Simulation Study of an Estimator of Bivariate Survivor Function and its Variance Estimator

Yu Lan Jin

A Thesis
for
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science at
Concordia University
Montréal, Québec, Canada

March 2008

©Yu Lan Jin, 2008

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-40818-6
Our file Notre référence
ISBN: 978-0-494-40818-6

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

Simulation Study of an Estimator of Bivariate Survivor Function and Estimator of its Variance

by Yu Lan Jin

Bivariate survival data arises when we have either a pair of observation times for each individual or times on two related individuals, such as infection times for the two kidneys of a person or death times of twins. Such data are also often subject to censoring - bivariate censoring - i.e., exact observations may not be available on one or both of components because of drop-out or other reasons. Hence it is important to have an efficient, nonparametric bivariate survivor function estimator under censoring, i.e., a bivariate Kaplan-Meier estimator. In this thesis we carry out an extensive simulation study of an estimator proposed by Sen and Stute(2007), which involves solving for an eigenvector of a certain matrix. A comparison of the estimator with two other existing but unsatisfactory ones is also given using a small data-set. Moreover, variance of the former is computed using a bivariate analogue of Greenwood's formula, which involves solving a matrix equation of the form **AXB=C**

Acknowledgements

I would like to express my gratitude to all people who have helped and inspired me during my graduate study.

I am deeply indebted to my supervisor, Dr. Arusharka Sen, Associate Professor in the Department of Mathematics and Statistics at Concordia University, whose expert supervision, stimulating suggestion and numerous guidance helped me in all the time of graduate study and thesis work, and gave lots of valuable advise in selecting graduate courses.

I would like to express my gratitude to the Department of Mathematics and Statistics, Concordia University, for giving me the permission to my graduate study, and giving me financial support.

I also thank Dr. Y. P. Chaubey and Dr. W. Jiang serving on my thesis committee.

I owned thanks to my friend, Care Fung, who encouraged me to apply the graduate study, and kept motivating me to finish my thesis.

I would like to thank my friends, especially, Yong Ling and Zhou Wei, for their encouragement and suggestion.

I wish to thank my entire extended family for providing a loving environment and constant support.

My thank goes to my sister, Xing Hua Jin, my brother in law, Xiu Xian Wu, and my sister-in-law, Fan Yang, who gave me academic writing suggestion and encouragement.

I would like to give my loving thanks to my husband, Ming Yang, who financially and spiritually supported me. Without his support, this thesis would certainly not have existed.

Also I warmly thank my intelligent son, Chi Yu Yang, who allowed me to spend more of the time on my thesis. For the past six months, my son kept on reminding me: "Mom, finish your thesis first"!

I wish to extend my warmest thanks to my father, Shan Lu Jin, who always was proud of me, my mother, Cui Jing Ji, who is the role model to face the difficulties and gave me emotional support. Unfortunately, my father is no longer with us to celebrate the completion of my thesis.

Dedication

I would like to dedicate this thesis to my mother Mrs. Jing Ji Cui and my father Mr. Shan Lu Jin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 How did the problem arise?

For univariate survival data analysis, we have many efficient estimators and inference proceeding.

Correlated failure time data arise in diverse application areas such as disease occurrence studies between pairs of family members in genetic epidemiology. Such data are often subject to (bi-variate) censoring. For the purpose of inference without parametric models, or for

Model checking (for a parametric model), we need an efficient, computationally convenient nonparametric bivariate survivor function estimator. In other words, it is desirable to have an analogue of the Kaplan-Meier estimator for bivariate failure time data.

1.2 Notation

Some notations are involved in my thesis as below.

Survival data

Survival data is a term used for describing data that measure the time to some event. The event is a transition from one state to another.

For example, death is a transition from the state alive to the state dead. Occurrence of disease is a transition from a state of being healthy to a state of presence of disease. In the economic example, it is a transition from a state unemployed to a state employed to a state unemployed.

