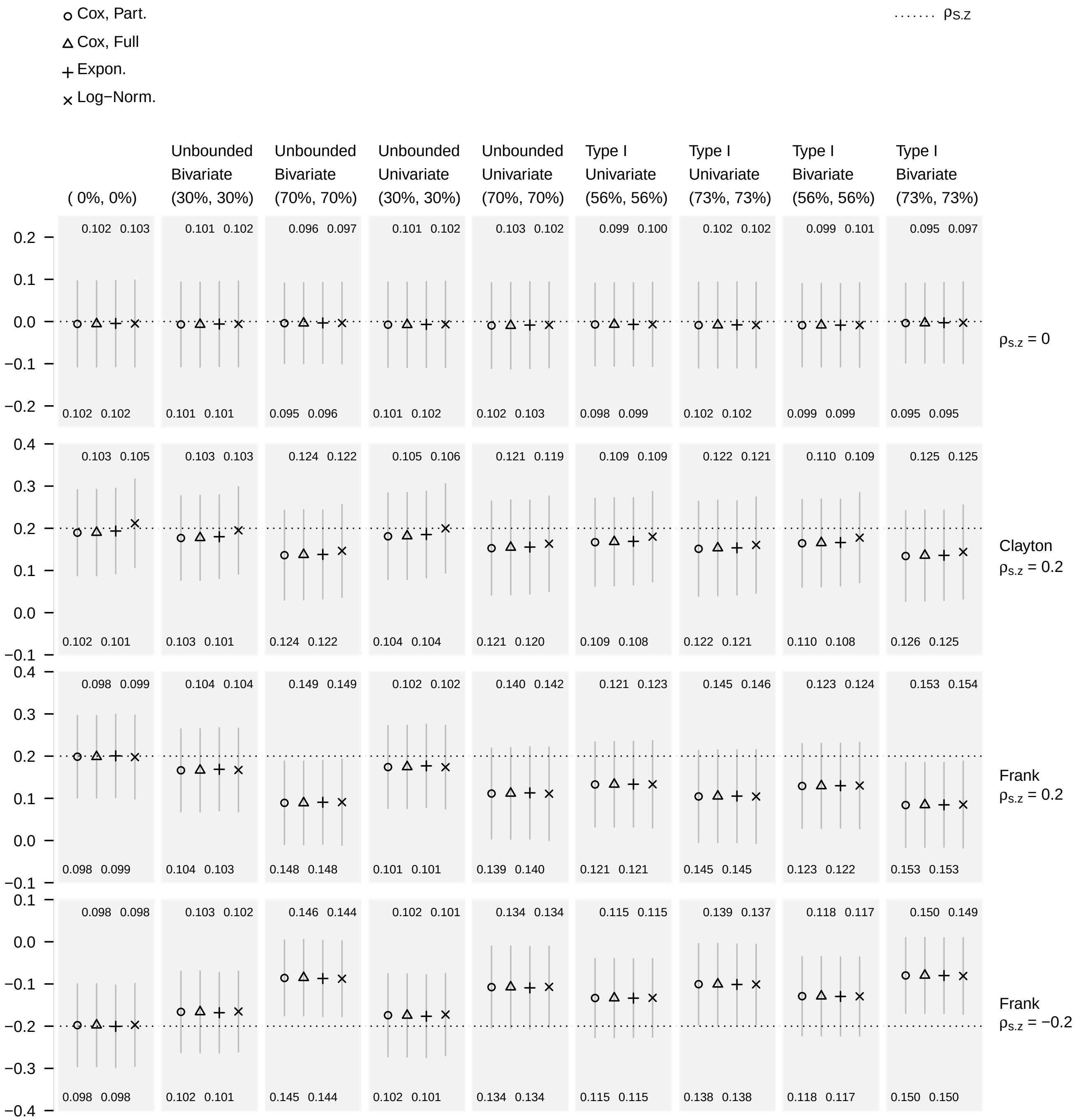

Figure 3.11: Point estimate $\pm SD$ for partial correlation and sample size of 100. The numbers above and below the point estimates are RMSEs.

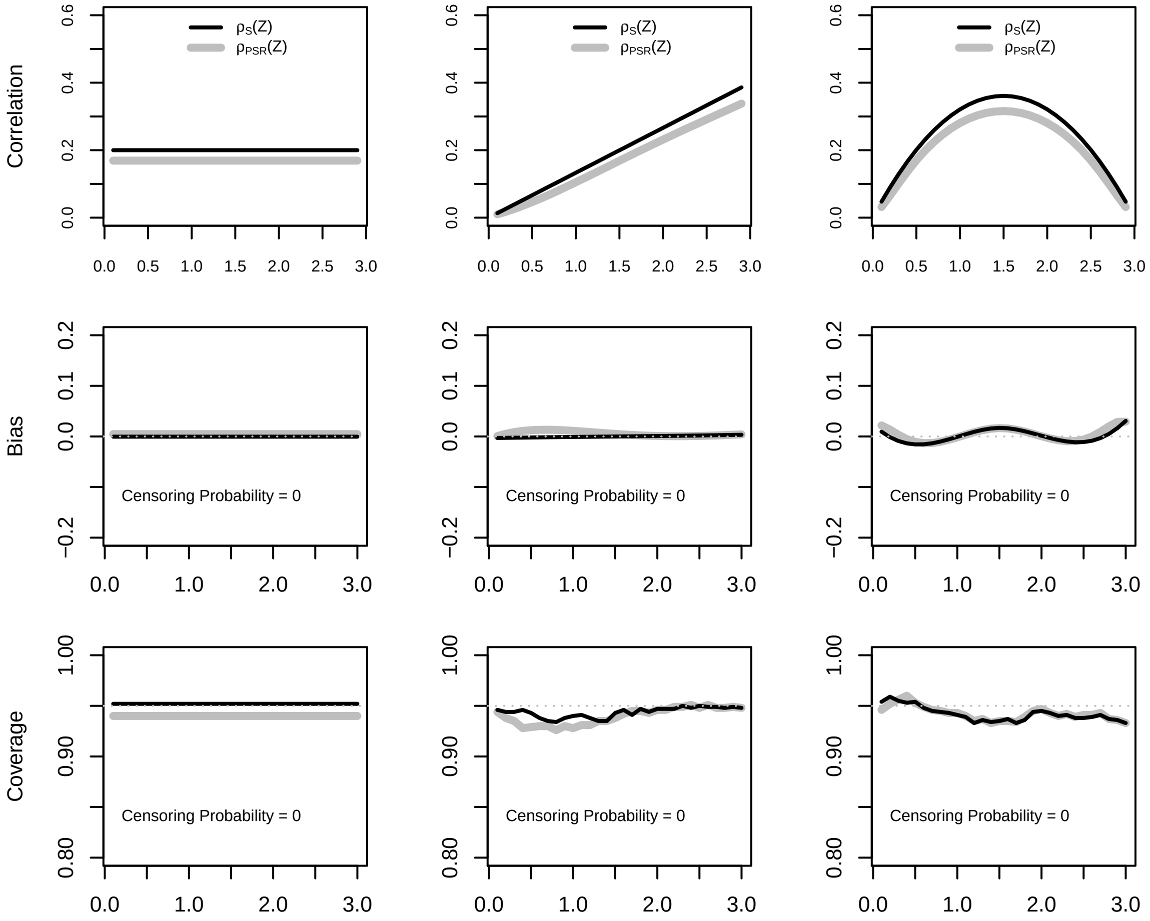

Figure 3.12: Top row: the population parameters for $\rho_S(Z)$ (in black) and $\rho_{PSR}(Z)$ (in gray) as functions of Z. Middle row: bias of $\widehat{\rho}_{PSR}$ for ρ_S (in black) and $\widehat{\rho}_{PSR}$ for ρ_{PSR} (in gray) as functions of Z. Bottom row: coverage probability of $\widehat{\rho}_{PSR}$ for ρ_S (in black) and $\widehat{\rho}_{PSR}$ for ρ_{PSR} (in gray) as functions of Z. Survival probabilities were modeled using Cox proportional hazards regression (true model) with variance estimated using full likelihood score equations. For bell-shaped conditional correlation, linear models were fit with restricted cubic splines with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles. The data was simulated 1000 times with a sample size of 500. No censoring was applied.

3.8.2 Tables

Table 3.2: Type I error rate and power for unadjusted correlation.

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
100	No Censoring	$\widehat{\rho}_{PSR}$; Stute	0.067	0.531	0.531	0.537
		$\widehat{\rho}_S^H$	0.056	0.497	0.491	0.497
		$\widehat{S}_N$; Dabrowska	0.046	0.428	0.468	0.570
		$\widehat{T}_N$; Dabrowska	0.028	0.190	0.326	0.293
		$\widehat{U}_N$; Shih, Louis	0.040	0.170	0.303	0.412
		$\widehat{V}_N$; Shih, Louis	0.041	0.504	0.536	0.539
	$C_1 \equiv C_2,\ 30\%$	$\widehat{\rho}_{PSR}$; Stute	0.066	0.477	0.495	0.469
		$\widehat{\rho}_S^H$	0.048	0.396	0.436	0.403
		$\widehat{S}_N$; Dabrowska	0.056	0.406	0.363	0.477
		$\widehat{T}_N$; Dabrowska	0.049	0.219	0.297	0.281
		$\widehat{U}_N$; Shih, Louis	0.064	0.227	0.307	0.368
		$\widehat{V}_N$; Shih, Louis	0.060	0.510	0.435	0.428
	$C_1 \equiv C_2,\ 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.064	0.396	0.265	0.286
		$\widehat{\rho}_S^H$	0.036	0.152	0.125	0.128
		$\widehat{S}_N$; Dabrowska	0.055	0.239	0.113	0.341
		$\widehat{T}_N$; Dabrowska	0.031	0.184	0.117	0.252
		$\widehat{U}_N$; Shih, Louis	0.045	0.292	0.219	0.240
		$\widehat{V}_N$; Shih, Louis	0.060	0.425	0.248	0.225
	$C_1 \perp C_2,\ 30\%, 30\%$	$\widehat{\rho}_{PSR}$; Stute	0.068	0.484	0.438	0.432
		$\widehat{\rho}_S^H$	0.047	0.361	0.366	0.360
		$\widehat{S}_N$; Dabrowska	0.070	0.381	0.307	0.476
		$\widehat{T}_N$; Dabrowska	0.049	0.207	0.258	0.278
		$\widehat{U}_N$; Shih, Louis	0.067	0.208	0.271	0.349
		$\widehat{V}_N$; Shih, Louis	0.076	0.479	0.399	0.425
	$C_1 \perp C_2,\ 30\%, 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.056	0.444	0.319	0.284
		$\widehat{\rho}_S^H$	0.028	0.158	0.176	0.140
		$\widehat{S}_N$; Dabrowska	0.050	0.294	0.166	0.372
		$\widehat{T}_N$; Dabrowska	0.037	0.213	0.166	0.247
		$\widehat{U}_N$; Shih, Louis	0.043	0.253	0.210	0.271
		$\widehat{V}_N$; Shih, Louis	0.045	0.439	0.277	0.287
	$C_1 \perp C_2,\ 70\%, 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.065	0.369	0.214	0.252
		$\widehat{\rho}_S^H$	0.032	0.101	0.082	0.090
		$\widehat{S}_N$; Dabrowska	0.060	0.185	0.074	0.309
		$\widehat{T}_N$; Dabrowska	0.047	0.159	0.090	0.244
		$\widehat{U}_N$; Shih, Louis	0.054	0.284	0.187	0.189
		$\widehat{V}_N$; Shih, Louis	0.048	0.411	0.200	0.182
	$C_1 \equiv C_2,\ 30\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.052	0.509	0.380	0.422
		$\widehat{\rho}_S^H$	0.044	0.439	0.359	0.379
		$\widehat{S}_N$; Dabrowska	0.058	0.362	0.267	0.440
		$\widehat{T}_N$; Dabrowska	0.045	0.296	0.273	0.412
		$\widehat{U}_N$; Shih, Louis	0.046	0.367	0.337	0.374
		$\widehat{V}_N$; Shih, Louis	0.057	0.487	0.379	0.368
	$C_1 \equiv C_2,\ 70\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.076	0.409	0.240	0.322
		$\widehat{\rho}_S^H$	0.062	0.235	0.186	0.210
		$\widehat{S}_N$; Dabrowska	0.068	0.237	0.129	0.350
		$\widehat{T}_N$; Dabrowska	0.067	0.194	0.132	0.299
		$\widehat{U}_N$; Shih, Louis	0.050	0.313	0.228	0.212
		$\widehat{V}_N$; Shih, Louis	0.057	0.423	0.261	0.208
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.062	0.502	0.351	0.374
		$\widehat{\rho}_S^H$	0.048	0.427	0.325	0.343
		$\widehat{S}_N$; Dabrowska	0.048	0.344	0.234	0.456
		$\widehat{T}_N$; Dabrowska	0.047	0.291	0.255	0.392
		$\widehat{U}_N$; Shih, Louis	0.057	0.365	0.329	0.341
		$\widehat{V}_N$; Shih, Louis	0.051	0.499	0.341	0.337
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.073	0.406	0.178	0.246
		$\widehat{\rho}_S^H$	0.039	0.182	0.115	0.116
		$\widehat{S}_N$; Dabrowska	0.078	0.203	0.078	0.322
		$\widehat{T}_N$; Dabrowska	0.064	0.181	0.084	0.288
		$\widehat{U}_N$; Shih, Louis	0.052	0.314	0.188	0.187

Continued on next page

Table 3.2: Type I error rate and power for unadjusted correlation. ***(continued)***

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
		$\widehat{V}_N$; Shih, Louis	0.053	0.416	0.204	0.172
200	No Censoring	$\widehat{\rho}_{PSR}$; Stute	0.048	0.814	0.822	0.828
		$\widehat{\rho}_S^H$	0.042	0.802	0.803	0.819
		$\widehat{S}_N$; Dabrowska	0.046	0.767	0.777	0.817
		$\widehat{T}_N$; Dabrowska	0.040	0.338	0.544	0.549
		$\widehat{U}_N$; Shih, Louis	0.045	0.309	0.509	0.623
		$\widehat{V}_N$; Shih, Louis	0.041	0.795	0.808	0.793
	$C_1 \equiv C_2,\ 30\%$	$\widehat{\rho}_{PSR}$; Stute	0.055	0.758	0.747	0.725
		$\widehat{\rho}_S^H$	0.043	0.683	0.711	0.686
		$\widehat{S}_N$; Dabrowska	0.052	0.709	0.696	0.759
		$\widehat{T}_N$; Dabrowska	0.036	0.386	0.540	0.532
		$\widehat{U}_N$; Shih, Louis	0.043	0.381	0.540	0.600
		$\widehat{V}_N$; Shih, Louis	0.052	0.758	0.743	0.723
	$C_1 \equiv C_2,\ 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.059	0.656	0.419	0.445
		$\widehat{\rho}_S^H$	0.030	0.234	0.238	0.219
		$\widehat{S}_N$; Dabrowska	0.059	0.555	0.340	0.498
		$\widehat{T}_N$; Dabrowska	0.049	0.378	0.320	0.406
		$\widehat{U}_N$; Shih, Louis	0.048	0.481	0.415	0.396
		$\widehat{V}_N$; Shih, Louis	0.052	0.688	0.453	0.412
	$C_1 \perp C_2,\ 30\%, 30\%$	$\widehat{\rho}_{PSR}$; Stute	0.057	0.787	0.698	0.723
		$\widehat{\rho}_S^H$	0.043	0.666	0.651	0.675
		$\widehat{S}_N$; Dabrowska	0.055	0.678	0.652	0.715
		$\widehat{T}_N$; Dabrowska	0.054	0.374	0.531	0.515
		$\widehat{U}_N$; Shih, Louis	0.060	0.374	0.524	0.568
		$\widehat{V}_N$; Shih, Louis	0.057	0.753	0.703	0.684
	$C_1 \perp C_2,\ 30\%, 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.066	0.699	0.512	0.530
		$\widehat{\rho}_S^H$	0.041	0.266	0.284	0.260
		$\widehat{S}_N$; Dabrowska	0.044	0.643	0.414	0.599
		$\widehat{T}_N$; Dabrowska	0.036	0.430	0.370	0.443
		$\widehat{U}_N$; Shih, Louis	0.043	0.460	0.403	0.462
		$\widehat{V}_N$; Shih, Louis	0.050	0.735	0.514	0.533
	$C_1 \perp C_2,\ 70\%, 70\%$	$\widehat{\rho}_{PSR}$; Stute	0.074	0.634	0.359	0.379
		$\widehat{\rho}_S^H$	0.038	0.145	0.134	0.138
		$\widehat{S}_N$; Dabrowska	0.040	0.539	0.247	0.450
		$\widehat{T}_N$; Dabrowska	0.041	0.408	0.253	0.389
		$\widehat{U}_N$; Shih, Louis	0.046	0.523	0.346	0.339
		$\widehat{V}_N$; Shih, Louis	0.043	0.688	0.383	0.342
	$C_1 \equiv C_2,\ 30\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.050	0.794	0.659	0.662
		$\widehat{\rho}_S^H$	0.047	0.756	0.652	0.639
		$\widehat{S}_N$; Dabrowska	0.060	0.732	0.557	0.702
		$\widehat{T}_N$; Dabrowska	0.058	0.607	0.547	0.649
		$\widehat{U}_N$; Shih, Louis	0.059	0.641	0.596	0.615
		$\widehat{V}_N$; Shih, Louis	0.061	0.797	0.624	0.641
	$C_1 \equiv C_2,\ 70\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.051	0.681	0.409	0.457
		$\widehat{\rho}_S^H$	0.053	0.457	0.343	0.352
		$\widehat{S}_N$; Dabrowska	0.055	0.577	0.299	0.493
		$\widehat{T}_N$; Dabrowska	0.055	0.458	0.291	0.449
		$\widehat{U}_N$; Shih, Louis	0.052	0.556	0.403	0.377
		$\widehat{V}_N$; Shih, Louis	0.046	0.699	0.416	0.375
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.050	0.772	0.641	0.663
		$\widehat{\rho}_S^H$	0.045	0.711	0.618	0.640
		$\widehat{S}_N$; Dabrowska	0.046	0.730	0.536	0.651
		$\widehat{T}_N$; Dabrowska	0.037	0.587	0.520	0.628
		$\widehat{U}_N$; Shih, Louis	0.041	0.635	0.577	0.584
		$\widehat{V}_N$; Shih, Louis	0.046	0.790	0.602	0.597
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau = median$	$\widehat{\rho}_{PSR}$; Stute	0.069	0.603	0.346	0.385
		$\widehat{\rho}_S^H$	0.050	0.268	0.223	0.226
		$\widehat{S}_N$; Dabrowska	0.072	0.510	0.216	0.466
		$\widehat{T}_N$; Dabrowska	0.059	0.430	0.232	0.431
		$\widehat{U}_N$; Shih, Louis	0.052	0.565	0.334	0.328
		$\widehat{V}_N$; Shih, Louis	0.055	0.672	0.356	0.331

Table 3.3: Bias and RMSE for unadjusted correlation.