Univariate survival data

This term means that all time variables describing the time to the same type of event and individuals are assumed to be independent. The event considered will be called death for theoretical evaluations, event though it also can be other events. Therefore, the data consist of n independent times, $T_1, \ldots, T_n$, with corresponding death indicators $D_1, \ldots, D_n$. Thus, in the case of death,$D = 1$, T is time to death; in the case of censoring $D = 0$, and T is the observation time. A subscript i is used to denote the individuals.

Univariate censoring

The survival variables $Y_1, Y_2, Y_3, \ldots, Y_n$ are right-censored by fixed constants $t_1, t_2, t_3, \ldots, t_n$, if the observed sample consists of the ordered pairs (Z_i, δ_i), for $i = 1, 2, \ldots, n$, where for each $Z_i = min\{Y_i, t_i\}$,

$$\delta_i = \left[\begin{array}{lll} 1 & ifY_i \leq t_i & (uncensored) \\ 0 & ifY_i > t_i & (censored) \end{array} \right.$$

wheret_i is the fixed censor time and δ_i the censor indicator for Y_i.

Survival random variable

A random variable X is a survival random variable if an observed outcome x of X lies in the interval $[0, \infty)$.

Suppose that X has probability density function f and cumulative distribution function F. Then $F(x) = P(X \leq xy) = \int_0^y f(u)du$.

Survivor function

The survival function,$\bar{F}$, is defined for all values of x by $\bar{F}(x) = 1 - F(x)$.
i.e. $\mathbf{F}(x) = P(X > x) = \int_x^\infty f(u)du$

Empirical survivor function

Given n observations $X_1, X_2, X_3, \ldots, X_n$ independently and identically distributed (i.i.d.) with the same distribution as Y, the empirical survivor function $\bar{F}_n$ is defined for all values of y by

$$\bar{F}_n(x) = \frac{number of observation > x}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(x,\infty)}(X_i)$$

and is an estimate of the survival function $\bar{F}$.

Hazard function

A survival random variable Y has hazard function, or hazard rate or force of mortality, defined for $y > 0$ by

$$h(y) = \lim_{\Delta y \to 0} \frac{P(y < Y < y + \Delta y)}{\Delta y P(Y > y)} = \frac{f(y)}{S(y)}$$

Greenwood's Formula

In terms of notation for jth interval $I_j = [a_{j-1}, a_j)$, N_j as the number at risk in I_j; we write:

$$\begin{aligned} p_j &= P(Surviving\ through\ I_j|\ Alive\ at\ the\ start\ of\ I_j) \\ &= P(X > a_j | X > a_{j-1}) \end{aligned} \tag{1.1}$$

where $\tilde{p}_j$ as the actuarial estimate of p_j and $q_j = 1 - p_j$, $\tilde{\bar{F}}(a_j)$ is lifetable estimator.
We will concentrate on the derivation which approximates $\tilde{S}(a_j)$ by a product of independent binomial proportions for the intof the lifetable prior to a_j. This will require positive sample sizes (effective numbers at risk) in each of the intervals concerned. If we condition on this, then the result is exact rather than an approximation; unconditionally, the standard error result is an approximation.
We begin by noting that if $N'_j > 0$, the effective number at risk,N'_j, for $j > r$ depends on past effective numbers at risk, N'_l with $l < r$ only through the value of N'_r. Of course, if $N'_j = 0$, then we cannot tell at a previous interval identified by $l < j$ whether $N'_l = 0$ or whether $N'_l > 0$. The discussion begins by showing that the lifetable estimate $\tilde{\bar{F}}(a_j)$ is 'approximately' unbiased for $S(a_j)$.