N	Censoring	Method		Clayton	Frank	Frank
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
100	No Censoring	$\widehat{\rho_{PSR}}$	-0.001 [0.103]	-0.001 [0.100]	-0.005 [0.099]	0.002 [0.100]
		$\widehat{\rho_S^H}$	-0.001 [0.103]	-0.001 [0.100]	-0.005 [0.099]	0.002 [0.100]
	$C_1 \equiv C_2,\ 30\%$	$\widehat{\rho_{PSR}}$	0.000 [0.101]	-0.008 [0.104]	-0.008 [0.101]	0.015 [0.099]
		$\widehat{\rho_S^H}$	-0.001 [0.111]	-0.009 [0.114]	0.006 [0.112]	0.000 [0.108]
	$C_1 \equiv C_2,\ 70\%$	$\widehat{\rho_{PSR}}$	0.004 [0.115]	-0.009 [0.120]	-0.051 [0.124]	0.066 [0.127]
		$\widehat{\rho_S^H}$	0.003 [0.230]	-0.007 [0.218]	-0.003 [0.211]	0.013 [0.228]
	$C_1 \perp C_2,\ 30\%, 30\%$	$\widehat{\rho_{PSR}}$	0.002 [0.101]	-0.011 [0.101]	-0.026 [0.101]	0.028 [0.102]
		$\widehat{\rho_S^H}$	0.003 [0.121]	-0.003 [0.116]	-0.002 [0.116]	0.003 [0.117]
	$C_1 \perp C_2,\ 30\%, 70\%$	$\widehat{\rho_{PSR}}$	0.006 [0.097]	-0.022 [0.104]	-0.060 [0.116]	0.070 [0.117]
		$\widehat{\rho_S^H}$	0.015 [0.191]	-0.012 [0.202]	0.002 [0.198]	0.013 [0.194]
	$C_1 \perp C_2,\ 70\%, 70\%$	$\widehat{\rho_{PSR}}$	0.001 [0.100]	-0.031 [0.111]	-0.089 [0.137]	0.096 [0.134]
		$\widehat{\rho_S^H}$	-0.010 [0.271]	-0.009 [0.281]	-0.029 [0.287]	0.015 [0.269]
	$C_1 \equiv C_2,\ 30\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.003 [0.100]	0.003 [0.102]	-0.033 [0.106]	0.032 [0.101]
		$\widehat{\rho_S^H}$	-0.003 [0.104]	0.004 [0.106]	-0.026 [0.109]	0.024 [0.103]
	$C_1 \equiv C_2,\ 70\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.010 [0.116]	-0.003 [0.121]	-0.057 [0.127]	0.059 [0.123]
		$\widehat{\rho_S^H}$	-0.013 [0.163]	0.000 [0.158]	-0.023 [0.159]	0.016 [0.159]
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.002 [0.099]	-0.003 [0.099]	-0.043 [0.105]	0.043 [0.103]
		$\widehat{\rho_S^H}$	-0.002 [0.111]	0.006 [0.110]	-0.028 [0.109]	0.026 [0.110]
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.001 [0.099]	-0.022 [0.112]	-0.096 [0.142]	0.093 [0.133]
		$\widehat{\rho_S^H}$	0.003 [0.227]	0.019 [0.223]	-0.039 [0.236]	0.025 [0.224]
200	No Censoring	$\widehat{\rho_{PSR}}$	0.000 [0.068]	0.000 [0.068]	-0.003 [0.067]	0.001 [0.065]
		$\widehat{\rho_S^H}$	0.000 [0.068]	0.000 [0.068]	-0.003 [0.067]	0.001 [0.065]
	$C_1 \equiv C_2,\ 30\%$	$\widehat{\rho_{PSR}}$	0.002 [0.073]	-0.003 [0.073]	-0.011 [0.072]	0.016 [0.072]
		$\widehat{\rho_S^H}$	0.002 [0.080]	-0.004 [0.078]	0.001 [0.077]	0.002 [0.077]
	$C_1 \equiv C_2,\ 70\%$	$\widehat{\rho_{PSR}}$	0.004 [0.081]	-0.001 [0.084]	-0.054 [0.099]	0.064 [0.099]
		$\widehat{\rho_S^H}$	0.009 [0.162]	-0.004 [0.157]	0.001 [0.165]	0.011 [0.158]
	$C_1 \perp C_2,\ 30\%, 30\%$	$\widehat{\rho_{PSR}}$	-0.001 [0.071]	-0.008 [0.070]	-0.026 [0.074]	0.026 [0.073]
		$\widehat{\rho_S^H}$	-0.003 [0.084]	-0.001 [0.081]	-0.001 [0.082]	-0.002 [0.081]
	$C_1 \perp C_2,\ 30\%, 70\%$	$\widehat{\rho_{PSR}}$	0.000 [0.073]	-0.022 [0.077]	-0.062 [0.093]	0.062 [0.092]
		$\widehat{\rho_S^H}$	0.006 [0.145]	-0.010 [0.145]	-0.005 [0.145]	0.007 [0.143]
	$C_1 \perp C_2,\ 70\%, 70\%$	$\widehat{\rho_{PSR}}$	0.004 [0.076]	-0.025 [0.083]	-0.085 [0.112]	0.095 [0.117]
		$\widehat{\rho_S^H}$	0.007 [0.235]	0.008 [0.213]	-0.006 [0.221]	0.005 [0.218]
	$C_1 \equiv C_2,\ 30\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.004 [0.073]	0.003 [0.075]	-0.029 [0.079]	0.032 [0.076]
		$\widehat{\rho_S^H}$	-0.005 [0.076]	0.003 [0.078]	-0.021 [0.079]	0.023 [0.077]
	$C_1 \equiv C_2,\ 70\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.001 [0.080]	0.001 [0.080]	-0.060 [0.100]	0.066 [0.101]
		$\widehat{\rho_S^H}$	0.000 [0.113]	0.002 [0.101]	-0.027 [0.112]	0.029 [0.115]
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	-0.002 [0.069]	-0.003 [0.070]	-0.038 [0.080]	0.041 [0.078]
		$\widehat{\rho_S^H}$	-0.002 [0.077]	0.004 [0.077]	-0.022 [0.081]	0.024 [0.078]
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau,\ type\ I$	$\widehat{\rho_{PSR}}$	0.001 [0.072]	-0.031 [0.083]	-0.087 [0.114]	0.096 [0.117]
		$\widehat{\rho_S^H}$	-0.002 [0.153]	-0.001 [0.147]	-0.017 [0.159]	0.021 [0.150]

Table 3.4: Type I error rate and power for partial correlation.

N	Censoring	Method		Clayton	Frank	Frank
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
100	No Censoring	Cox; Partial L.	0.069	0.485	0.540	0.541
		Cox; Full L.	0.065	0.495	0.541	0.532
		Exponential	0.065	0.498	0.554	0.552
		Log-normal	0.066	0.525	0.521	0.532
	$C_1 \equiv C_2,\ 30\%$	Cox; Partial L.	0.061	0.437	0.436	0.435
		Cox; Full L.	0.054	0.430	0.418	0.423
		Exponential	0.059	0.447	0.441	0.445
		Log-normal	0.058	0.470	0.422	0.425
	$C_1 \equiv C_2,\ 70\%$	Cox; Partial L.	0.073	0.293	0.189	0.243

Continued on next page

Table 3.4: Type I error rate and power for partial correlation. ***(continued)***

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
		Cox; Full L.	0.050	0.264	0.154	0.192
		Exponential	0.072	0.289	0.200	0.235
		Log-normal	0.080	0.294	0.181	0.247
	$C_1 \perp C_2,\ 30\%, 30\%$	Cox; Partial L.	0.061	0.433	0.414	0.414
		Cox; Full L.	0.051	0.414	0.404	0.401
		Exponential	0.053	0.428	0.419	0.427
		Log-normal	0.065	0.456	0.390	0.408
	$C_1 \perp C_2,\ 30\%, 70\%$	Cox; Partial L.	0.063	0.336	0.253	0.251
		Cox; Full L.	0.053	0.314	0.215	0.229
		Exponential	0.061	0.347	0.263	0.262
		Log-normal	0.066	0.344	0.237	0.268
	$C_1 \perp C_2,\ 70\%, 70\%$	Cox; Partial L.	0.084	0.264	0.156	0.219
		Cox; Full L.	0.062	0.243	0.118	0.174
		Exponential	0.088	0.280	0.165	0.212
		Log-normal	0.091	0.273	0.149	0.217
	$C_1 \equiv C_2,\ 30\%,\ \tau = median$	Cox; Partial L.	0.062	0.382	0.286	0.333
		Cox; Full L.	0.060	0.375	0.271	0.313
		Exponential	0.067	0.394	0.287	0.337
		Log-normal	0.066	0.395	0.269	0.324
	$C_1 \equiv C_2,\ 70\%,\ \tau = median$	Cox; Partial L.	0.076	0.309	0.183	0.232
		Cox; Full L.	0.061	0.275	0.155	0.190
		Exponential	0.077	0.299	0.189	0.235
		Log-normal	0.075	0.302	0.183	0.238
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau = median$	Cox; Partial L.	0.070	0.391	0.294	0.318
		Cox; Full L.	0.064	0.377	0.270	0.301
		Exponential	0.061	0.382	0.292	0.317
		Log-normal	0.064	0.389	0.274	0.314
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau = median$	Cox; Partial L.	0.083	0.268	0.144	0.197
		Cox; Full L.	0.066	0.245	0.118	0.166
		Exponential	0.085	0.268	0.158	0.206
		Log-normal	0.089	0.262	0.139	0.222
	$C_1 \equiv C_2,\ \tau = median$	Cox; Partial L.	0.063	0.418	0.344	0.364
		Cox; Full L.	0.057	0.404	0.325	0.344
		Exponential	0.063	0.421	0.336	0.358
		Log-normal	0.065	0.421	0.312	0.355
200	No Censoring	Cox; Partial L.	0.043	0.815	0.828	0.826
		Cox; Full L.	0.043	0.815	0.825	0.825
		Exponential	0.045	0.821	0.837	0.839
		Log-normal	0.046	0.859	0.818	0.818
	$C_1 \equiv C_2,\ 30\%$	Cox; Partial L.	0.058	0.762	0.702	0.711
		Cox; Full L.	0.053	0.765	0.693	0.713
		Exponential	0.051	0.763	0.705	0.732
		Log-normal	0.053	0.796	0.685	0.714
	$C_1 \equiv C_2,\ 70\%$	Cox; Partial L.	0.065	0.533	0.331	0.361
		Cox; Full L.	0.058	0.507	0.308	0.335
		Exponential	0.064	0.523	0.335	0.369
		Log-normal	0.060	0.544	0.326	0.376
	$C_1 \perp C_2,\ 30\%, 30\%$	Cox; Partial L.	0.043	0.746	0.684	0.700
		Cox; Full L.	0.040	0.743	0.678	0.684
		Exponential	0.045	0.746	0.694	0.706
		Log-normal	0.048	0.777	0.671	0.685
	$C_1 \perp C_2,\ 30\%, 70\%$	Cox; Partial L.	0.072	0.617	0.463	0.438
		Cox; Full L.	0.063	0.604	0.432	0.426
		Exponential	0.073	0.622	0.467	0.444
		Log-normal	0.073	0.634	0.443	0.432
	$C_1 \perp C_2,\ 70\%, 70\%$	Cox; Partial L.	0.062	0.499	0.280	0.298
		Cox; Full L.	0.053	0.479	0.252	0.280
		Exponential	0.064	0.502	0.284	0.302
		Log-normal	0.065	0.510	0.260	0.313
	$C_1 \equiv C_2,\ 30\%,\ \tau = median$	Cox; Partial L.	0.060	0.701	0.504	0.507
		Cox; Full L.	0.057	0.702	0.496	0.502
		Exponential	0.053	0.698	0.501	0.511
		Log-normal	0.050	0.716	0.481	0.522
	$C_1 \equiv C_2,\ 70\%,\ \tau = median$	Cox; Partial L.	0.051	0.527	0.307	0.335
		Cox; Full L.	0.041	0.516	0.284	0.312
		Exponential	0.053	0.532	0.317	0.338
		Log-normal	0.054	0.537	0.305	0.336
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau = median$	Cox; Partial L.	0.057	0.677	0.495	0.500
		Cox; Full L.	0.057	0.673	0.483	0.488
		Exponential	0.056	0.669	0.491	0.504

Continued on next page

Table 3.4: Type I error rate and power for partial correlation. ***(continued)***

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
		Log-normal	0.053	0.690	0.474	0.501
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau = median$	Cox; Partial L.	0.059	0.488	0.244	0.289
		Cox; Full L.	0.054	0.464	0.220	0.272
		Exponential	0.061	0.487	0.242	0.300
		Log-normal	0.061	0.494	0.225	0.304
	$C_1 \equiv C_2,\ \tau = median$	Cox; Partial L.	0.057	0.719	0.558	0.572
		Cox; Full L.	0.053	0.713	0.546	0.572
		Exponential	0.057	0.719	0.563	0.578
		Log-normal	0.048	0.741	0.542	0.580

Table 3.5: Bias and RMSE for partial correlation.

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
100	No Censoring	Exponential	-0.005 [0.102]	-0.007 [0.101]	0.001 [0.099]	0.000 [0.098]
		Log-normal	-0.005 [0.103]	0.012 [0.105]	-0.002 [0.099]	0.003 [0.098]
		Cox; Partial L.	-0.006 [0.102]	-0.010 [0.102]	-0.001 [0.098]	0.002 [0.098]
		Cox; Full L.	-0.006 [0.102]	-0.010 [0.103]	-0.001 [0.098]	0.002 [0.098]
	$C_1 \equiv C_2,\ 30\%$	Exponential	-0.007 [0.102]	-0.015 [0.104]	-0.023 [0.101]	0.023 [0.101]
		Log-normal	-0.007 [0.102]	0.000 [0.106]	-0.026 [0.102]	0.027 [0.101]
		Cox; Partial L.	-0.007 [0.101]	-0.019 [0.104]	-0.026 [0.101]	0.026 [0.102]
		Cox; Full L.	-0.008 [0.101]	-0.018 [0.105]	-0.025 [0.102]	0.026 [0.102]
	$C_1 \equiv C_2,\ 70\%$	Exponential	-0.008 [0.103]	-0.045 [0.120]	-0.087 [0.140]	0.091 [0.134]
		Log-normal	-0.008 [0.102]	-0.037 [0.119]	-0.089 [0.142]	0.093 [0.134]
		Cox; Partial L.	-0.009 [0.102]	-0.047 [0.121]	-0.089 [0.139]	0.093 [0.134]
		Cox; Full L.	-0.010 [0.103]	-0.045 [0.121]	-0.088 [0.140]	0.093 [0.134]
	$C_1 \perp C_2,\ 30\%, 30\%$	Exponential	-0.006 [0.101]	-0.020 [0.101]	-0.031 [0.103]	0.032 [0.101]
		Log-normal	-0.006 [0.102]	-0.005 [0.103]	-0.033 [0.104]	0.035 [0.102]
		Cox; Partial L.	-0.007 [0.101]	-0.023 [0.103]	-0.034 [0.104]	0.034 [0.102]
		Cox; Full L.	-0.007 [0.101]	-0.023 [0.103]	-0.034 [0.104]	0.034 [0.103]
	$C_1 \perp C_2,\ 30\%, 70\%$	Exponential	-0.007 [0.098]	-0.047 [0.113]	-0.078 [0.126]	0.076 [0.120]
		Log-normal	-0.006 [0.099]	-0.035 [0.113]	-0.079 [0.128]	0.077 [0.122]
		Cox; Partial L.	-0.007 [0.098]	-0.049 [0.114]	-0.080 [0.127]	0.078 [0.122]
		Cox; Full L.	-0.007 [0.098]	-0.048 [0.114]	-0.080 [0.127]	0.079 [0.122]
	$C_1 \perp C_2,\ 70\%, 70\%$	Exponential	-0.003 [0.096]	-0.062 [0.122]	-0.109 [0.148]	0.113 [0.144]
		Log-normal	-0.004 [0.097]	-0.053 [0.122]	-0.109 [0.149]	0.112 [0.144]
		Cox; Partial L.	-0.004 [0.095]	-0.064 [0.124]	-0.111 [0.148]	0.114 [0.145]
		Cox; Full L.	-0.004 [0.096]	-0.063 [0.124]	-0.111 [0.149]	0.115 [0.146]
	$C_1 \equiv C_2,\ 30\%,\ \tau,\ type\ I$	Exponential	-0.007 [0.099]	-0.031 [0.108]	-0.067 [0.121]	0.066 [0.115]
		Log-normal	-0.007 [0.100]	-0.020 [0.109]	-0.067 [0.123]	0.067 [0.115]
		Cox; Partial L.	-0.007 [0.098]	-0.033 [0.109]	-0.067 [0.121]	0.067 [0.115]
		Cox; Full L.	-0.007 [0.099]	-0.032 [0.109]	-0.067 [0.121]	0.067 [0.115]
	$C_1 \equiv C_2,\ 70\%,\ \tau,\ type\ I$	Exponential	-0.008 [0.102]	-0.046 [0.121]	-0.095 [0.145]	0.099 [0.138]
		Log-normal	-0.008 [0.102]	-0.039 [0.121]	-0.096 [0.146]	0.099 [0.137]
		Cox; Partial L.	-0.009 [0.102]	-0.049 [0.122]	-0.096 [0.145]	0.099 [0.138]
		Cox; Full L.	-0.008 [0.102]	-0.047 [0.122]	-0.095 [0.145]	0.099 [0.139]
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau,\ type\ I$	Exponential	-0.009 [0.099]	-0.034 [0.108]	-0.070 [0.122]	0.070 [0.117]
		Log-normal	-0.008 [0.101]	-0.022 [0.109]	-0.070 [0.124]	0.071 [0.117]
		Cox; Partial L.	-0.009 [0.099]	-0.036 [0.110]	-0.071 [0.123]	0.071 [0.118]
		Cox; Full L.	-0.009 [0.099]	-0.034 [0.110]	-0.071 [0.123]	0.071 [0.118]
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau,\ type\ I$	Exponential	-0.003 [0.095]	-0.064 [0.125]	-0.115 [0.153]	0.120 [0.150]
		Log-normal	-0.003 [0.097]	-0.056 [0.125]	-0.115 [0.154]	0.119 [0.149]
		Cox; Partial L.	-0.004 [0.095]	-0.066 [0.126]	-0.116 [0.153]	0.120 [0.150]
		Cox; Full L.	-0.004 [0.095]	-0.064 [0.125]	-0.116 [0.153]	0.120 [0.150]
200	No Censoring	Exponential	0.000 [0.068]	-0.001 [0.067]	0.000 [0.066]	0.000 [0.066]
		Log-normal	-0.001 [0.069]	0.018 [0.071]	-0.003 [0.067]	0.003 [0.067]
		Cox; Partial L.	0.000 [0.069]	-0.002 [0.068]	-0.002 [0.066]	0.002 [0.067]
		Cox; Full L.	0.000 [0.068]	-0.002 [0.068]	-0.002 [0.066]	0.002 [0.067]
	$C_1 \equiv C_2,\ 30\%$	Exponential	-0.002 [0.070]	-0.008 [0.070]	-0.023 [0.073]	0.024 [0.072]
		Log-normal	-0.002 [0.070]	0.007 [0.072]	-0.026 [0.075]	0.027 [0.073]
		Cox; Partial L.	-0.002 [0.070]	-0.009 [0.071]	-0.025 [0.074]	0.026 [0.072]
		Cox; Full L.	-0.002 [0.070]	-0.009 [0.071]	-0.025 [0.074]	0.026 [0.072]
	$C_1 \equiv C_2,\ 70\%$	Exponential	-0.005 [0.074]	-0.039 [0.089]	-0.087 [0.117]	0.091 [0.116]
		Log-normal	-0.005 [0.074]	-0.030 [0.087]	-0.088 [0.117]	0.094 [0.117]

Continued on next page

Table 3.5: Bias and RMSE for partial correlation. ***(continued)***

N	**Censoring**	**Method**		**Clayton**	**Frank**	**Frank**
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.2$	$\rho = -0.2$
		Cox; Partial L.	-0.005 [0.074]	-0.039 [0.090]	-0.089 [0.118]	0.093 [0.118]
		Cox; Full L.	-0.005 [0.075]	-0.039 [0.089]	-0.089 [0.118]	0.093 [0.118]
	$C_1 \perp C_2,\ 30\%, 30\%$	Exponential	-0.002 [0.069]	-0.015 [0.071]	-0.033 [0.075]	0.032 [0.074]
		Log-normal	-0.002 [0.069]	0.000 [0.071]	-0.034 [0.076]	0.034 [0.075]
		Cox; Partial L.	-0.002 [0.068]	-0.016 [0.071]	-0.034 [0.076]	0.034 [0.075]
		Cox; Full L.	-0.002 [0.068]	-0.015 [0.071]	-0.034 [0.076]	0.034 [0.075]
	$C_1 \perp C_2,\ 30\%, 70\%$	Exponential	-0.002 [0.070]	-0.038 [0.080]	-0.075 [0.103]	0.077 [0.102]
		Log-normal	-0.002 [0.071]	-0.026 [0.078]	-0.075 [0.103]	0.079 [0.103]
		Cox; Partial L.	-0.002 [0.070]	-0.039 [0.082]	-0.076 [0.104]	0.079 [0.103]
		Cox; Full L.	-0.002 [0.070]	-0.038 [0.081]	-0.076 [0.104]	0.079 [0.103]
	$C_1 \perp C_2,\ 70\%, 70\%$	Exponential	-0.004 [0.068]	-0.058 [0.093]	-0.108 [0.128]	0.112 [0.129]
		Log-normal	-0.004 [0.069]	-0.048 [0.091]	-0.107 [0.128]	0.112 [0.129]
		Cox; Partial L.	-0.004 [0.068]	-0.058 [0.094]	-0.109 [0.129]	0.113 [0.130]
		Cox; Full L.	-0.004 [0.068]	-0.057 [0.093]	-0.109 [0.129]	0.114 [0.130]
	$C_1 \equiv C_2,\ 30\%,\ \tau,\ type\ I$	Exponential	-0.001 [0.070]	-0.022 [0.076]	-0.066 [0.096]	0.071 [0.097]
		Log-normal	-0.002 [0.070]	-0.011 [0.076]	-0.065 [0.096]	0.071 [0.097]
		Cox; Partial L.	-0.001 [0.070]	-0.023 [0.076]	-0.066 [0.096]	0.071 [0.097]
		Cox; Full L.	-0.001 [0.070]	-0.022 [0.076]	-0.066 [0.096]	0.071 [0.097]
	$C_1 \equiv C_2,\ 70\%,\ \tau,\ type\ I$	Exponential	-0.005 [0.072]	-0.042 [0.090]	-0.094 [0.122]	0.100 [0.122]
		Log-normal	-0.005 [0.072]	-0.034 [0.088]	-0.095 [0.123]	0.100 [0.122]
		Cox; Partial L.	-0.004 [0.072]	-0.043 [0.091]	-0.095 [0.122]	0.100 [0.123]
		Cox; Full L.	-0.004 [0.072]	-0.042 [0.090]	-0.095 [0.122]	0.100 [0.123]
	$C_1 \perp C_2,\ 30\%, 30\%,\ \tau,\ type\ I$	Exponential	-0.003 [0.068]	-0.027 [0.076]	-0.068 [0.096]	0.071 [0.097]
		Log-normal	-0.003 [0.068]	-0.016 [0.075]	-0.067 [0.097]	0.071 [0.097]
		Cox; Partial L.	-0.003 [0.068]	-0.028 [0.077]	-0.068 [0.097]	0.072 [0.098]
		Cox; Full L.	-0.003 [0.068]	-0.028 [0.076]	-0.068 [0.097]	0.072 [0.098]
	$C_1 \perp C_2,\ 70\%, 70\%,\ \tau,\ type\ I$	Exponential	-0.004 [0.066]	-0.059 [0.094]	-0.115 [0.134]	0.119 [0.134]
		Log-normal	-0.004 [0.067]	-0.051 [0.091]	-0.113 [0.134]	0.118 [0.134]
		Cox; Partial L.	-0.004 [0.066]	-0.060 [0.094]	-0.115 [0.135]	0.119 [0.135]
		Cox; Full L.	-0.004 [0.066]	-0.059 [0.094]	-0.115 [0.135]	0.119 [0.135]

3.9 Appendix 3.B

3.9.1 Notations

Let $U = F_X(T_X)$, $V = F_Y(T_Y)$, and $H(du, dv)$ be the joint probability mass function of U and V. Let $U_0 = F_X(C_X)$, $V_0 = F_Y(C_Y)$ and u_0 and v_0 be realizations of U_0 and V_0, respectively. We use U, V, U_0, and V_0 instead of T_X, T_Y, C_X, and C_Y to simplify the proofs.

3.9.2 Proofs

3.9.2.1 Variance and covariance of PSRs for continuous T_X and T_Y conditionally on censoring values

For continuous T_X, the variance of PSRs conditionally on $U_0 = u_0$ is

$$\begin{aligned}\operatorname{Var}\left\{r(X, F_X, \Delta_X)|u_0\right\} &= \int_0^{u_0} (2U-1)^2 dU + \int_{u_0}^1 u_0^2 dU = \frac{1}{2}\frac{(2U-1)^3}{3}\bigg|_0^{u_0} + u_0^2(1-u_0) \\ &= \frac{1}{3}u_0^3 - u_0^2 + u_0. \end{aligned} \quad (3.13)$$

For continuous T_X and T_Y, the covariance of PSRs conditionally on $U_0 = u_0$ and $V_0 = v_0$ is

$$\text{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y) | u_0, v_0\right\} = A + B + C + D, \tag{3.14}$$

$$\begin{aligned} &\text{where} \\ A &= \int_0^{v_0} \int_0^{u_0} (2u-1)(2v-1) H(du, dv), \\ B &= \int_{v_0}^{1} \int_0^{u_0} v_0 (2u-1) H(du, dv), \\ C &= \int_0^{v_0} \int_{u_0}^{1} u_0 (2v-1) H(du, dv), \\ D &= \int_{v_0}^{1} \int_{u_0}^{1} u_0 v_0 H(du, dv). \end{aligned}$$

3.9.2.2 ρ_{PSR} under general censoring scenario

The population parameter of ρ_{PSR} can be obtained from equations (3.13) and (3.14), variance and covariance of ρ_{PSR} conditionally on censoring variables $U_0 = F_X(C_X)$, $V_0 = F_Y(C_Y)$,

$$\rho_{PSR} = \frac{\text{E}_{U_0,V_0}\text{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y) | U_0, V_0\right\}}{\sqrt{\left\{\text{E}_{U_0}\text{Var}\left\{r(X, F_X, \Delta_X) | U_0\right\}\right\}\left\{\text{E}_{V_0}\text{Var}\left\{r(Y, F_Y, \Delta_Y) | V_0\right\}\right\}}}, \tag{3.15}$$

From (3.15), it is clear that ρ_{PSR} depends on the censoring distribution. To obtain the expression for the bias of ρ_{PSR} for ρ_S we would have had to factor out the term representing ρ_S. Although the numerator of the above expression can be presented as a sum of terms similar to A, B, C, and D in (3.14), because of the denominator, such representation does not help to factor out ρ_S.

Note that (3.15) can be considered as a general expression for continuous and discrete T_X and T_Y. Either way, ρ_{PSR} depends on the censoring distribution, and so is the bias of ρ_{PSR} for ρ_S.

3.9.2.3 ρ_{PSR} for $T_X \perp T_Y$

For continuous and independent T_X and T_Y, $H(dx, dy) = dxdy$, and the conditional covariance (3.14) is

$$\begin{aligned} \text{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y) | u_0, v_0\right\} &= (u_0^2 - u_0)(v_0^2 - v_0) + v_0(1 - v_0)(u_0^2 - u_0) \\ &+ u_0(1 - u_0)(v_0^2 - v_0) + v_0(1 - v_0)u_0(1 - u_0) \\ &= 0. \end{aligned}$$

Because the above is true for any u_0 and v_0, we have $\mathrm{E}_{U_0,V_0}\mathrm{Cov}\left\{r(X,F_X,\Delta_X),\ r(Y,F_Y,\Delta_Y)|U_0,V_0\right\} = 0$, and it follows from (3.15) that $\rho_{PSR} = 0$ regardless of the censoring distribution.

For discrete and independent T_X and T_Y, we use our usual notations T_X, T_Y, C_X, and C_Y and let the support of T_X be t_i, where $i = 0, ..., \infty$ and of T_Y be s_j, where $j = 0, ..., \infty$. Let c_X and c_Y be realizations of C_X and C_Y respectively. Let $P_{X,i} = \Pr(T_X = t_i) = F_X(t_i) - F_X(t_{i-1})$ and $P_{Y,j} = \Pr(T_Y = s_j) = F_Y(s_j) - F_Y(s_{j-1})$. For discrete T_X and T_Y, expression (3.14) can be rewritten as

$$
\begin{aligned}
&\mathrm{Cov}\left\{r(X,F_X,\Delta_X),\ r(Y,F_Y,\Delta_Y)|c_X,c_Y\right\}\\
&\quad= \sum_{i:t_i \le c_X}\sum_{j:s_j \le c_Y}\left\{F_X(t_i)+F_X(t_{i-1})-1\right\}\left\{(F_Y(s_j)+F_Y(s_{j-1}-1\right\}P_{X,i}P_{Y,j}\\
&\qquad+\sum_{i:t_i \le c_X}\sum_{j:s_j > c_Y}F_Y(c_Y)\left\{F_X(t_i)+F_X(t_{i-1})-1\right\}P_{X,i}P_{Y,j}+\\
&\qquad+\sum_{i:t_i > c_X}\sum_{j:s_j \le c_Y}F_X(c_X)\left\{F_Y(s_j)+F_Y(s_{j-1}-1\right\}P_{X,i}P_{Y,j}\\
&\qquad+\sum_{i:t_i > c_X}\sum_{j:s_j > c_Y}F_X(c_X)F_Y(c_Y)P_{X,i}P_{Y,j}\\
&\quad=\left[\sum_{i:t_i \le c_X}\left\{F_X(t_i)+F_X(t_{i-1})-1\right\}P_{X,i}+F_X(c_X)\sum_{i:t_i > c_X}P_{X,i}\right]\\
&\qquad\times\left[\sum_{j:s_j \le c_Y}\left\{(F_Y(s_j)+F_Y(s_{j-1}-1\right\}P_{Y,j}+F_Y(c_Y)\sum_{j:s_j > c_Y}P_{Y,j}\right].
\end{aligned}
$$

Because $F_X(t_i) = \sum_{i:t_k \le t_i} P_{X,k}$, it is straight forward to show that $\sum_{i:t_i \le c_X} F_X(t_i)P_{X,i}$ $= \frac{1}{2}\left\{\left(\sum_{i:t_i \le c_X} P_{X,i}\right)^2 + \sum_{i:t_i \le c_X} P_{X,i}^2\right\}$ and $\sum_{i:t_i \le c_X} F_X(t_{i-1})P_{X,i} = \frac{1}{2}\left\{\left(\sum_{i:t_i \le c_X} P_{X,i}\right)^2 - \sum_{i:t_i \le c_X} P_{X,i}^2\right\}$.
Therefore, term $\sum_{i:t_i \le c_X}\left\{F_X(t_i)+F_X(t_i)-1\right\}P_{X,i}+F_X(c_X)\sum_{i:t_i > c_X}P_{X,i}$ can be rewritten as

$$
\begin{aligned}
&\frac{1}{2}\left\{\left(\sum_{i:t_i \le c_X}P_{X,i}\right)^2+\sum_{i:t_i \le c_X}P_{X,i}^2\right\}+\frac{1}{2}\left\{\left(\sum_{i:t_i \le c_X}P_{X,i}\right)^2-\sum_{i:t_i \le c_X}P_{X,i}^2\right\}-\sum_{i:t_i \le c_X}P_{X,i}+F_X(c_X)\left\{1-F_X(c_X)\right\}\\
&=F_X(c_X)^2-F_X(c_X)+F_X(c_X)\left\{1-F_X(c_X)\right\}\\
&=0.
\end{aligned}
$$

Similarly, it can be shown that $\left[\sum_{j:s_j \le c_Y}\left\{(F_Y(s_j)+F_Y(s_{j-1}-1\right\}P_{Y,j}+F_Y(c_Y)\sum_{j:s_j > c_Y}P_{Y,j}\right] = 0$. Because the above is true for any c_X and c_Y, we have $\mathrm{E}_{C_X,C_Y}\mathrm{Cov}\left\{r(X,F_X,\Delta_X),\ r(Y,F_Y,\Delta_Y)|C_X,C_Y\right\} =$

0, and it follows from (3.15) that $\rho_{PSR} = 0$ regardless of the censoring distribution.

3.9.3 ρ_{PSR} for continuous T_X and T_Y under strict type I censoring

3.9.3.1 ρ_{PSR} for perfectly positively correlated T_X and T_Y

Because $T_X \equiv T_Y$, we have: $U = F_X(T_X) = F_Y(T_Y) = V$. The marginal distributions of U and V are uniform on $[0,1]$ and their joint cumulative distribution function (CDF) is $H(u,v) = \min(u,v)$ (see Nelsen (2007)). Because the entire probability mass is concentrated on the diagonal $u = v$ of length $\sqrt{2}$, the probability density function (PDF) is $h(u,v) = \mathbb{1}(u = v)(1/\sqrt{2})$ and the integration is done over the curve $u(t) = t$ and $v(t) = t$. The integral includes term $\sqrt{\left(\frac{dU(t)}{dt}\right)^2 + \left(\frac{dV(t)}{dt}\right)^2} = \sqrt{(t')^2 + (t')^2} = \sqrt{1^2 + 1^2} = \sqrt{2}$, so $\sqrt{2}$ and $1/\sqrt{2}$ cancel out in the following derivations. Using equation (3.14) we first compute the case when $u_0 < v_0$. The terms A, B, C, and D from (3.14) are computed in the following way:

$$\begin{aligned}
A &= \int_0^{u_0} (2t-1)^2 dt = \frac{1}{6}\left\{(2u_0 - 1)^3 - (0-1)^3\right\} = \frac{1}{6}\left\{(2u_0-1)^3 + 1\right\},\\
C &= \int_{u_0}^{v_0} (2t-1)u_0 dt = \frac{u_0}{2}\int_{u_0}^{v_0} (2t-1)d(2t-1) = \frac{u_0}{4}\left\{(2v_0-1)^2 - (2u_0-1)^2\right\},\\
B &= \frac{v_0}{4}\left\{(2u_0-1)^2 - (2v_0-1)^2\right\},\\
D &= \int_{v_0}^{1} v_0 u_0 dt = u_0 v_0 (1 - v_0).
\end{aligned}$$

When $u_0 \geq v_0$, the expressions for A, B, C, and D are computed similarly. Summing up A, B, C, and D, and denoting $\mathrm{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)|u_0, v_0\right\}$ and $\mathrm{Cor}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)|u_0, v_0\right\}$ as $\mathrm{Cov}^+_{PSR}(u_0, v_0)$ as $\rho^+_{PSR}(u_0, v_0)$, respectively, for perfectly positively correlated variables we have

$$\begin{aligned}
\mathrm{Cov}^+_{PSR}(u_0,v_0) &= \frac{1}{6}\left\{(2u_0-1)^3+1\right\} + \frac{u_0}{4}\left\{(2v_0-1)^2 - (2u_0-1)^2\right\} + u_0v_0(1-v_0), \quad \textit{when } u_0 < v_0,\\
&= \frac{1}{6}\left\{(2v_0-1)^3+1\right\} + \frac{v_0}{4}\left\{(2u_0-1)^2 - (2v_0-1)^2\right\} + v_0u_0(1-u_0), \quad \textit{when } u_0 \geq v_0.
\end{aligned}$$

Keeping in mind (3.13) for conditional variance, we have

$$\rho^+_{PSR}(u_0, v_0) = \frac{\mathrm{Cov}^+_{PSR}(u_0, v_0)}{\sqrt{\left(\frac{1}{3}u_0^3 - u_0^2 + u_0\right)\left(\frac{1}{3}v_0^3 - v_0^2 + v_0\right)}}.$$

To obtain $\rho^+_{PSR}(u_0, v_0)$ in terms of censoring proportions γ_X and γ_Y, we use the following variable transformations: $u_0 = 1 - \gamma_X$ and $v_0 = 1 - \gamma_Y$.

3.9.3.2 ρ_{PSR} for perfectly negatively correlated T_X and T_Y

When T_X and T_Y are perfectly negatively correlated, the probability mass is concentrated on the curve $u = 1 - v$ or $u(t) = t$ and $v(t) = 1 - t$, and similar to Section 3.9.3.1, the integration is done over this curve. We first integrate the area $u_0 + v_0 < 1$, where integral A (see Section 3.9.3.1) is 0, and

$$\begin{aligned}
B &= \int_{1-v_0}^{1} (-2t+1)u_0 dt = -\frac{u_0}{2}\int_{1-v_0}^{1} (2t-1)d(2t-1) = -\frac{u_0}{4}\left\{(2v_0-1)^2 - 1\right\} = u_0v_0(u_0-1),\\
C &= \int_0^{u_0} (2t-1)u_0 dt = u_0v_0(v_0-1),\\
D &= \int_{u_0}^{1-v_0} u_0v_0 dt = u_0v_0(1-u_0-v_0).
\end{aligned}$$

Integration over the area $u_0 + v_0 > 1$, gives $D = 0$ and

$$\begin{aligned}
A &= \int_{1-v_0}^{u_0} -(2t-1)^2 dt = -\frac{1}{2}\int_{1-v_0}^{u_0} (2t-1)^2 d(2t-1) = -\frac{1}{6}\left\{(2u_0-1)^3 + (2v_0-1)^3\right\},\\
B &= \int_{u_0}^{1} (-2t+1)u_0 dt = u_0^2(u_0-1),\\
C &= \int_0^{1-v_0} (2t-1)u_0 dt = v_0^2(v_0-1).
\end{aligned}$$

When $u_0 + v_0 = 1$, the derivations are similar except that $A = D = 0$. Summing up all four integrals, and denoting $\text{Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)|u_0, v_0\right\}$ and $\text{Cor}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)|u_0, v_0\right\}$ as $\text{Cov}^-_{PSR}(u_0, v_0)$ as $\rho^-_{PSR}(u_0, v_0)$, respectively, for perfectly negatively correlated variables, we have we have the following expressions:

$$\begin{aligned}
\text{Cov}^-_{PSR}(u_0, v_0) &= -u_0v_0, && \textit{when } u_0 + v_0 < 1,\\
&= -\frac{1}{6}\left\{(2u_0-1)^3 + (2v_0-1)^3\right\} + u_0^2(u_0-1) + v_0^2(v_0-1), && \textit{when } u_0 + v_0 \geq 1,\\
\rho^-_{PSR}(u_0, v_0) &= \frac{\text{Cov}^-_{PSR}(u_0, v_0)}{\sqrt{\left(\frac{1}{3}u_0^3 - u_0^2 + u_0\right)\left(\frac{1}{3}v_0^3 - v_0^2 + v_0\right)}}.
\end{aligned}$$

To obtain $\rho^-_{PSR}(u_0, v_0)$ in terms of censoring proportions γ_X and γ_Y, we use the following variable transformations: $u_0 = 1 - \gamma_X$ and $v_0 = 1 - \gamma_Y$.