Theorem1

$$E[\bar{\tilde{F}}(a_j)] \approx p_1 p_2 p_3 \cdots p_j = \bar{F}(a_j).j = 1, 2, 3, \ldots, k+1.$$

Theorem2 (Greenwood's Formula)

The standard error of the lifetable estimate if given by

$$Var[\bar{\tilde{F}}(a_j)] \approx \bar{F}(a_j)^2 \sum_{i=1}^{j} \frac{q_i}{p_i N_i'}, j = 1, 2, 3, \ldots, k+1.$$

The Kaplan-Meier Estimator

The most commonly used for survival data is the Kaplan- Meier (1958) product limit estimate . The Kaplan-Meier estimator is aimed at estimation of the survival function from censored life-time data. The value of the survival function between successive distinct uncensored observations is taken as constant, and the graph of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude.

If π_j is the probability of having an event until then, that is, on surviving to that time, the likelihood function is $L(\pi) = \prod_{j=1}^{k} \pi_j^{d_j}(1-\pi_j)^{n_j - d_j}$ where n_j is the number having survived and still under observation, and hence still known to be at risk just prior to t_j, called the risk set, d_j, is the number having the event at time t_j, and π_j is the hazard or intensity at t_j. This is a special application of binomial distribution, with maximum likelihood estimates,$\hat{\pi}_j = d_j/n_j$. Then, the product limit estimate of survivor function is just the product of the estimated probabilities of not having the event at all time points up to the one of interest:

$$\hat{\bar{F}}(t) = \prod_{j|t_j<t} \frac{n_j - d_j}{n_j} = \prod_{i|Z_i \leq x} \left[1 - \frac{\delta_i}{n\bar{H}_n(Z_i)}\right]$$

where $\bar{H}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i > x)$.

Bivariate survival times, Survival function, Hazard function

X means $X = (X_1, X_2) \in \mathbb{R}^{+2}$ and if we write $\leq, \geq, >$ and $<$, then this should hold componentwise: for example, if $x, y \in \mathbb{R}^{+2}$, then $x \leq y \iff x_1 \leq y_1, x_2 \leq y_2$. We will write

$X_i, i = 1, \ldots, n$, as notation for n i.i.d. bivariate survival times with the same distribution as T, while we write X_1 and X_2 for survival times with the same distribution as X, while we write X_1 and X_2 for the components of X.
Bivariate right randomly censored data can be modeled as follows: T is a positive bivariate lifetime vector with bivariate distribution F_0 and survival function $\bar{F}_0 : \bar{F}_0(t) = Pr(X \leq x)$ and $\bar{F}_0(t) = Pr(X > x)$. Let C be a positive bivariate censoring vector with bivariate distribution G_0 and survivor function bivariate censoring vector with bivariate distribution G_0 and survivor function $H_0 : G_0(Y) = Pr(Y \leq y)$ and $\bar{G}_0(Y) = Pr(Y > y)$.

Handling probabilities is more complicated in the bivariate case than in the univariate. In the univariate case, the probability of an interval (a, b), that is, $Pr(T(a, b))$, is found as $\bar{F}(b) - \bar{F}(a)$, but in bivariate case, the corresponding formula is

$$Pr(X_i(a_1, b_1), X_2(a_2, b_2)) = \bar{F}(b_1, b_2) - \bar{F}(a_1, b_2) - \bar{F}(b_1, a_2) + \bar{F}(a_1, a_2).$$

We may define a bivariate hazard function as

$$h(t_1, t_2) = \frac{f(t_1, t_2)}{\bar{F}(t_1, t_2)}$$

which describes the probability that both coordinates will experience an event given that they are both alive. This naturally extends the univariate expression $h(t) = \frac{f(t)}{\bar{F}(t)}$, which alternatively can be written as $h(t) = -\frac{dlog\bar{F}(t)}{dt}$. Thus $h(t_1, t_2) = \frac{\frac{d^2\bar{F}(t_1,t_2)}{dt_1 dt_2}}{\bar{F}(t_1,t_2)}$, but it cannot be simply formulated by means of the derivative of $log\bar{F}(t_1, t_2)$. In fact, the relation is

$$\frac{d^2}{dt_1, dt_2} log\bar{F}(t_1, t_2) = h(t_1, t_2) - \{\frac{d}{dt_1} log\bar{F}(t_1, t_2)\}\{\frac{d}{dt_2} log\bar{F}(t_1, t_2)\}.$$

Observations are described in the standard parallel way, as (X_1, X_2) with corresponding death indicators (Y_1, Y_2). There are three types of observations - the double deaths, that is, known times; single deaths, where one individual is observed to die and the other is censored; and double censoring.