3.9.4 ρ_{PSR} for continuous T_X and T_Y under unbounded censoring

We design the unbounded censoring mechanism in such a way that first, we choose proportions γ_X^* and γ_Y^* of observations U and V, respectively, for potential censoring. Second, we sample the censoring time for the selected observations using the same marginal distributions as their corresponding time to event. As a result, the probabilities of being actually censored for U and V are $\gamma_X = \gamma_X^*/2$ and $\gamma_Y = \gamma_Y^*/2$, respectively. For simplicity of derivations, we set $\gamma_X^* = \gamma_Y^* = \gamma^*$. Formally, this censoring design implies that U_0 and V_0 have the following marginal distribution:

$$\begin{aligned} &= 1, && \text{with probability } 1-\gamma^*; \\ &\sim Unif(0,1), && \text{with probability } \gamma^*. \end{aligned}$$

Because U_0 and V_0 are independent, their joint PDF is

$$\begin{aligned} h(u_0,v_0) &= (1,1), \text{ with probability } (1-\gamma^*)^2, \\ &= (1,v_0), \text{ where } v_0 \sim \gamma^* Unif(0,1), \\ &= (u_0,1), \text{ where } u_0 \sim \gamma^* Unif(0,1), \\ &= (u_0,v_0), \text{ where } u_0 \sim (\gamma^*)^2 Unif(0,1)\cdot Unif(0,1). \end{aligned} \tag{3.16}$$

3.9.4.1 ρ_{PSR} for perfectly positively correlated T_X and T_Y

Section 3.9.3.1 provides the expression for conditional covariance of PSRs, $\mathrm{Cov}^+_{PSR}(U_0,V_0)$. Taking its expectation with respect to the joint distribution (3.16) of U_0 and V_0, we get

$$\begin{aligned} \mathrm{Cov}^+_{PSR}(\gamma^*) = &\int_0^1\int_0^1 \mathrm{Cov}^+_{PSR}(u_0,v_0)h(u_0,v_0)du_0dv_0 = \\ &= (1-\gamma^*)^2\cdot\frac{1}{3} + (1-\gamma^*)\gamma^*\int_0^1\left[\frac{1}{6}\left\{(2v_0-1)^3+1\right\}+\frac{v_0}{4}\left\{1-(2v_0-1)^2\right\}\right]dv_0 + \\ &+ (1-\gamma^*)\gamma^*\int_0^1\left[\frac{1}{6}\left\{(2u_0-1)^3+1\right\}+\frac{u_0}{4}\left\{1-(2u_0-1)^2\right\}\right]du_0 + \\ &+ (\gamma^*)^2\int_0^1\int_0^1 \mathrm{Cov}^+_{PSR}(u_0,v_0)du_0dv_0. \end{aligned}$$

The second and third terms in the above sum are equal to $\frac{1}{4}$. The last term is

$$\begin{aligned}
&\int_0^1 \int_0^1 \text{Cov}^+_{PSR}(u_0, v_0) du_0 dv_0 \\
&\quad = 2\int_0^1 \int_0^{v_0} \left[\frac{1}{6}\left\{(2u_0-1)^3+1\right\} + \frac{u_0}{4}\left\{(2v_0-1)^2-(2u_0-1)^2\right\} + u_0 v_0(1-v_0)\right] du_0 dv_0 \\
&\quad = 2\int_0^1 \left(\frac{1}{6}\left[\frac{1}{8}\left\{(2v_0-1)^4-1\right\}+v_0\right] + \frac{1}{4}\left\{(4v_0^2-4v_0+1)\frac{v_0^2}{2} - \left(v_0^4 - \frac{4}{3}v_0^3 + \frac{v_0^2}{2}\right)\right\} + v_0(1-v_0)\frac{v_0^2}{2}\right) dv_0 \\
&\quad = \frac{2}{15} + \frac{2}{120} + \frac{2}{40} = \frac{1}{5}.
\end{aligned}$$

The resulting covariance is

$$\text{Cov}^+_{PSR}(\gamma^*) = (1-\gamma^*)^2 \cdot \frac{1}{3} + 2(1-\gamma^*)\gamma^*\frac{1}{4} + (\gamma^*)^2 \cdot \frac{1}{5}.$$

According to Shepherd et al. (2016), under unbounded censoring, the mean of PSRs is 0, and the variance is $\frac{1}{3} - E_{C.}\left[(1-F(C.))^3\right]$. If the distribution of the censored data is the same as the distribution of the time to event, then $E_{C.}\left[(1-F(C.))^3\right] = \frac{1}{12}$, which is the case here except only $\gamma^{*\ th}$ proportion of all observations is subject to censoring, so the resulting variance is $\frac{1}{3} - \gamma^*\frac{1}{12}$, and the correlation of PSRs is

$$\rho^+_{PSR}(\gamma^*) = \frac{(1-\gamma^*)^2 \cdot \frac{1}{3} + 2(1-\gamma^*)\gamma^*\frac{1}{4} + (\gamma^*)^2 \cdot \frac{1}{5}}{\frac{1}{3} - \gamma^*\frac{1}{12}}.$$

Similarly, it is straightforward to show that for $\gamma^*_X \neq \gamma^*_Y$

$$\rho^+_{PSR}(\gamma^*_X, \gamma^*_Y) = \frac{(1-\gamma^*_X)(1-\gamma^*_Y) \cdot \frac{1}{3} + (1-\gamma^*_X)\gamma^*_Y \cdot \frac{1}{4} + (1-\gamma^*_Y)\gamma^*_X \cdot \frac{1}{4} + \gamma^*_X\gamma^*_Y \cdot \frac{1}{5}}{\sqrt{\left(\frac{1}{3} - \gamma^*_X\frac{1}{12}\right)\left(\frac{1}{3} - \gamma^*_Y\frac{1}{12}\right)}}.$$

To obtain $\rho^+_{PSR}(\gamma^*_X, \gamma^*_Y)$ in terms of actual censoring proportions γ_X and γ_Y, we use the following variable transformations: $\gamma^*_X = 2\gamma_X$ and $\gamma^*_Y = 2\gamma_Y$.

3.9.4.2 ρ_{PSR} *for perfectly negatively correlated* T_X *and* T_Y

Using the results of Section 3.9.3.2, we have the following:

$$\begin{aligned}\mathrm{Cov}^-_{PSR}(\gamma^*) &= \int_0^1\int_0^1 \mathrm{Cov}^-(u_0,v_0)h(u_0,v_0)du_0dv_0 = \\ &= -(1-\gamma^*)^2\cdot\frac{1}{3} + (1-\gamma^*)\gamma^*\int_0^1\left[-\frac{1}{6}\left\{1+(2v_0-1)^3\right\}+v_0^2(v_0-1)\right]dv_0 - \\ &+ (1-\gamma^*)\gamma^*\int_0^1\left[-\frac{1}{6}\left\{1+(2u_0-1)^3\right\}+u_0^2(u_0-1)\right]du_0 + \\ &+ (\gamma^*)^2\int_0^1\int_0^1 \mathrm{Cov}^-_{PSR}(u_0,v_0)du_0dv_0.\end{aligned}$$

After some tedious math, we obtain

$$\rho^-_{PSR}(\gamma^*) = -\frac{(1-\gamma^*)^2\cdot\frac{1}{3}+2(1-\gamma^*)\gamma^*\cdot\frac{1}{4}+(\gamma^*)^2\cdot\frac{7}{40}}{\frac{1}{3}-\gamma^*\frac{1}{12}}.$$

When $\gamma^*_X \neq \gamma^*_Y$, it is straight forward to show that

$$\rho^-(\gamma^*_X,\gamma^*_Y) = -\frac{(1-\gamma^*_X)(1-\gamma^*_Y)\cdot\frac{1}{3}+(1-\gamma^*_X)\gamma^*_Y\cdot\frac{1}{4}+(1-\gamma^*_Y)\gamma^*_X\cdot\frac{1}{4}+\gamma^*_X\gamma^*_Y\cdot\frac{7}{40}}{\sqrt{\left(\frac{1}{3}-\gamma^*_X\frac{1}{12}\right)\left(\frac{1}{3}-\gamma^*_Y\frac{1}{12}\right)}}.$$

To obtain $\rho^-_{PSR}(\gamma^*_X,\gamma^*_Y)$ in terms of actual censoring proportions γ_X and γ_Y, we use the following variable transformations: $\gamma^*_X = 2\gamma_X$ and $\gamma^*_Y = 2\gamma_Y$.

3.10 Appendix 3.C

3.10.1 Notations

For brevity, we denote $r(y_i, F_Y, \delta_{Y,i})$ as $r_{Y,i}$.

3.10.2 Estimating equations and derivatives of PSRs for unadjusted ρ_{PSR}

We assume that the number of unique time points is k. The estimating equations for outcomes T_X and T_Y are denoted as ψ_X and ψ_Y, respectively. Let k be the number of unique time points, so we have k estimating equations. Following the example of

Shepherd et al. (2007), we have:

$$\begin{aligned}
\psi_Y(\widehat{F}_{1,Y}(y_i), y_i) &= V_{1,i,Y} - \left(1 - \widehat{S}_{1,Y}(y_i)\right), \\
\psi_Y(\widehat{F}_{2,Y}(y_i), y_i) &= V_{2,i,Y} - \left(1 - \widehat{S}_{2,Y}(y_i)\right), \\
&\dots \\
\psi_Y(\widehat{F}_{k,Y}(y_i), y_i) &= V_{k,i,Y} - \left(1 - \widehat{S}_{k,Y}(y_i)\right),
\end{aligned}$$

where functions $V_{j,i} = \phi_j(Y_i)\gamma_0(Y_i)\epsilon_i + \gamma_{j1}(Y_i)(1-\epsilon_i) - \gamma_{j2}(Y_i)$ are derived by Stute (1995) and

$$\begin{aligned}
\phi_j(Y_i) &= I_{\{Y_i \le t_j\}}, \\
\gamma_0(Y_i) &= exp\left(\int_{-\infty}^{Y_i^-} \frac{H^0(dw)}{1-H(w)}\right), \\
\gamma_{j,1}(Y_i) &= \frac{1}{1-H(Y_i)} \int I_{\{Y_i < w\}} \phi_j(w)\gamma_0(w) H^1(dw), \\
\gamma_{j,2}(Y_i) &= \int\int \frac{I_{\{v<Y_i,\ v<w\}}\phi_j(w)\gamma_0(w)}{[1-H(v)]^2} H^0(dv)H^1(dw), \\
H^0(y) &= P\{Y \le y,\ \epsilon = 0\} = \frac{\sum_i (1-\epsilon_i) I_{\{Y_i \le y\}}}{N}, \\
H^1(y) &= P\{Y \le y,\ \epsilon = 1\} = \frac{\sum_i \epsilon_i I_{\{Y_i \le y\}}}{N}, \\
H(y) &= P\{Y \le y\} = \frac{\sum_i I_{\{Y_i \le y\}}}{N}, \\
\gamma_{j,2}(Y_i) &= \int\int \frac{I_{\{v<Y_i,\ v<w\}}\phi_j(w)\gamma_0(w)}{[1-H(v)]^2} H^0(dv)H^1(dw).
\end{aligned}$$

To reduce computational time, we save $\gamma_{j,1}(v)$ to compute $\gamma_{j,2}(Y_i)$:

$$\begin{aligned}
\gamma_{j,2}(Y_i) &= \int\int \frac{I_{\{v<Y_i,\ v<w\}}\phi_j(w)\gamma_0(w)}{[1-H(v)]^2} H^0(dv)H^1(dw) \\
&= \int \frac{I_{\{v<Y_i\}}}{1-H(v)} \left[\frac{1}{1-H(v)} \int I_{\{v<w\}}\phi_j(w)\gamma_0(w)H^1(dw)\right] H^0(dv) \\
&= \int \frac{I_{\{v<Y_i\}}\gamma_{j,1}(v)}{1-H(v)} H^0(dv).
\end{aligned}$$

The derivatives of PSRs according to $\widehat{F}_i$ are

$$\frac{\partial r_{Y,i}}{\partial \widehat{F}_{Y,i}} = \frac{\partial \left[\widehat{F}_Y(y_i) + \widehat{F}_Y(y_{i-1})\delta_i - \delta_i\right]}{\partial \widehat{F}_{Y,i}} = 1, \quad \text{if } \widehat{F}_{Y,i} = \widehat{F}_Y(y_i),$$
$$= \delta_i, \quad \text{if } \widehat{F}_{Y,i} = \widehat{F}_Y(y_{i-1})$$
$$= 0, \quad \text{otherwise.}$$

3.10.3 Derivatives of PSRs for parametric survival models

Because we use the estimating equations provided by **R**, it is important to make sure that the derivatives of PSRs are taken according to the correct parameters. In the following sections, we review the parametrization of the parametric survival models in **R** and derive the derivatives of PSRs based on this parametrization. According to Zhang (2020) parametric survival models can be parametrized as following:

$$\begin{aligned} log(T_i) &= \beta_0 + \beta_1 z_{i1} + ... + \beta_p z_{ip} + \sigma\epsilon_i \text{ or } T_i = e^{\beta_0+\beta_1 z_{i1}+...+\beta_p z_{ip}+\sigma\epsilon_i}, \\ S_{T_i}(t) &= P(T_i > t) = P\left(e^{\beta_0+\beta_1 z_{i1}+...+\beta_p z_{ip}+\sigma\epsilon_i} > t\right) = P\left(e^{\sigma\epsilon_i} > e^{-\beta_0-\beta_1 z_{i1}-...-\beta_p z_{ip}}t\right) \\ &= S_{e^{\sigma\epsilon_i}}(e^{-\boldsymbol{z}_i\boldsymbol{\beta}}t), \end{aligned}$$

where $e^{\sigma\epsilon_i}$ can be Weibull, exponential, gamma, log-logistic, or log-normal distribution.

3.10.3.1 Exponential survival model

For exponential survival model, $\sigma = 1$, $e^{\epsilon_i} \sim Exponential(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})$, and $F_{e^{\epsilon_i}}(y_i) = 1 - e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i}$. The PSRs derivatives are

$$\begin{aligned} r_{Y,i} &= 1 - (1+\delta_i)e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i}, \\ \frac{\partial r_{Y,i}}{\partial \beta_0} &= -(1+\delta_i)e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i}\left[-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i\right](-1) = -(1+\delta_i)y_i e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i-\boldsymbol{z}_i\boldsymbol{\beta}}, \\ \frac{\partial r_{Y,i}}{\partial \beta_1} &= -(1+\delta_i)\ y_i\ z_{i1}\ e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i-\boldsymbol{z}_i\boldsymbol{\beta}}, \\ \frac{\partial r_{Y,i}}{\partial \beta_2} &= -(1+\delta_i)\ y_i\ z_{i2}\ e^{-(e^{-\boldsymbol{z}_i\boldsymbol{\beta}})y_i-\boldsymbol{z}_i\boldsymbol{\beta}}. \end{aligned}$$

3.10.3.2 Weibull survival model

In **R**, Weibull distribution (`rweibull()`) is parameterized as $F(y) = 1 - exp\left(-\left(\frac{y}{scale}\right)^{shape}\right)$, but the parameterization for `survreg()` is different:

$$
\begin{aligned}
scale = e^{\boldsymbol{z}_i\boldsymbol{\beta}}, \qquad\qquad shape = \frac{1}{\sigma},\\
F(y_i) = 1 - exp\left(-\left(\frac{y}{e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}\right).
\end{aligned}
$$

The PSRs and their derivatives by β_j are

$$
\begin{aligned}
r_{Y,i} &= (1+\delta_i)\left[1 - exp\left(-\left(\frac{y}{e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}\right)\right] - \delta_i,\\
\frac{\partial r_{Y,i}}{\partial \beta_j} &= -(1+\delta_i)\frac{z_{ji}}{\sigma} y_i^{\frac{1}{\sigma}} e^{\left(-y_i \cdot e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{\frac{1}{\sigma}} - \frac{\boldsymbol{z}_i\boldsymbol{\beta}}{\sigma}}.
\end{aligned}
$$

It appears that **R** provides score equations based on $log(\sigma)$, so we compute the derivatives accordingly:

$$
\begin{aligned}
\gamma = ln(\sigma), \qquad \sigma = e^{\gamma}, \qquad \frac{1}{\sigma} = e^{-\gamma},&\\
F(y_i) &= 1 - exp\left(-\left(\frac{y}{e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\right)^{e^{-\gamma}}\right),\\
r_{Y,i} &= (1+\delta_i)\left[1 - exp\left(-\left(\frac{y_i}{e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\right)^{e^{-\gamma}}\right)\right] - \delta_i,\\
\frac{\partial r_{Y,i}}{\partial \gamma} &= (1+\delta_i)\left[-exp\left(-\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{e^{-\gamma}}\right)\frac{\partial}{\partial\gamma}\left(-\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{e^{-\gamma}}\right)\right]\\
&= (1+\delta_i) exp\left(-\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{e^{-\gamma}}\right) ln\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{e^{-\gamma}}\left(-e^{-\gamma}\right)\\
&= -(1+\delta_i)\left(ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}\right) y_i^{e^{-\gamma}} exp\left\{-\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{e^{-\gamma}} - \boldsymbol{z}_i\boldsymbol{\beta}e^{-\gamma} - \gamma\right\},\\
\left.\frac{\partial r_{Y,i}}{\partial \gamma}\right|_{\gamma = log(\sigma)} &= -(1+\delta_i)\left(ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}\right) y_i^{\frac{1}{\sigma}} exp\left\{-\left(y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{\frac{1}{\sigma}} - \frac{\boldsymbol{z}_i\boldsymbol{\beta}}{\sigma} - ln(\sigma)\right\}.
\end{aligned}
$$

3.10.3.3 Log-logistic survival model

According to DaÌĹtwyler (2011), for log-logistic distribution, **R** uses parametrization of *accelerated failure time* (AFT) model, $F_{Y_i}(y_i) = \frac{1}{1 + y_i^{-\gamma}\left(e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right)^{-\gamma}}$. When computing score equations, **R** uses $\gamma^* = -log(\gamma)$, so the PSRs derivatives are

$$
\begin{aligned}
r_{Y,i} &= 1 - (1+\delta_i)\frac{1}{1+[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}]^{\gamma}}, \\
\frac{\partial r_{Y,i}}{\partial \beta_0} &= -\gamma(1+\delta_i)\frac{\left[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right]^{\gamma}}{\left(1+[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}]^{\gamma}\right)^2}, \\
\frac{\partial r_{Y,i}}{\partial \beta_1} &= -\gamma(1+\delta_i)\frac{\left[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right]^{\gamma}}{\left(1+[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}]^{\gamma}\right)^2} z_{i1}, \\
\left.\frac{\partial r_{Y,i}}{\partial \gamma^*}\right|_{\gamma^*=-log(\gamma)} &= \quad \gamma(1+\delta_i)\frac{\left[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}\right]^{\gamma}}{\left(1+[y_i e^{-\boldsymbol{z}_i\boldsymbol{\beta}}]^{\gamma}\right)^2}(log(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}).
\end{aligned}
$$