Uncensored

If $D_i = (1, 1)$, then the observation Y_i is called uncensored.

Singly censored

If $D_i = (0, 1)$ or $D_i = (1, 0)$, then the observation Y_i is called singly censored.

Doubly censored

If $D_i = (0, 0)$, then the observation Y_i is called doubly censored.
The uncensored observations are the complete observations and singly censored and doubly censored are incomplete observations.

1.3 History of non-parametric bivariate survivor function estimator

For univariate survival function estimator, we have Kaplan-Meier eatimator and Nelson-Aalen estimator. Especially, Kaplan-Meier Estimator is successfully to express the masses of survival function. We use the graphs of the Kaplan-Meier estimator to compare different group of survival data.
It is a long history to find an efficient bivariate survival estimator. Many proposals for estimation of the bivariate survival function have been made in bivariate censored data . There are some main remarkable estimators.

Hanley and parnes (1983) estimator is a maximum likelihood estimate. They suggested this estimate and made an explicit evaluation under homogeneous censoring and described an iterative solution in the general case.
Their estimation method for homogeneous case has an interpretation like the multi-state model, because they split the problem into the distribution of minimum, the distribution of which component(s) fails at the minimum given the minimum, and then an arbitrary distribution for the second event given the first. When the censoring pattern is not homogeneous, this simple derivation is not possible. therefore, the two cases are treated separately. This method is limited to solve some cases.

Pruitt(1991) proposed an interesting implicitly defined estimator which is the solution of an ad hoc modification of the self-consistency equation. The Pruitt method instead distributes the mass according to a Kaplan-Meier method applied to the observed events in a neighborhood of the observation.

Dabrowska(1988) and Van der Laan (1995) found the notable estimators. Dabrowska's multivariate product-limit estimator, based on a very clever representation of a multivariate survival function in terms of its conditional multivariate hazard measure.
The Dabrowska method has the problem that it assigns negative probability masses to some points. As demonstrated in the example, this happens at a very large number of points and the mass is non-ignorable. A further problem that makes us insure about the approach is that it supplies an estimate for full bivariate distribution,even when this distribution does not make sense, e.g., for bivariate data for different events with one of events being death. Note that the former can assign negative values to some events whereas the latter is inexplicit, although asymptotically efficient under some strong conditions such as complete observation of the censoring variables.

The Dabrowska estimate

An interesting estimate of the bivariate survivor function was suggested by Dabrowska (1988). It was derived by a consideration of bivatiate hazard functions. The estimate is as follows. First find the bivariate risk set

$$R(t_1, t_2) = \sum_i \{T_{i1} \geq t_1, T_{i2} \geq t_2\}.$$

Then we need the number of bivariate events at each time

$$K_{11}(t_1, t_2) = \sum_i D_{i1} D_{i2} 1\{T_{i1} = t_1,\ T_{i2} = t_2\}$$

and the number of events for coordinate 1, among those where
the second component is alive at time t_2

$$K_{10}(t_1,t_2) = \sum D_{i1}1\{T_{i1} = t_1, T_{i2} \geq t_2\}$$

The quantities are seen relative to the risk set

$$\begin{aligned} L_{11}(t_1,t_2) &= K_{11}(t_1,t_2)/R(t_1,t_2) \\ L_{10}(t_1,t_2) &= K_{10}(t_1,t_2)/R(t_1,t_2) \\ L_{01}(t_1,t_2) &= K_{01}(t_1,t_2)/R(t_1,t_2) \end{aligned}$$