3.10.3.4 Log-normal survival model

For log-normal model, we have $f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(log(y)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}}$ and $F(y) = \Phi\left(\frac{ln(y)-\boldsymbol{z}_i\boldsymbol{\beta}}{\sigma}\right)$, where $\Phi\left(\frac{ln(y)-\boldsymbol{z}_i\boldsymbol{\beta}}{\sigma}\right)$ is the CDF of normal distribution. The PSRs and their derivatives according to β_j are

$$
r_{Y,i} = (1+\delta_i)F(y) - \delta_i = (1+\delta_i)\Phi\left(\frac{ln(y_i)-\boldsymbol{z}_i\boldsymbol{\beta}}{\sigma}\right) - \delta_i,
$$

$$
\frac{\partial r_{Y,i}}{\partial \beta_j} = -(1+\delta_i)z_{ji}\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(ln(y_i)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}}.
$$

To compute $\frac{\partial r_{Y,i}}{\partial \sigma}$ we use *Leibniz* rule (see Wikipedia contributors (2020)) keeping in mind that **R** seems to use parameter $\gamma = log(\sigma)$,

$$
\begin{aligned}
\gamma &= ln(\sigma); \qquad \sigma = e^{\gamma} \\
\frac{\partial r_{Y,i}}{\partial \gamma} &= (1+\delta_i)\int_0^{y_i} \frac{\partial}{\partial \gamma}\frac{e^{-\gamma}}{t\sqrt{2\pi}} e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2}e^{-2\gamma}} dt \\
&= (1+\delta_i)\int_0^{y_i} \frac{1}{t\sqrt{2\pi}} e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2}e^{-2\gamma}-\gamma}\left[\cancel{-}\frac{\cancel{1}}{2}\left(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta}\right)^2 e^{-2\gamma}(\cancel{-2}) - 1\right] dt \\
&= (1+\delta_i)\left[\int_0^{y_i} \frac{1}{t\sqrt{2\pi}}\left(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta}\right)^2 e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2}e^{-2\gamma}-3\gamma} dt - \int_0^{y_i} \frac{1}{t\sqrt{2\pi}} e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2}e^{-2\gamma}-\gamma} dt\right] \\
&= (1+\delta_i)\left[A - B\right].
\end{aligned}
$$

It is straight forward to show (using $\gamma = ln(\sigma)$) that $B = F_Y(y_i)$ (the CDF of the corresponding log-normal distribution). To compute A, we replace γ with $ln(\sigma)$ and keep in mind that $\frac{dt}{t} = d\left(ln(t)\right)$:

$$A = \int_0^{y_i} \frac{1}{t\sqrt{2\pi}\sigma^3}\left(ln(t) - \boldsymbol{z}_i\boldsymbol{\beta}\right)^2 e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}} dt = \int_0^{y_i} \frac{1}{\sqrt{2\pi}\sigma^3}\left(ln(t) - \boldsymbol{z}_i\boldsymbol{\beta}\right)^2 e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}} d(ln(t) - \boldsymbol{z}_i\boldsymbol{\beta}).$$

Using the following variable substitution:

$$x = \frac{ln(t) - \boldsymbol{z}_i\boldsymbol{\beta}}{\sigma}, \quad e^{-\frac{(ln(t)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}} = e^{-\frac{x^2}{2}}, \quad (ln(t) - \boldsymbol{z}_i\boldsymbol{\beta})^2 = x^2\sigma^2, \quad d(ln(t) - \boldsymbol{z}_i\boldsymbol{\beta}) = \sigma dx,$$

if $t = 0$ then $x = -\infty$,

if $t = y_i$ then $x = \dfrac{ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}}{\sigma} = x(y_i)$,

we have

$$A = \int_{-\infty}^{x(y_i)} \frac{1}{\sqrt{2\pi}\cancel{\sigma^3}} x^2 \cancel{\sigma^2} e^{-\frac{x^2}{2}} \cancel{\sigma} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x(y_i)} x^2 e^{-\frac{x^2}{2}} dx.$$

We apply integration by parts with

$$dV = e^{-\frac{x^2}{2}} x dx = -e^{-\frac{x^2}{2}} d\left(-\frac{x^2}{2}\right), \qquad V = -e^{-\frac{x^2}{2}} U = x, \qquad dU = dx,$$

and get

$$\begin{aligned} A &= \frac{1}{\sqrt{2\pi}} \left\{ -x \cdot e^{-\frac{x^2}{2}} \Big|_{-\infty}^{x(y_i)} + \int_{-\infty}^{x(y_i)} e^{-\frac{x^2}{2}} dx \right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \left[-x(y_i) \cdot e^{-\frac{x(y_i)^2}{2}} + 0 \right] + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x(y_i)} e^{-\frac{x^2}{2}} dx \\ &= \frac{-(ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta})}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(ln(y_i)-\boldsymbol{z}_i\boldsymbol{\beta})^2}{2\sigma^2}} + \Phi_{St.N.}\left(\frac{ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}}{\sigma} \right), \end{aligned}$$

where $\Phi_{St.N.}(\cdot)$ is the CDF of the standard normal distribution. Therefore,

$$\begin{aligned} \frac{\partial r_{Y,i}}{\partial \beta_j} &= -(1 + \delta_i) \frac{z_{ji}}{\sigma\sqrt{2\pi}} e^{-\frac{\left(ln(y_i)-\boldsymbol{z}_i\boldsymbol{\beta}\right)^2}{2\sigma^2}}, \\ \frac{\partial r_{Y,i}}{\partial \gamma}\Bigg|_{\gamma = ln(\sigma)} &= (1 + \delta_i) \left[-\left(ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}\right) \cdot \phi_{St.N.}\left(\frac{ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}}{\sigma} \right) \right. \\ &\qquad \left. + \Phi_{St.N.}\left(\frac{ln(y_i) - \boldsymbol{z}_i\boldsymbol{\beta}}{\sigma} \right) - F_Y(y_i) \right], \end{aligned}$$

where $\phi_{St.N.}$ and $\Phi_{St.N.}$ are the PDF and CDF of the standard Normal distribution, and $F_Y(y_i)$ is the CDF of the log-normal distribution with $\mu = \boldsymbol{z}_i\boldsymbol{\beta}$ and σ.

3.10.4 Estimating equations and derivatives of PSRs for Cox model

For Cox model, we have

$$\begin{aligned}
S(y_i) &= e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}},\\
\Lambda(y_i) &= \sum_{k:t_{(k)}\leq y_i} \left(t_{(k)} - t_{(k-1)}\right)\lambda_k(y_i),\\
f(y_i) &= -\frac{\partial S(y_i)}{\partial y_i} = e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\cdot e^{\boldsymbol{z}_i\boldsymbol{\beta}}\cdot\frac{\partial\Lambda(y_i)}{\partial y_i} = e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\cdot e^{\boldsymbol{z}_i\boldsymbol{\beta}}\cdot\lambda_k(y_i),\\
\text{Ł}_i &= f(y_i)^{\delta_i}S(y_i)^{1-\delta_i} = \left[\frac{f(y_i)}{S(y_i)}\right]^{\delta_i}S(y_i) = \left[\lambda_k(y_i)\cdot e^{\boldsymbol{z}_i\boldsymbol{\beta}}\right]^{\delta_i}e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}},\\
\text{ł}_i &= \delta_i[ln(\lambda_k(y_i)) + \boldsymbol{z}_i\boldsymbol{\beta}] - \Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}} = \delta_i[ln(\lambda_k(y_i)) + \boldsymbol{z}_i\boldsymbol{\beta}] - \sum_{k:t_{(k)}\leq y_i}\left(t_{(k)} - t_{(k-1)}\right)\lambda_k(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}.
\end{aligned}$$

The first derivatives are

$$\begin{aligned}
\frac{\partial \text{ł}_i}{\partial\beta_j} &= \delta_i z_{ij} - \left[\sum_{k:t_{(k)}\leq y_i}(t_{(k)} - t_{(k-1)})\lambda_k(y_i)\right]e^{\boldsymbol{z}_i\boldsymbol{\beta}}z_{ij},\\
\frac{\partial \text{ł}_i}{\partial\lambda_k(y_i)} &= \frac{\delta_i}{\lambda_k(y_i)} - \left(t_{(k)} - t_{(k-1)}\right)e^{\boldsymbol{z}_i\boldsymbol{\beta}} && i,k: \quad y_i = t_{(k)},\\
&= -\left(t_{(k)} - t_{(k-1)}\right)e^{\boldsymbol{z}_i\boldsymbol{\beta}} && i,k: \quad y_i > t_{(k)},\\
&= 0 && i,k: \quad y_i < t_{(k)}.
\end{aligned}$$

The second derivatives are

$$\begin{aligned}
\frac{\partial^2 \text{ł}_i}{\partial\beta_j\partial\beta_{j'}} &= -\left[\sum_{k:t_{(k)}\leq y_i}\left(t_{(k)} - t_{(k-1)}\right)\lambda_k(y_i)\right]e^{\boldsymbol{z}_i\boldsymbol{\beta}}z_{ij}z_{ij'},\\
\frac{\partial^2 \text{ł}_i}{\partial\lambda_k(y_i)\partial\lambda_{k'}(y_i)} &= -\frac{\delta_i}{\lambda_k^2(y_i)} && i,k: \quad k = k' \cap y_i = t_{(k)},\\
&= 0 && \text{otherwise},\\
\frac{\partial^2 \text{ł}_i}{\partial\beta_j\partial\lambda_k(y_i)} &= -\left(t_{(k)} - t_{(k-1)}\right)e^{\boldsymbol{z}_i\boldsymbol{\beta}}z_{ij} && i,k: \quad y_i \geq t_{(k)},\\
&= 0 && \text{otherwise}.
\end{aligned}$$

We recall that $\widehat{r}_{Y,i} = \widehat{F}(y_i) + \delta_i \widehat{F}(y_i^-) - \delta_i = 1 - \widehat{S}(y_i) - \delta_i \widehat{S}(y_i^-)$ and

$$\begin{aligned}
\frac{\partial S_{Y_i}}{\partial \beta_j} &= -e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}} \cdot \Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}} \cdot z_{ij} \\
\frac{\partial S_{Y_i}}{\partial \lambda_k} &= -e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}} \cdot e^{\boldsymbol{z}_i\boldsymbol{\beta}}(t_{(k)} - t_{(k-1)}) \cdot \mathbb{1}(y_i \geq t_k).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial r_{Y,i}}{\partial \beta_j} &= \left(e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\Lambda(y_i) + \delta_i e^{-\Lambda(y_i^-)e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\Lambda(y_i^-)\right) \cdot e^{\boldsymbol{z}_i\boldsymbol{\beta}} \cdot z_{ij}, \\
\frac{\partial r_{Y,i}}{\partial \lambda_k} &= \left(e^{-\Lambda(y_i)e^{\boldsymbol{z}_i\boldsymbol{\beta}}} + \delta_i e^{-\Lambda(y_i^-)e^{\boldsymbol{z}_i\boldsymbol{\beta}}}\right) e^{\boldsymbol{z}_i\boldsymbol{\beta}}(t_{(k)} - t_{(k-1)}) \qquad i,k: \quad y_i \geq t_{(k)}, \\
&= 0 \qquad \text{otherwise.}
\end{aligned}$$

If we assume continuity and denote martingale residuals as $m(y_i)$, we have:

$$\begin{aligned}
\frac{\partial r_{Y,i}}{\partial \beta_j} &= (1 + \delta_i)\left(\delta_i - m(y_i)\right) e^{m(y_i) - \delta_i} \cdot z_{ij} \\
\frac{\partial r_{Y,i}}{\partial \lambda_k} &= (1 + \delta_i)e^{m(y_i) - \delta_i} e^{\boldsymbol{z}_i\boldsymbol{\beta}}(t_{(k)} - t_{(k-1)}) \qquad i,k: \quad y_i \geq t_{(k)}, \\
&= 0 \qquad \text{otherwise.}
\end{aligned}$$

CHAPTER 4

EXTENSION TO R PACKAGE PRESIDUALS: UNADJUSTED, PARTIAL, AND CONDITIONAL CORRELATION WITH BIVARIATE SURVIVAL DATA

4.1 Abstract

We present new methods for the analysis of correlation with bivariate survival data. First, we review our methods and then illustrate their use by analyzing the correlation of time to retinopathy in treated and untreated eyes in patients with diabetes. These methods are implemented as part of the R package **PResiduals** (Dupont et al., 2018).

Keywords: Spearman's correlation, Bivariate survival data, Probability Scale Residuals, covariate-adjustment.

4.2 Introduction

In Chapters 2 and 3, we have proposed several methods for estimating Spearman's rank correlation with right-censored data. The first method, restricted Spearman's correlation $\rho_{S|\Omega_R}$, computes the correlation only within a restricted region supported by the data. The second method, the highest rank Spearman's correlation ρ_S^H, computes the correlation on the entire probability space; it implicitly assigns the observations censored beyond the maximum observed event time to be the same highest rank value. Both $\widehat{\rho}_{S|\Omega_R}$ and $\widehat{\rho}_S^H$ (see Chapter 2) require estimating the joint bivariate survival distribution non-parametrically. We also proposed a third non-parametric correlation estimator that does not require estimating the joint survival distributions (see Chapter 3). Using marginal distributions only, this method computes the correlation of probability scale residuals (PSRs), ρ_{PSR}, with censored data. Lastly, following the semi-parametric approach of Liu et al. (2018), we extended ρ_{PSR} to compute adjusted (partial), $\rho_{PSR \cdot \boldsymbol{Z}}$, and conditional, $\rho_{PSR|\boldsymbol{Z}}$, estimators of correlations for right-censored data (see Chapter 3). Although estimators $\rho_{PSR \cdot \boldsymbol{Z}}$ and $\rho_{PSR|\boldsymbol{Z}}$ are semi-parametric, their only assumptions are about the form of the marginal distributions as functions of the adjustment covariates.

These estimators were extensively studied in Chapters 2 and 3 and shown to have various strengths and limitations. We have implemented these methods in the

PResiduals package of the R statistical language (Dupont et al., 2018; R Core Team, 2017). Here, we introduce these methods and software so that other analysts can apply them. In Section 4.3, we review our methods. In Section 4.4, we provide detailed examples of their use with the **PResiduals** package by estimating in various manners the rank correlation between the times to retinopathy in the right and left eyes of patients with diabetes (see package `SurvCorr`, Ploner et al. (2013)). We conclude with Section 4.5.

4.3 Review of methods

We are interested in the correlation between two time-to-event variables, T_X and T_Y, possibly right-censored. Times to events can be observed on a single subject or a pair of subjects. We assume independence between (T_X, T_Y) and times to censoring (C_X, C_Y). However, C_X and C_Y can be dependent. As a result of censoring, we only observe $X = \min(T_X, C_X)$ and $Y = \min(T_Y, C_Y)$ and event indicators $\Delta_X = \mathbb{1}(T_X \leq C_X)$ and $\Delta_Y = \mathbb{1}(T_Y \leq C_Y)$. For events observed on a single subject, censoring is likely to be *univariate* (C_X equals C_Y with probability one). When $C_X \neq C_Y$, the censoring is *bivariate.* In practice, follow-up is often bounded by design. We denote the maximum follow-up times for T_X and T_Y as τ_X and τ_Y, respectively. We refer to end-of-study censoring as type I censoring. Here we do not distinguish between *strict* type I censoring (all subjects start the study at the same calendar time and are observed until the end of the study) and *generalized* type I censoring (subjects can start the study at different calendar times and can be censored before the end of the study). Although rare in practice, some studies can have *unbounded* censoring when $\tau_X = \infty$ and $\tau_Y = \infty$ and censoring may happen throughout the study period. We denote marginal and joint cumulative distribution functions of T_X and T_Y as $F_X(x) = \Pr(T_X \leq x)$, $F_Y(y) = \Pr(T_Y \leq y)$, $F(x, y) = \Pr(T_X \leq x, T_Y \leq y)$, and marginal and joint survival functions as $S_X(x) = \Pr(T_X > x)$, $S_Y(y) = \Pr(T_Y > y)$, $S(x, y) = \Pr(T_X > x, T_Y > y)$. We define $F_X(x^-) = \lim_{t \uparrow x} F_X(t)$ and $F(x^-, y) = \lim_{t \uparrow x} F(t, y)$; functions $F_Y(y^-)$ and $F(x, y^-)$ are defined similarly.

The population parameter of Spearman's correlation can be presented as

$$\begin{aligned} \rho_S / c_\rho &= \mathrm{E}_{T_X, T_Y}\left[\left\{F_X(T_X) + F_X(T_X^-) - 1\right\}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right] \\ &= \int_0^\infty \int_0^\infty \left\{F_X(x) + F_X(x^-) - 1\right\}\left\{F_Y(y) + F_Y(y^-) - 1\right\} F(dx, dy), \end{aligned} \tag{4.1}$$

where $c_\rho = \left[\mathrm{Var}\left\{F_X(T_X) + F_X(T_X^-) - 1\right\} \mathrm{Var}\left\{F_Y(T_Y) + F_Y(T_Y^-) - 1\right\}\right]^{-1/2}$, and $c_\rho = 3$ when T_X and T_Y are continuous (Liu et al., 2018). The right-hand side of (4.1) is the covariance of probability-scale residuals (PSRs) proposed and studied by Li and Shepherd (2012) and Shepherd et al. (2016) and defined as:

$$r_X(t_X, F_X) = \mathrm{E}\left\{\mathrm{sign}(t_X - T_X)\right\} = \Pr(T_X < t_X) - \Pr(T_X > t_X) = F_X(t_X^-) + F_X(t_X) - 1,$$

where $\mathrm{sign}(t_X - T_X)$ is -1, 0, and 1 for $t_X < T_X$, $t_X = T_X$, and $t_X > T_X$ respectively. We can rewrite definition (4.1) in terms of survival functions:

$$\rho_S/c_\rho = \int_0^\infty \int_0^\infty \left\{1 - S_X(x) - S_X(x^-)\right\}\left\{1 - S_Y(y) - S_Y(y^-)\right\} S(dx, dy).$$

4.3.1 Restricted region Spearman's correlation, $\rho_{S|\Omega_R}$

When the follow-up period is limited either by design or by the fact that the last observed time to event is censored, non-parametric estimation of the overall correlation is not possible. It is possible, however, to estimate the correlation within the follow-up period, $\Omega = [0, \tau_X) \times [0, \tau_Y)$, or within some subregion of the follow-up period, $\Omega_R \subseteq \Omega$, (see Chapter 2):

$$\rho_{S|\Omega_R}/c_{\rho|\Omega_R} = \iint\limits_{\Omega_R} \left\{1 - S_X(x|\Omega_R) - S_X(x^-|\Omega_R)\right\}\left\{1 - S_Y(y|\Omega_R) - S_Y(y^-|\Omega_R)\right\} S(dx, dy|\Omega_R), \tag{4.2}$$

where $c_{\rho|\Omega_R} = \left[\mathrm{Var}\left\{1 - S_X(x|\Omega_R) - S_X(x^-|\Omega_R\right\}\mathrm{Var}\left\{1 - S_Y(y|\Omega_R) - S_Y(y^-|\Omega_R)\right\}\right]^{-1/2}$; and $S_X(x|\Omega_R)$, $S_Y(y|\Omega_R)$, and $S(x, y|\Omega_R)$ are conditional marginal and joint survival functions of T_X and T_Y. In general, parameter $\rho_{S|\Omega_R}$ is not the same as the overall Spearman's correlation, but it is well defined and interpreted as Spearman's correlation within Ω_R. $\rho_{S|\Omega_R}$ is computed by plugging in modified estimators of the marginal and joint distributions conditional on Ω_R using estimators proposed by Kaplan and Meier (1958) and Dabrowska (1988). Confidence intervals for $\rho_{S|\Omega_R}$ are estimated using the bootstrap.

4.3.2 Highest rank Spearman's correlation, ρ_S^H

Although the correlation structure outside of Ω is not estimable, we can approximate the overall correlation using ρ_S^H, the highest rank Spearman's correlation. The name *highest rank* Spearman's correlation originates from the fact that for strict type I censoring, this estimator is equivalent to Spearman's correlation that assigns highest ranks to the observations censored outside of the restricted region. Estimator ρ_S^H keeps track of the observations censored outside of Ω by accounting for their probability mass, which is the left-over mass from the total mass of 1 and the mass within Ω. This left-over mass can be compared to dark matter in the universe. Dark matter cannot be observed but is an essential part of the gravitational fabric of the universe. So is the left-over probability mass of the observations censored outside of the restricted region. Although not identifiable for each observation, this aggregated mass provides critical information about the overall correlation. Technically, we account for the left-over probability mass by replacing $S(dx, dy)$ with a probability mass function $S^H(dx, dy)$ that is defined as $S(dx, dy)$ inside Ω, as $S(\tau_X^-, dy)$ for $x = \tau_X$ and $y < \tau_Y$, as $S(dx, \tau_Y^-)$ for $x < \tau_X$ and $y = \tau_Y$, as $S(\tau_X^-, \tau_Y^-)$ for $x = \tau_X$ and $y = \tau_Y$, and zero otherwise. Estimator ρ_S^H is defined as the correlation of PSRs for this new distribution:

$$\rho_S^H / c_\rho^H = \int_0^{\tau_X} \int_0^{\tau_Y} \left\{1 - S_X^H(x) - S_X^H(x^-)\right\} \left\{1 - S_Y^H(y) - S_Y^H(y^-)\right\} S^H(dx, dy),$$

where $S_X^H(x)$ and $S_Y^H(y)$ are the marginal survival functions of $S^H(x, y)$, and

$$c_\rho^H = \left[\mathrm{Var}\left\{1 - S_X^H(T_X) - S_X^H(T_X^-)\right\} \mathrm{Var}\left\{1 - S_Y^H(T_Y) - S_Y^H(T_Y^-)\right\}\right]^{-1/2}.$$

The estimator of ρ_S^H is consistent in the presence of unbounded censoring. Again, ρ_S^H is estimated by plugging in modified Kaplan-Meier and Dabrowska's estimators of the marginal and joint survival functions $S^H(x, y)$. For type I censoring, however, $\rho_S^H \neq \rho_S$. Still, ρ_S^H can be considered as an approximation of the overall Spearman's correlation - the best guess made without any parametric assumptions. The confidence intervals of ρ_S^H are estimated using the bootstrap.

4.3.3 Correlation of probability scale residuals, ρ_{PSR}

Another way to approximate Spearman's correlation is to use estimator ρ_{PSR}, the correlation of the probability scale residuals for right-censored data. As shown by Liu et al. (2018), in the absence of censoring, the population parameter of ρ_{PSR} is the Spearman's rank correlation defined in (4.1). For censored observations, Shepherd et al. (2016) extended the definition of PSRs to the expectation of PSRs over the area after the censoring point, $r(x, F_X, \Delta_X = 0) = E\{r(T_X, F_X) | T_X > x\}$; therefore, for right-censored data, $r(x, F_X, \delta_X) = F_X(x) - \delta_X(1 - F_X(x-))$, where (x, δ_X) is a realization of (X, Δ_X). So the population parameter of ρ_{PSR} in the presence of right-censoring is

$$\rho_{PSR}/c_{PSR} = \text{ Cov}\left\{r(X, F_X, \Delta_X),\ r(Y, F_Y, \Delta_Y)\right\}$$

where $c_{PSR} = \left[\text{Var}\left\{r(X, F_X, \Delta_X)\right\}\text{Var}\left\{r(Y, F_Y, \Delta_Y)\right\}\right]^{-1/2}$, and the covariance/variances are taken over the distribution of (X, Y) and (Δ_X, Δ_Y). Estimation of ρ_{PSR}, unlike ρ_S^H and $\rho_{S|\Omega_R}$ does not require estimating the joint bivariate survival distribution, which can be difficult in practice, particularly with small sample sizes. Hence, ρ_{PSR} is estimated using plug-in estimates based on Kaplan-Meier estimators. Unlike for $\widehat{\rho}_{S|\Omega_R}$ and $\widehat{\rho}_S^H$, confidence intervals for $\widehat{\rho}_{PSR}$ can be computed using M-estimation (Stefanski and Boos, 2002) (see the details in Chapter 3). When $\rho_S = 0$, ρ_{PSR} is also zero regardless of the censoring distribution. When $\rho_S \neq 0$ and censoring is present, $\rho_{PSR} \neq \rho_S$. However, with moderate censoring, the bias of $\widehat{\rho}_{PSR}$ is small, and its standard error is much lower compared to that of $\widehat{\rho}_S^H$; so $\widehat{\rho}_{PSR}$ can outperform $\widehat{\rho}_S^H$. Finally, ρ_{PSR} allows for easy extensions to partial and conditional correlations.

4.3.4 Partial correlation of probability scale residuals, $\rho_{PSR\cdot\boldsymbol{Z}}$

To evaluate whether or to what extent the correlation between two variables is induced by other variables, it is helpful to compute adjusted or *partial* rank correlation,

$$\rho_{PSR\cdot\boldsymbol{Z}}/c_{\rho\cdot\boldsymbol{Z}} = \text{ Cov}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)\right\},$$

where $c_{\rho\cdot\boldsymbol{Z}} = \left[\text{Var}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X)\right\}\text{Var}\left\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)\right\}\right]^{-1/2}$ and $F_{X|\boldsymbol{Z}},\ F_{Y|\boldsymbol{Z}}$ are the distributions of T_X and T_Y, respectively, conditional on covariates $\boldsymbol{Z}$ and can be easily estimated using available survival models.

When estimating $\widehat{F}_{X|\boldsymbol{Z}}$ and $\widehat{F}_{Y|\boldsymbol{Z}}$, the decision on what covariates to include should be made carefully. Since the goal of computing partial correlation is to remove the confounding effects of $\boldsymbol{Z}$ from the association, it makes sense to include covariates that affect both variables. When observations are collected on the same subject, both survival models will naturally have the same covariates with equal value for one pair of observations. When observations are collected on two different subjects, users have to be more careful. For example, each twin can have a body mass index (BMI) recorded, and the BMI may be different for each of the twins. Adjusting each model only for one twin's BMI means that the other twin's BMI does not affect the correlation, which defeats the purpose of adjusting away the effect of BMI from the correlation. Including both BMI variables in both models with result in correlation adjusted for both BMI variables.

Our software supports several survival regressions: exponential, Weilbull, log-normal, log-logistic, and Cox proportional hazards regressions. The confidence intervals of $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$ are computed using M-estimation (see Stefanski and Boos (2002); Liu et al. (2018); Chapter 3), which requires score equations for model parameters. These score equations can be obtained from the corresponding regression objects. For Cox regression, we implemented two ways of obtaining score equations depending on whether the variability of the estimated baseline hazard is taken account: 1) with partial likelihood and 2) with full likelihood suggested by Breslow (1972). The problem with the full likelihood approach is that the dimension of M-estimation matrices may grow linearly with the number of observations, which may significantly increase the computational time or result in singular matrices. When singularity occurs, the software switches to the partial likelihood approach. Although computing $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$ with partial likelihood will result in underestimating its variance, our simulations showed that these two approaches perform very similarly (see Chapter 3).

4.3.5 Conditional correlation of probability scale residuals, $\rho_{PSR|\boldsymbol{Z}}$

There may be interest in estimating the rank correlation between two time-to-event variables conditional (i.e., as a function of) a third variable. For example, we might be interested in the correlation of time to retinopathy as a function of $A1C$ level (a protein in red blood cells that carries oxygen and is coated with sugar). Higher $A1C$ levels mean poorer blood sugar control, which may affect the correlation. The correlation of PSRs can also be used to estimate the conditional correlation defined as

$$\rho_{PSR|\boldsymbol{Z}}/c_{\rho|\boldsymbol{Z}} = \text{Cov}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}\right\},$$

where $c_{\rho|\boldsymbol{Z}} = \left[\text{Var}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X)|\boldsymbol{Z}\right\}\text{Var}\left\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}\right\}\right]^{-1/2}$.
Partial-conditional correlation can also be defined as
$\rho_{PSR\cdot\boldsymbol{Z}_1|\boldsymbol{Z}_2}/c_{\rho\cdot\boldsymbol{Z}_1|\boldsymbol{Z}_2} = \text{Cov}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X),\ r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y)|\boldsymbol{Z}_2\right\}$, where
$c_{\rho\cdot\boldsymbol{Z}_1|\boldsymbol{Z}_2} = \left[\text{Var}\left\{r(X, F_{X|\boldsymbol{Z}}, \Delta_X|\boldsymbol{Z}_2)\right\}\text{Var}\left\{r(Y, F_{Y|\boldsymbol{Z}}, \Delta_Y|\boldsymbol{Z}_2)\right\}\right]^{-1/2}$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$.

Similar to $\rho_{PSR\cdot\boldsymbol{Z}}$, when estimating partial-conditional correlation, one should carefully consider what covariates to include. Regarding $\boldsymbol{Z}_2$, both models should include the same covariates. The method of computing confidence intervals for $\rho_{PSR|\boldsymbol{Z}}$ and $\rho_{PSR\cdot\boldsymbol{Z}_1|\boldsymbol{Z}_2}$ is similar to $\rho_{PSR\cdot\boldsymbol{Z}}$, but more involved (Liu et al., 2018).

4.4 Analysis of rank correlation

4.4.1 Diabetic retinopathy data

To illustrate our methods, we use the `diabetes` dataset available with the package `SurvCorr` (Ploner et al., 2015). This data contains 197 subjects diagnosed with diabetes mellitus who underwent a laser treatment on one of their eyes to prevent retinopathy, a disease that affects the retina's blood vessels and may result in loss of vision. These subjects represent a subset of a larger cohort of the Diabetic Retinopathy Study (DRS) (see National Eye Institute (1981)). According to the study protocol, one eye of each patient was randomly assigned to treatment with photocoagulation. The eye chosen for treatment was randomly assigned to either using an argon laser or a xenon photocoagulator. Patients were tested for visual acuity at 4-month intervals. The data contains time to diabetic retinopathy or censoring (in months), the eye (right (1) or left (2)) that the treatment was applied to, the type of treatment (xenon (1) or argon (2)), and age at diabetes diagnosis. The median follow-up time was 42.23 and 32.63 months in the treated and untreated eye, respectively. The proportions of censoring were 72.6% in the right eye, 48.7% in the left eye, 40.6% in both eyes; only 19.3% of subjects had events for both eyes. Figures 4.1 visualizes events and censoring events.

Our methods are temporarily contained in the `survSpearman` package (available on GitHub) and will become part of the package `PResiduals`. The package is dependent on the `survival` package (Therneau, 2015). We also call `SurvCorr` package (Ploner et al., 2013) to access the diabetic retinopathy data. The first five rows of the dataset are printed below.

```
>   library(survival)
>   library(SurvCorr)
>   library(survSpearman)
>   data(diabetes, package = "SurvCorr")

>  diabetes[1:5,]
  ID LASER TRT_EYE AGE_DX ADULT TIME1 STATUS1 TIME2 STATUS2
1  5     2       2     28     2 46.23       0 46.23       0
2 14     2       1     12     1 42.50       0 31.30       1
3 16     1       1      9     1 42.27       0 42.27       0
4 25     2       2      9     1 20.60       0 20.60       0
```

```
5 29     1       2      13      1 38.77       0  0.30       1
```

The function `visualBivarTimeToEvent()` plots the bivariate survival data shown in Figure 4.1, distinguishing between fully observed, singly censored, and doubly censored observations.

```
>   visualBivarTimeToEvent(timeX = diabetes$TIME1, deltaX = diabetes$STATUS1,
+                          timeY = diabetes$TIME2, deltaY = diabetes$STATUS2,
+                          labelX = "Time to retinopathy in the treated eye",
+                          labelY = "Time to retinopathy in the untreated eye",
+                          dotSize = 1, segLength=2, legendCex = .8,
+                          scaleLegendGap = 2.5, xlim = c(0, 90), ylim= c(0, 90))
```

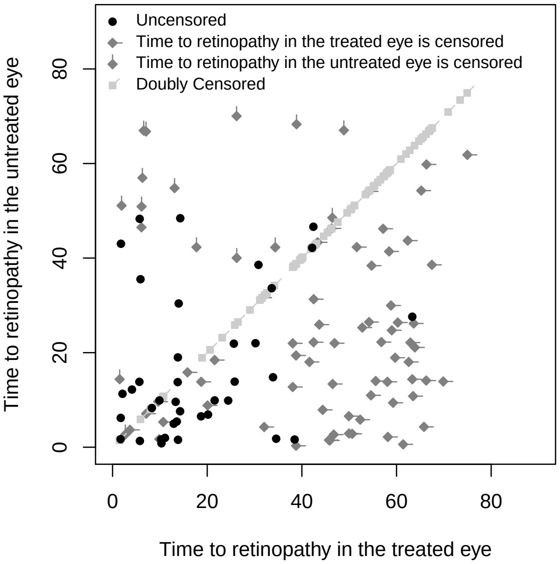

Figure 4.1: Diabetic retinopathy data.

From Figure 4.1 it is hard to see if the data are correlated because of censoring. Viewing the bivariate probability mass function may help visualize the correlation structure better. The `survPMFPlot()` function plots the estimated bivariate probability mass function of Dabrowska (1988). This function takes as input an estimate of Dabrowska's bivariate survival distribution, which is obtained by function `survDabrowska()`, illustrated in the code below. The `survPMFPlot()` function aggregates probability mass into user-defined cells using arguments `gridWidthX` and `gridWidthY`, which in the code below, resulting in Figure 4.2, aggregates probability mass into cells of 2 × 2 months.

The darker shade in the figure corresponds to regions with higher probability mass. The estimated marginal probability mass is shown in the bottom and left margins of the plot. The arguments scaleHistX and scaleHistY regulate the height of the marginal histograms, and arguments scaleGapX and scaleGapY regulate the gap between the main plot and the axes' labels.

```
>   dabrSurface = survDabrowska(X = diabetes$TIME1, Y = diabetes$TIME2,
+                  deltaX = diabetes$STATUS1, deltaY = diabetes$STATUS2)$DabrowskaEst
>   survPMFPlot(dabrSurface, gridWidthX = 2, gridWidthY = 2,
+                XAxisLabel = "Time to retinopathy for the treated eye",
+                YAxisLabel = "Time to retinopathy for the untreated eye",
+                scaleHistX = 20, scaleHistY = 20,
+                scaleGapX = 1.5, scaleGapY = 1.5,
+                labelSkipX = 1, labelSkipY = 1,
+                lineXAxisLabel = -0.5, lineYAxisLabel = -0.25)
```

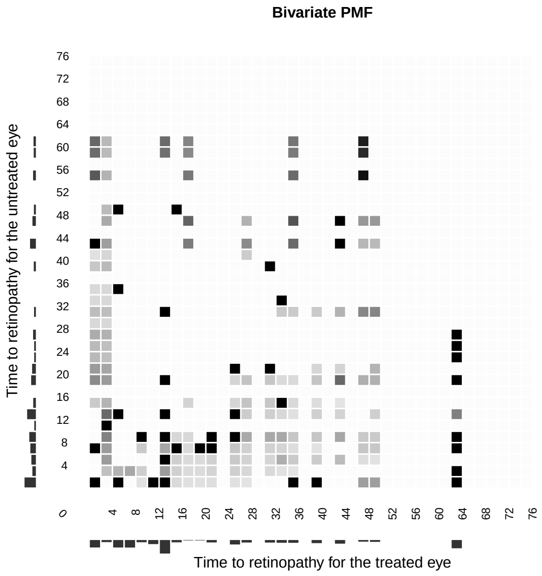

Figure 4.2: Probability mass function for diabetic retinopathy data. The data was aggregated into cells of 2 × 2 months.

Figure 4.2 shows that larger probability mass is loosely concentrated along the diagonal, suggesting some positive correlation. The lighter gray points located closer to the axes represent the probability mass estimated mainly from the censored events. The Kaplan-Meier plot helps visualize the marginal survival distributions and is shown in Figure 4.3. According to Figure 4.3, the risk of having retinopathy event appears to be smaller in the treated compared to the untreated eye.

```
>     fit1 <- survfit(Surv(TIME1, STATUS1)~1, data = diabetes)
>     fit2 <- survfit(Surv(TIME2, STATUS2)~1, data = diabetes)
>     plot(0, 0, type = "n", xlab = "Time to event", cex = 1, col = "black",
+         ylab = "Survival probability", xlim = range(c(fit1$time, fit2$time)),
+         ylim = c(0, 1), axes = TRUE)
>     polygon(x=c(fit1$time, fit1$time[length(fit1$time):1]),
+         y=c(fit1$upper, fit1$lower[length(fit1$lower):1]),
+         col = "#22222222", border="#22222222")
>     lines(fit1$time, fit1$surv, col="black")
>     polygon(x=c(fit2$time, fit2$time[length(fit2$time):1]),
+         y=c(fit2$upper, fit2$lower[length(fit2$lower):1]), col = "#22222222", border="#22222222")
>     lines(fit2$time, fit2$surv, col="black")
>     text(x = c(max(fit1$time), max(fit1$time)) - 7,
+         y = c(min(fit1$surv), min(fit2$surv))+ 0.05, labels = c("Treated", "Untreated"))
```

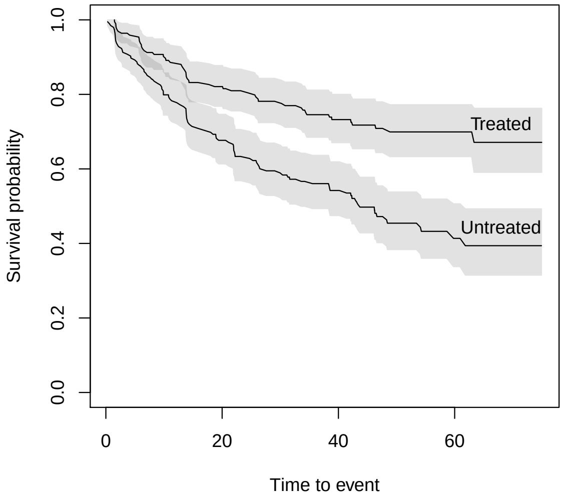

Figure 4.3: Kaplan-Meier estimates of the probability of not having a retinopathy in the treated and untreated eyes.