The marginal survivor functions are found as

$$\begin{aligned} S_1(t_1) &= \prod_{u\leq t_1}\{1 - L_{10}(u,0)\} \\ S_2(t_2) &= \prod_{u\leq t_2}\{1 - L_{01}(u,0)\}. \end{aligned}$$

In fact, they are just Kaplan-Meier estimates based on each coordinate separately. At all times without events, the factor is 1 and can be neglected. At times with event, there is a term below 1, which contributes to the estimate. Then the estimate is

$$S(t_1,t_2) = S_1(t_1)S_2(t_2) \prod_{0\leq u\ \leq t_1, 0\leq u\ \leq t_2} \{1 - H(u,v)\} \quad (2.1)$$

where H is given by

$$H(t_1,t_2) = \frac{L_{10}(t_1,t_2)L_{01}(t_1,t_2) - L_{11}(t_1,t_2)}{\{1 - L_{10}(t_1,t_2)\}\{1 - L_{01}(t_1,t_2)\}}$$

It can be seen that Equation(2.1) has a strong interpretation as the product of the marginal survivor functions, modified by the product of H terms, which then describe the dependence. If we want to assume symmetry, R should be substituted by $R(t_1,t_2)+R(t_2,t_1)$ and similarly K_{11} should be substituted by $K_{11}(t_1,t_2) + K_{11}(t_2,t_1)$. Furthermore, K_{01} should be substituted by $K_{01}(t_1,t_2) + K_{01}(t_2,t_1)$, and $K_{10}(t_1,t_2)$ should be the transpose of sum.

Prentice and Cai (1992) suggested an estimator based on representation of the survivor function by Peano series which is a nice estimator.

Prentice et al (2004) obtained one estimator of survival function with the empirical matrix eigenvector, but it has incorrect solution.

Sen and Stute (2007) derived a bi-variate (or, multivariate) survivor function estimator with a general solution to the empirical version of the eigenfunction equation by using a simple matrix eigenvector calculation. The estimator is linearized by the functional $\Delta-$method.

In brief, the Dabrowska method gives negative mass in some points. The Prentice et al (1992) method has the incorrect solution which is shown by Sen and Stute (2007).

Generally, expressions for the variance are not available. Variance estimate has been derived only for the Hanley and Parnes approach, using Greenwood's formula.

1.4 Content

The aim of my thesis is to carry out a simulation study of Sen. and Stute's(2007) estimator as well as the associated variance estimator formula.

In Chapter 2, computation and simulation of the estimator of Sen and Stute (2007) under different survival joint distributions and censored joint distributions, are given. We also checked the estimator with the real data (twins, kidney), and compared with Dabrowska's and Hanley and Parnes's methods.

Chapter 3 gives the estimator of variance of bivariate survival function and the simulation results.

Chapter 4 shows the conclusion of the simulation study.

Further study is in Chapter 5.

Chapter 2

Calculation and Simulation of the Estimator of Sen and Stute(2007)

2.1 The estimator of Sen and Stute (2007)

Simulation studies are presented to assess the moderate sample performance of a bi-variate Kaplan-Meier estimator, denoted $\bar{F}_e$, derived by Sen and Stute. We present the mean squared-error (MSE) of $\bar{F}_e$ under different degrees censoring, with failure times and random censoring times generated from several joint distributions $F(x_1, x_2)$ and $G(y_1, y_2)$ respectively. A comparison with Dabrowska and Hanley-Parnes estimators are also provided in a small, real-life data-set.