4.4.1.1 Unadjusted correlation with $\widehat{\rho}_{S|\Omega_R}$, $\widehat{\rho}^H_S$, and $\widehat{\rho}_{PSR}$

In this exercise, we are interested in estimating the rank correlation between the times to retinopathy in the treated and untreated eyes. In spite of the fact that the treatment appears to be effective, we expect that the time to retinopathy in one eye is positively correlated with that in the other eye. To evaluate this, we compute the unadjusted rank correlation between the times to retinopathy in the treated and untreated eyes. Because the maximum observed time for each eye corresponds to a censored event,

```
>   diabetes[diabetes$TIME1 >= max(diabetes$TIME1), c("TIME1", "STATUS1")]
    TIME1 STATUS1
183 74.97       0
```

```
>   diabetes[diabetes$TIME2 >= max(diabetes$TIME2), c("TIME2", "STATUS2")]
    TIME2 STATUS2
141 74.93       0
```

the overall Spearman's correlation, ρ_S, is not estimable non-parametrically. We can either approximate the overall correlation by computing $\widehat{\rho}_S^H$ or focus on estimating Spearman's correlation in an estimable region and compute $\widehat{\rho}_{S|\Omega_R}$. Function `survSpearman()` computes both estimates:

```
>   est = survSpearman(X = diabetes$TIME1, Y = diabetes$TIME2,
+                deltaX = diabetes$STATUS1, deltaY = diabetes$STATUS2,
+                tauX = Inf, tauY = Inf)
>   est
$‘Restricted region set by user‘
 tauX  tauY
"Inf" "Inf"

$‘Effective restricted region‘
     tauX      tauY
"63.33 +" "61.83 +"

$Correlation
HighestRank  Restricted
  0.2675704   0.2511821
```

The output of function `survSpearman()` reports the estimated highest rank and restricted region rank correlations. The output also includes the restricted region set by the user, which defaults to no restrictions, $[0, \infty] \times [0, \infty]$, and the effective restricted region, which is defined by the maximum observed event times in the restricted region set by the user. In this example, these values are 63.33 and 61.83 months. The plus signs denote the fact that the highest rank correlation assigns probability mass to values just outside this region. Estimate $\widehat{\rho}_S^H$ incorporates this information as the probability mass of the observations censored outside of the restricted region. Estimate $\widehat{\rho}_{S|\Omega_R}$ focuses only on the effective restricted region and disregards the partial information provided by the observations censored outside of Ω.

Confidence intervals of these estimates are computed using the bootstrap:

```
>   bootCI = matrix(NA, nrow = 200, ncol = 2)
>   colnames(bootCI) = c("HighestRank", "Restricted")
>   set.seed(2020)
>   for (i in 1:nrow(bootCI)){
+     bsample = diabetes[sample(1:nrow(diabetes), nrow(diabetes), replace = TRUE), ]
+     bootCI[i, ] =  survSpearman(X = bsample$TIME1, Y = bsample$TIME2,
+                                 deltaX = bsample$STATUS1, deltaY = bsample$STATUS2,
+                                 tauX = Inf, tauY = Inf)$Correlation
+   }
>   CIs = apply(bootCI, 2, quantile, prob = c(0.025, 0.975))
>   res = rbind(est, CIs)
>   t(round(res, 3))
              est   2.5% 97.5%
HighestRank 0.268  0.091 0.412
Restricted  0.251 -0.149 0.589
```

In this example, the two estimates of correlation are quite similar, with both suggesting a moderate correlation between the time to retinopathy. Notice that the confidence interval for $\rho_{S|\Omega_R}$ is wider than that of ρ_S^H because the former ignores information outside the restricted region, and this region is likely to be different for different bootstrap samples.

We can also compute $\widehat{\rho}_{S|\Omega_R}$ in a smaller restricted region if such restricted correlation makes clinical sense. For example, suppose we are interested in the correlation during the first four years, $\Omega_{48} = [0, 48) \times [0, 48)$. This is implemented by specifying `tauX` = `tauY` = 48 in `survSpearman()`.

```
>   tauX = tauY = 48
>   est = survSpearman(X = diabetes$TIME1, Y = diabetes$TIME2,
+                deltaX = diabetes$STATUS1, deltaY = diabetes$STATUS2,
+                tauX = tauX, tauY = tauY)
>   bootCI = rep(NA, 200)
>   set.seed(2021)
>   for (i in 1:length(bootCI)){
+     bsample = diabetes[sample(1:nrow(diabetes), nrow(diabetes), replace = TRUE), ]
+     bootCI[i] =  survSpearman(X = bsample$TIME1, Y = bsample$TIME2,
+                deltaX = bsample$STATUS1, deltaY = bsample$STATUS2,
+                tauX = tauX, tauY = tauY)$Correlation["Restricted"]
+   }
>   res = c(est = est$Correlation["Restricted"], quantile(bootCI, prob = c(0.025, 0.975)))
>   round(res, 3)
est.Restricted           2.5%          97.5%
         0.299         -0.084          0.649
```

The resulting estimate is about 0.299, but the confidence interval contains zero. Note that the function also outputs $\widehat{\rho}_S^H$ within the more narrow restricted region, but computing $\widehat{\rho}_S^H$ while artificially disregarding the information about the overall correlation makes less sense.

Similarly to $\widehat{\rho}_S^H$, estimator $\widehat{\rho}_{PSR}$ provides a way to approximate the overall Spearman's correlation. The function `unadjusted.CorPSRs()` computes $\widehat{\rho}_{PSR}$ and its 95% confidence interval.

```
>   res = unadjusted.CorPSRs(X = diabetes$TIME1, Y = diabetes$TIME2,
+                    deltaX = diabetes$STATUS1, deltaY = diabetes$STATUS2)
>   round(res[ c("est", "lower.CI", "upper.CI") ], 3)
     est lower.CI upper.CI
   0.271    0.132    0.410
```

Estimators $\widehat{\rho}_S^H$ and $\widehat{\rho}_{PSR}$ have similar point estimates with $\widehat{\rho}_{PSR}$ having a narrower confidence interval because it avoids estimating the bivariate survival surface, which adds to the variability of $\widehat{\rho}_S^H$. Obtaining the confidence interval of $\widehat{\rho}_{PSR}$ using M-estimation also saves computational time.

4.4.1.2 Partial correlation with $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$

The correlation can be affected by different factors, e.g., age, what eye was treated, and treatment type. We can adjust for these other variables by estimating the partial correlation, $\widehat{\rho}_{PSR\cdot Z}$. To compute the partial correlation, we first need to fit separate models for both of the time to event outcomes on the covariates. Here we fit Cox models:

```
>   survObjX = Surv(diabetes$TIME1, diabetes$STATUS1)
>   survObjY = Surv(diabetes$TIME2, diabetes$STATUS2)
>   modX = coxph(survObjX ~ TRT_EYE + AGE_DX + LASER,
+                data=diabetes, method = "breslow", timefix = FALSE)
>   modY = coxph(survObjY ~ TRT_EYE + AGE_DX + LASER,
+               data=diabetes, method = "breslow", timefix = FALSE)
```

The function `partial.corPSRs()` then takes these model objects as input and computes the partial correlation:

```
>   round(partial.corPSRs(modX, modY)[ c("est", "lower.CI", "upper.CI") ], 3)
     est lower.CI upper.CI
   0.301    0.162    0.441
```

In this example, the correlation slightly increased after adjusting for age, treatment eye, and type of treatment.

When the survival probabilities are modeled using Cox proportional hazards model, the confidence interval of $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}}$ is computed using score equations from the Cox regression partial likelihood. Using partial likelihood results in slightly underestimated variability of ρ_{PSR} because it does not take into account the variability of the baseline hazard. The user can choose the full likelihood option instead:

```
>   round(partial.corPSRs(modX, modY, likelihood = "full")[c("est", "lower.CI", "upper.CI") ], 3)
     est lower.CI upper.CI
   0.301    0.163    0.439
```

The confidence intervals obtained from the full and partial likelihoods are almost the same. Note that Cox regression should be fit with `timefix = FALSE` because of issues related to the floating point round-off error of the time to event stored in the Cox regression object.

Instead of Cox proportional hazards, parametric survival models can be used, for example, the log-logistic model:

```
>   modX = survreg(survObjX ~ AGE_DX, data=diabetes, dist = "loglogistic")
>   modY = survreg(survObjY ~ AGE_DX, data=diabetes, dist = "loglogistic")
>   round(partial.corPSRs(modX, modY)[ c("est", "lower.CI", "upper.CI") ], 3)
      est lower.CI upper.CI
    0.290    0.153    0.427
```

The correlation computed using log-logistic model is very similar.

4.4.1.3 Conditional and partial-conditional correlation with $\widehat{\rho}_{PSR|\boldsymbol{Z}}$ *and* $\widehat{\rho}_{PSR\cdot\boldsymbol{Z}_1|\boldsymbol{Z}_2}$

In addition to the partial correlation, there is an interest in computing correlation conditional on other variables. Suppose we are interested in estimating the rank correlation conditional on age at diagnosis. The function `conditional.corPSRs()` can be used to estimate $\rho_{PSR|Z}$. The function requires inputting models of the times-to-event conditional on age; Cox models are a natural choice and are implemented below. Since age is a continuous variable, `conditional.corPSRs()` allows the user to specify how to model the rank correlation. Specifically, $\rho_{PSR|Z}$ can be modeled linearly (`numKnots` = 0) or using restricted cubic splines with `numKnots` set to a whole number between 3 and 7. The location of the spline knots is defined in terms of quantiles suggested by Harrell Jr (2015). The code below estimates and plots $\rho_{PSR|Z}$ with restricted cubic splines with three knots.

```
>   modXC = coxph(survObjX ~ AGE_DX, data=diabetes, method = "breslow", timefix = FALSE)
>   modYC = coxph(survObjY ~ AGE_DX, data=diabetes, method = "breslow", timefix = FALSE)
>   z = diabetes[["AGE_DX"]]
>   newZ = diabetes[["AGE_DX"]]
>   par(mfrow = c(1, 2))
>   for(n_knots in c(0, 3)){
+     resultXY = conditional.corPSRs(modXC, modYC, z, newZ, numKnots = n_knots)
+     plotData = data.frame(x = newZ, y = resultXY$est,
+       yLower = resultXY$lower.CI, yUpper = resultXY$upper.CI)
+     plotData = plotData[order(plotData$x),]
+     plot(plotData$x, plotData$y, type = "n", ylim = c(0, 1))
+     points(plotData$x, plotData$y, pch = 19, col = plotCol, cex = .3)
+     lines(plotData$x, plotData$yLower, col = plotCol)
+     lines(plotData$x, plotData$yUpper, col = plotCol)
+     abline(h=0, lty = 3, col = plotCol)
+   }
```

Note that the current version of function `conditional.corPSRs()` computes correlation conditional only on one variable.

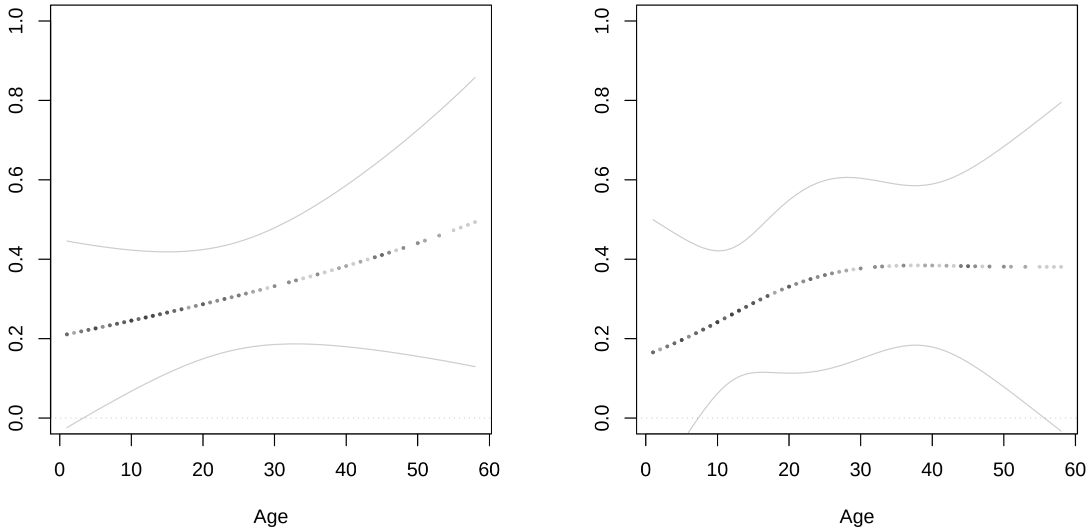

Figure 4.4: $\widehat{\rho}_{PSR|Z}$ conditional on age at diagnosis. Left: age is modeled as a linear variable. Right: age is modeled with restricted cubic splines with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles.

Note that argument `z` contains variable Z, and `newZ` contains only those values of Z, for which $\rho_{PSR|Z}$ is computed. Figure 4.4 shows $\widehat{\rho}_{PSR|Z}$ computed as a function of age, where age is modeled as a linear (left panel) and quadratic (right panel) variable. The panels generally show the rank correlation increasing with age, suggesting that the association between times to retinopathy for different eyes in the same patient is likely stronger for older patients. Larger sample sizes are needed to determine the functional form of the correlation with greater precision.

The same function `conditional.corPSRs()` can also compute the partial-conditional correlation. We may want to estimate the rank correlation conditional on age after adjusting for which eye was treated (`TRT_EYE`) and the type of treatment (`LASER`). This can be estimated using the same code except inputting objects from models that include `TRT_EYE` and `LASER` covariates:

```
>   plotCol = "#44444444"
>   modXC = coxph(survObjX ~ TRT_EYE + AGE_DX + LASER, data=diabetes,
+                 method = "breslow", timefix = FALSE)
>   modYC = coxph(survObjY ~ TRT_EYE + AGE_DX + LASER, data=diabetes,
+                 method = "breslow", timefix = FALSE)
```

```
>   z = diabetes[["AGE_DX"]]
>   newZ = diabetes[["AGE_DX"]]
>   par(mfrow = c(1, 2))
>   for(n_knots in c(0, 3)){
+     resultXY = conditional.corPSRs(modXC, modYC, z, newZ, numKnots = n_knots)
+     plotData = data.frame(x = newZ, y = resultXY$est,
+       yLower = resultXY$lower.CI, yUpper = resultXY$upper.CI)
+     plotData = plotData[order(plotData$x),]
+     plot(plotData$x, plotData$y, type = "n", ylim = c(0, 1))
+     points(plotData$x, plotData$y, pch = 19, col = plotCol, cex = .3)
+     lines(plotData$x, plotData$yLower, col = plotCol)
+     lines(plotData$x, plotData$yUpper, col = plotCol)
+     abline(h=0, lty = 3, col = plotCol)
+   }
```

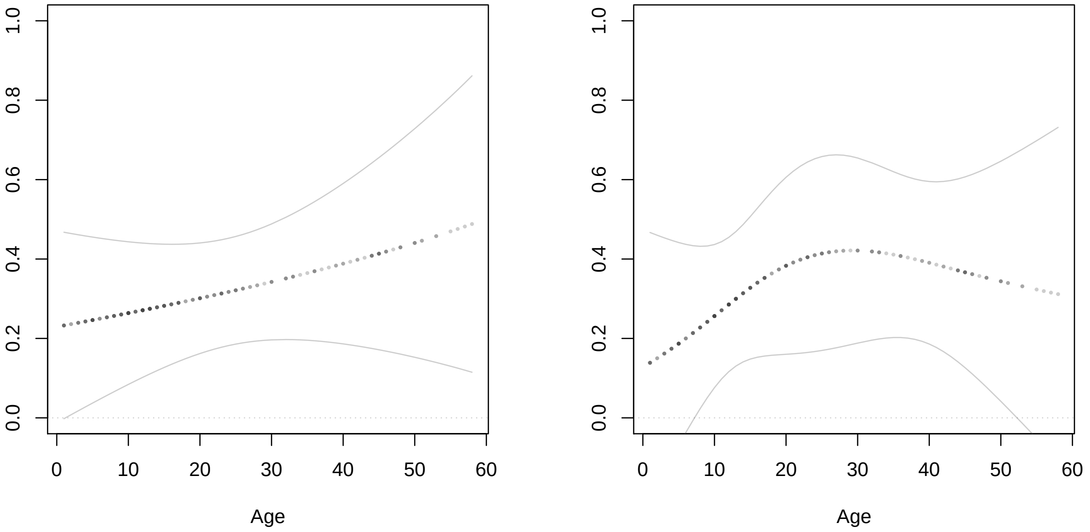

Figure 4.5: $\rho_{PSR \cdot Z_1 | Z_2}$ - correlation of PSRs adjusted for left/right eye and conditional on age at diagnosis. Left: age is modeled as a linear variable. Right: age is modeled as a restricted cubic spline with three knots at 0.1^{th}, 0.5^{th}, 0.9^{th} percentiles.

There is not much difference between the correlation conditional on age (Figure 4.4) and the partial correlation conditional on age (Figure 4.5).

To check if the correlation is the same, whether the right eye or left eye is treated or for the two types of treatment, we compute the partial rank correlation adjusted for age and conditional on the treated eye and treatment type.

```
>   par(mfrow = c(1, 2))
>   nameList = list(TRT_EYE = c("Right", "Left"), LASER = c("Xenon", "Argon"))
>   xList = list(TRT_EYE = 2:1, LASER = 1:2)
>   newZ = c(1, 2)
>   for(var_i in c("TRT_EYE", "LASER")){
+     z = diabetes[[var_i]]
+     resultXY = conditional.corPSRs(modXC, modYC, z, newZ, numKnots = 0)
+     resultXY[[var_i]] = newZ
+     resultXY[["NAMES"]] = nameList[[var_i]][newZ]
+     plot(0, 0, type = "n", xlim = c(0, 3), ylim = range(-0.2, 1),
+         axes = FALSE, xlab = "", ylab = "")
+     abline(h = 0, col = "gray", lty = 3)
+     box()
+     points(xList[[var_i]], resultXY[, "est"], pch = 18)
+     segments(x0 = xList[[var_i]], y0 = resultXY$lower.CI,
+             x1 = xList[[var_i]], y1 = resultXY$upper.CI)
+     axis(side = c(2), at = seq(-0.2, 1, .2))
+     axis(side = c(4), at = seq(-0.2, 1, .2))
+     mtext(resultXY$NAMES, side = 1, line = 1, at = xList[[var_i]])
+   }
```

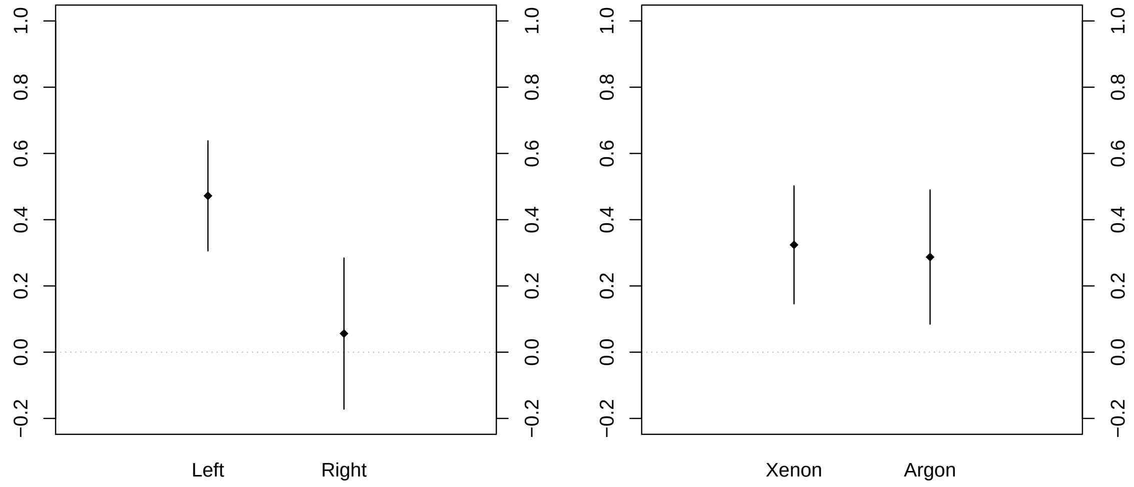

Figure 4.6: Partial-conditional correlation of PSRs. Left: $\rho_{PSR\cdot Z_1|Z_2}$, where Z_1 is age and treatment type, and Z_2 is whether the right or left eye was treated. Right: $\rho_{PSR\cdot Z_1|Z_2}$, where Z_1 is age and whether the right or left eye was treated, and Z_2 is the type of treatment.

Figure 4.6 shows estimates of the partial rank correlation conditional on eye (left panel) and treatment (right panel). Interestingly, it appears that the correlation between the times to retinopathy for different eyes within the same patient was higher when the left eye was treated than when the right eye was treated. In contrast, there appears to be no difference in correlation between the two treatment groups.

We can also compute $\widehat{\rho}_S^H$ for these subgroups; see the following code and Figure 4.7. From Figures 4.6 and 4.7, we observe that $\widehat{\rho}_{PSR\cdot Z_1|Z_2}$ and $\widehat{\rho}_S^H$ give consistent results. Note that the former estimate adjusts for covariates whereas the latter does not. Covariate adjustment for $\rho_{S|\Omega_R}$ and ρ_S^H has not been developed because it would require estimating bivariate survival distributions conditional on covariates, which is quite complicated.

```
> d1 = diabetes[diabetes$TRT_EYE == 1,]
> d2 = diabetes[diabetes$TRT_EYE == 2,]
> l1 = diabetes[diabetes$LASER == 1,]
> l2 = diabetes[diabetes$LASER == 2,]
>
> estD1 = survSpearman(X = d1$TIME1, Y = d1$TIME2, deltaX = d1$STATUS1, deltaY = d1$STATUS2)
> estD2 = survSpearman(X = d2$TIME1, Y = d2$TIME2, deltaX = d2$STATUS1, deltaY = d2$STATUS2)
> estL1 = survSpearman(X = l1$TIME1, Y = l1$TIME2, deltaX = l1$STATUS1, deltaY = l1$STATUS2)
> estL2 = survSpearman(X = l2$TIME1, Y = l2$TIME2, deltaX = l2$STATUS1, deltaY = l2$STATUS2)
>
> pEst = c(estD1$Correlation["HighestRank"], estD2$Correlation["HighestRank"],
+    estL1$Correlation["HighestRank"], estL2$Correlation["HighestRank"])
> names(pEst) = c("Right", "Left", "Xenon", "Argon")
>
> bootCI = matrix(NA, nrow = 200, ncol = 4)
> colnames(bootCI) = c("Right", "Left", "Xenon", "Argon")
> set.seed(238)
> for (i in 1:nrow(bootCI)){
+   bD1 = d1[sample(1:nrow(d1), nrow(d1), replace = TRUE), ]
+   bD2 = d2[sample(1:nrow(d2), nrow(d2), replace = TRUE), ]
+   bL1 = l1[sample(1:nrow(l1), nrow(l1), replace = TRUE), ]
+   bL2 = l2[sample(1:nrow(l2), nrow(l2), replace = TRUE), ]
+   estBD1 = survSpearman(X = bD1$TIME1, Y = bD1$TIME2, deltaX = bD1$STATUS1, deltaY = bD1$STATUS2)
+   estBD2 = survSpearman(X = bD2$TIME1, Y = bD2$TIME2, deltaX = bD2$STATUS1, deltaY = bD2$STATUS2)
+   estBL1 = survSpearman(X = bL1$TIME1, Y = bL1$TIME2, deltaX = bL1$STATUS1, deltaY = bL1$STATUS2)
+   estBL2 = survSpearman(X = bL2$TIME1, Y = bL2$TIME2, deltaX = bL2$STATUS1, deltaY = bL2$STATUS2)
+   bootCI[i, ] =  c(estBD1$Correlation["HighestRank"], estBD2$Correlation["HighestRank"],
+        estBL1$Correlation["HighestRank"], estBL2$Correlation["HighestRank"])
+ }
Restricted Spearman’s correlation was corrected.
Restricted Spearman’s correlation was corrected.
Warning messages:
1: In HighestRankAndRestrictedSpearman(bivarSurf, tauX = tauX, tauY = tauY) :
  Restricted Spearman’s correlation was corrected.
2: In HighestRankAndRestrictedSpearman(bivarSurfForHR, tauX = Inf,  :
  Restricted Spearman’s correlation was corrected.
>
> CIs = apply(bootCI, 2, quantile, prob = c(0.025, 0.975))
> res = rbind(pEst, CIs)
```

Note the warnings given by function `survSpearman()`. This means that the restricted Spearman’s correlation, $\widehat{\rho}_{S|\Omega_R}$ was corrected (see Chapter 2) to remedy the problem of negative mass of the Dabrowska’s estimator (Dabrowska, 1988; Pruitt, 1991). This problem did not affect $\hat{\rho}_S^H$, which is not as sensitive to the negative mass problem.

```
> pList = list(c("Right", "Left"), c("Xenon", "Argon"))
> par(mfrow = c(1, 2))
> xList = list(TRT_EYE = 2:1, LASER = 1:2)
> for(list_i in c(1, 2)){
+   x = xList[[list_i]]
+   plot(0, 0, type = "n", xlim = c(0, 3), ylim = range(-0.2, 1), axes = FALSE, xlab = "", ylab = "")
+   abline(h = 0, col = "gray", lty = 3)
+   listData = res[, pList[[list_i]]]
+   box()
+   points(x, listData["pEst", ], pch = 18)
+   segments(x0 = x, y0 = listData["2.5%", ], x1 = x, y1 = listData["97.5%", ])
+   axis(side = c(2), at = seq(-0.2, 1, .2))
+   axis(side = c(4), at = seq(-0.2, 1, .2))
+   mtext(colnames(listData), side = 1, line = 1, at = x)
+ }
```

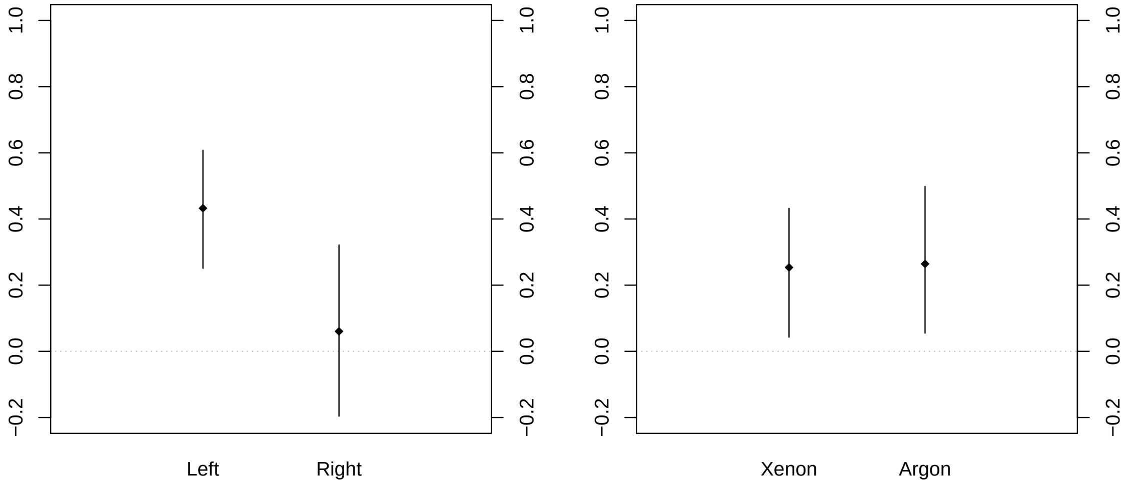

Figure 4.7: $\hat{\rho}_S^H$ for the right and left treated eye (left panel) and for Xenon and Argon treatment (right panel).

4.5 Summary

We have developed methods to estimate unadjusted, partial, and conditional rank correlations with bivariate survival data. These methods will be available as part of the R package **PResiduals**. We hope these methods will be a useful addition to the statistical toolbox of researchers working with bivariate survival data.

CHAPTER 5

CONCLUSION

In this dissertation, I proposed several non-parametric and semi-parametric methods of estimating unadjusted, partial, and conditional rank correlations with bivariate right-censored data. These methods are based on Spearman's rank correlation and, unlike previously suggested methods, do not make assumptions about the underlying correlation structure and, therefore, are less prone to bias.

Censoring presents serious technical and conceptual challenges when estimating correlation. For example, in the presence of end-of-study censoring, researchers may be faced with a dilemma: making more parametric assumptions and possibly getting a biased estimate (e.g., the semi-parametric method of Schemper et al. [2013]) or assuming less and being able to estimate an approximation to the desired parameter (e.g., the highest rank Spearman's correlation). There are situation when the parametric methods are justified and should be applied. However, estimating a well-interpreted approximation without making assumptions is also of practical value.

This work also illuminates bias-variance tradeoffs between various estimators. Specifically, when computing the highest rank Spearman's correlation by plugging in an estimator of the bivariate survival distribution, we obtain a consistent estimator of Spearman's rank correlation in the setting of unbounded censoring, but this estimator tends to have much greater variability and a larger mean-squared error than the correlation of probability scale residuals, which does not require an estimator of the bivariate survival distribution but is biased for Spearman's correlation. The choice between a more variable or a more biased estimator is not always obvious. Still, I hope that this work provides some guidance on how to navigate the analysis of correlation with bivariate survival data.

Lastly, I hope that making these methods available using open-source software will lead to their application in biomedical research, where bivariate survival data are not uncommon. These methods and software are useful tools to include in researchers' analytical toolboxes.

REFERENCES

Belzile, L. and Genest, C. (2017), *lcopula: Liouville Copulas*. R package version 1.0, https://CRAN.R-project.org/package=lcopula (accessed July 9, 2019). **URL:** *https://CRAN.R-project.org/package=lcopula*

Breslow, N. E. (1972), Discussion of professor coxâĂŹs paper, *J Royal Stat Soc B* **34**, 216–217.

Campbell, G. (1981), Nonparametric bivariate estimation with randomly censored data, *Biometrika* **68**(2), 417–422.

Carriere, J. F. (2000), Bivariate survival models for coupled lives, *Scandinavian Actuarial Journal* **2000**(1), 17–32.

Cesar, C., Jenkins, C. A., Shepherd, B. E., Padgett, D., Mejía, F., Ribeiro, S. R., Cortes, C. P., Pape, J. W., Madero, J. S., Fink, V. et al. (2015), Incidence of virological failure and major regimen change of initial combination antiretroviral therapy in the Latin America and the Caribbean: an observational cohort study, *The Lancet HIV* **2**(11), e492–e500.

Clayton, D. and Cuzick, J. (1985), Multivariate generalizations of the proportional hazards model, *Journal of the Royal Statistical Society. Series A (General)* 82–117.

Clayton, D. G. (1978), A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**(1), 141–151.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

Cuzick, J. (1982), Rank tests for association with right censored data, *Biometrika* **69**(2), 351–364.

Dabrowska, D. M. (1986), Rank tests for independence for bivariate censored data, *The Annals of Statistics* **14**(1), 250–264.

Dabrowska, D. M. (1988), Kaplan–Meier estimate on the plane, *The Annals of Statistics* **16**(4), 1475–1489.

Dabrowska, D. M. (1989), Kaplan–Meier estimate on the plane: weak convergence, LIL, and the bootstrap, *Journal of Multivariate Analysis* **29**(2), 308–325.

DaÌĹtwyler, C. (2011), 'Parametric survival models'. [Online; accessed 28-April-2020].
URL: *https://stat.ethz.ch/education/semesters/ss2011/seminar/contents/handout_9.pdf*

Ding, A. A. and Wang, W. (2004), Testing independence for bivariate current status data, *Journal of the American Statistical Association* **99**(465), 145–155.

Dupont, C., Horner, J., Li, C., Liu, Q. and Shepherd, B. (2018), *PResiduals: Probability-Scale Residuals and Residual Correlations.* R package version 0.2-6.
URL: *https://CRAN.R-project.org/package=PResiduals*

Fan, J., Hsu, L. and Prentice, R. L. (2000), Dependence estimation over a finite bivariate failure time region, *Lifetime Data Analysis* **6**(4), 343–355.

Gill, R. D., Laan, M. J. v. d. and Wellner, J. A. (1995), Inefficient estimators of the bivariate survival function for three models, *Annales de l'I.H.P. Probabilités et statistiques* **31**(3), 545–597.
URL: *http://www.numdam.org/item/AIHPB_1995___31_3_545_0*

Harrell Jr, F. E. (2015), *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, Springer.

Kalbfleisch, J. D. and Prentice, R. L. (2011), *The statistical analysis of failure time data*, John Wiley & Sons.

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

Kruskal, W. H. (1958), Ordinal measures of association, *Journal of the American Statistical Association* **53**(284), 814–861.

Li, C. and Shepherd, B. E. (2012), A new residual for ordinal outcomes, *Biometrika* **99**(2), 473–480.

Lin, D. and Ying, Z. (1993), A simple nonparametric estimator of the bivariate survival function under univariate censoring, *Biometrika* **80**(3), 573–581.

Liu, Q., Li, C., Wanga, V. and Shepherd, B. E. (2018), Covariate-adjusted spearman's rank correlation with probability-scale residuals, *Biometrics* **74**(2), 595–605.

McGowan, C. C., Cahn, P., Gotuzzo, E., Padgett, D., Pape, J. W., Wolff, M., Schechter, M. and Masys, D. R. (2007), Cohort profile: Caribbean, central and south america network for hiv research (ccasanet) collaboration within the international epidemiologic databases to evaluate aids (iedea) programme, *International journal of epidemiology* **36**(5), 969–976.

Narasimhan, B. and Johnson, S. G. (2017), *cubature: Adaptive Multivariate Integration over Hypercubes.* R package version 1.3-8, https://CRAN.R-project.org/package=cubature (accessed July 9, 2019).
URL: *https://CRAN.R-project.org/package=cubature*

National Eye Institute (1981), Diabetic retinopathy study (DRS). ClinicalTrials.gov Identifier: NCT00000160.
URL: *https://clinicaltrials.gov/ct2/show/study/NCT00000160*

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Oakes, D. (1982), A model for association in bivariate survival data, *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(3), 414–422.

Oakes, D. (1989), Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**(406), 487–493.

Ploner, M., Kaider, A. and Heinze, G. (2013), *Correlation Analysis of Survival Times by Iterative Multiple Imputation.* R package version 1.0.

Ploner, M., Kaider, A. and Heinze, G. (2015), *SurvCorr: Correlation of Bivariate Survival Times.* R package version 1.0, https://CRAN.R-project.org/package=SurvCorr (accessed July 9, 2019).
URL: *https://CRAN.R-project.org/package=SurvCorr*

Prentice, R. L. and Cai, J. (1992), Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**(3), 495–512.

Prentice, R. L. and Hsu, L. (1997), Regression on hazard ratios and cross ratios in multivariate failure time analysis, *Biometrika* **84**(2), 349–363.

Prentice, R. L. and Zhao, S. (2019), *The Statistical Analysis of Multivariate Failure Time Data: A Marginal Modeling Approach*, Chapman and Hall/CRC.

Pruitt, R. C. (1991), On negative mass assigned by the bivariate Kaplan–Meier estimator, *The Annals of Statistics* **19**(1), 443–453.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Version 3.4.3., https://www.R-project.org/ (accessed July 9, 2019).
URL: *https://www.R-project.org/*

Romeo, J. S., Tanaka, N. I. and Pedroso-de Lima, A. C. (2006), Bivariate survival modeling: a Bayesian approach based on copulas, *Lifetime Data Analysis* **12**(2), 205–222.

Royston, P. and Parmar, M. K. (2013), Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome, *BMC Medical Research Methodology* **13**(1), 152.

Schemper, M., Kaider, A., Wakounig, S. and Heinze, G. (2013), Estimating the correlation of bivariate failure times under censoring, *Statistics in Medicine* **32**(27), 4781–4790.

Shepherd, B. E., Gilbert, P. B. and Lumley, T. (2007), Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization, *Journal of the American Statistical Association* **102**(478), 573–582.

Shepherd, B. E., Li, C. and Liu, Q. (2016), Probability-scale residuals for continuous, discrete, and censored data, *Canadian Journal of Statistics* **44**(4), 463–479.

Shih, J. H. and Louis, T. A. (1995), Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* **51**(4), 1384–1399.

Shih, J. H. and Louis, T. A. (1996), Tests of independence for bivariate survival data, *Biometrics* **52**(4), 1440–1449.

Stefanski, L. A. and Boos, D. D. (2002), The calculus of m-estimation, *The American Statistician* **56**(1), 29–38.

Stute, W. (1993), Consistent estimation under random censorship when covariables are present, *Journal of Multivariate Analysis* **45**(1), 89–103.

Stute, W. (1995), The central limit theorem under random censorship, *The Annals of Statistics* 422–439.

Therneau, T. M. (2015), *A Package for Survival Analysis in S.* version 2.38.
URL: *https://CRAN.R-project.org/package=survival*

van der Laan, M. (1997), Nonparametric estimators of the bivariate survival function under random censoring, *Statistica Neerlandica* **51**(2), 178–200.

van der Laan, M. J. (1996), Efficient estimation in the bivariate censoring model and repairing npmle, *The Annals of Statistics* **24**(2), 596–627.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Wikipedia contributors (2020), 'Leibniz integral rule — Wikipedia, the free encyclopedia'. [Online; accessed 28-April-2020].
URL: *https://en.wikipedia.org/wiki/Leibniz_integral_rule*

Zeng, D. and Lin, D. (2007), Maximum likelihood estimation in semiparametric regression models with censored data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(4), 507–564.

Zhang, D. (2020), 'Modeling survival data with parametric regression models'. [Online; accessed 28-April-2020].
URL: *http://www4.stat.ncsu.edu/ dzhang2/st745/chap5.pdf*

Zhang, S. (2008), *Inference on the association measure for bivariate survival data with hybrid censoring and applications to an HIV study*, ProQuest.