The bivariate Kaplan-Meier estimator derived by Sen and Stute (2007)
Let (X_{i1}, X_{i2}), $1 \leq i \leq n$, be independent and identically distributed (i.i.d) nonnegative random vectors, each having a bi-variate distribution function (d.f.) $F(x_1, x_2)$ and representing a bi-variate failure or survival time, such as those for 'twins', or pairs of kidney. Suppose further that these vectors are subject to random censoring from the right by another, independent set of i.i.d random vectors $(Y_{i1}; Y_{i2}), 1 \leq i \leq n$, each having d.f. $G(y_1, y_2)$, so that we can only observe $(\delta_{i1}; \delta_{i2}; Z_{i1}; Z_{i2}); 1 \leq i \leq n$; where $\delta_{ij} = I\{X_{ij} \leq Y_{ij}, Z_{ij} = \min(X_{ij}; Y_{ij}); j = 1, 2; 1 \leq i \leq n$.
Let $\bar{F}(x_1, \ldots, x_m) = P\{X_1 > x_1, \ldots, X_m > x_m\}$ be the survivor function of an m-

dimensional random vector $X = (X_1, \ldots, X_m),\ m \geq 1$. Then $\bar{F}(\cdot, \ldots, \cdot)$ satisfies the integral equation

$$\bar{F}(x_1-, \ldots, x_m-) = \int_{[x_1,\infty)\times\cdots\times[x_m,\infty)} \bar{F}(t_1-, \ldots, t_m-) \frac{dF(t_1, \ldots, t_m)}{\bar{F}(t_1-, \ldots, t_m-)} \tag{2.1}$$

Let us look at $m = 2$ only. Now for censored data, we have

$$\frac{dF(t_1, t_2)}{\bar{F}(t_1-, t_2-)} = \frac{\bar{G}(t_1-, t_2-)dF(t_1, t_2)}{\bar{G}(t_1-, t_2-)\bar{F}(t_1-, t_2-)},$$

where $G(\cdot, \cdot)$ is the censoring distribution. Thus Eq.(2.1) becomes

$$\bar{F}(x_1-, x_2-) = \int_{[x_1,\infty)\times[x_2,\infty)} \bar{F}(t_1-, t_2-) \frac{dH^{11}(t_1, t_2)}{\bar{H}(t_1-, t_2-)}, \tag{2.2}$$

and $F(\cdot, \cdot)$ can be estimated as a *solution* to the empirical version of Eq.(2.2):

$$\bar{F}_n(x_1-, x_2-) = \int_{[x_1,\infty)\times[x_2,\infty)} \bar{F}_n(t_1-, t_2-) \frac{dH_n^{11}(t_1, t_2)}{\bar{H}_n(t_1-, t_2-)}, \tag{2.3}$$

where as usual, $H_n^{11}(t_1, t_2) = n^{-1}\sum_{i=1}^n \delta_{i1}\delta_{i2}\mathbf{1}\{Z_{i1} \leq t_1, Z_{i2} \leq t_2\}$, and $\bar{H}_n(t_1, t_2) = n^{-1}\sum_{i=1}^n \mathbf{1}\{Z_{i1} > t_1, Z_{i2} > t_2\}$.

Equations (2.1) and (2.3) obviously represent *eigenvalue* problems, i.e., $\bar{F}(x_1-, x_2-)$ and $\bar{F}_n(x_1-, x_2-)$ are *eigenvectors* corresponding to the eigenvalue 1 for the integral operators $\int_{[\cdot,\infty)\times[\cdot,\infty)}(dF(t_1, t_2)/\bar{F}(t_1-, t_2-))$ and $\int_{[\cdot,\infty)\times[\cdot,\infty)}(dH_n^{11}(t_1, t_2)/\bar{H}_n(t_1-, t_2-))$, respectively.

To solve Eq.(2.3), we may assume that the estimator gives mass $p_i \geq 0$ to the observation $(Z_{i1}, Z_{i2}),\ 1 \leq i \leq n$, so that

$$\bar{F}_i := \bar{F}_n(Z_{i1}-, Z_{i2}-) = \sum_{j=1}^n a_{ij}p_j,$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } Z_{j1} \geq Z_{i1},\ Z_{j2} \geq Z_{i2} \\ 0 & \text{otherwise;} \end{cases}$$

Further, let $b_i := \triangle H_n^{11}(Z_{i1}, Z_{i2})/\bar{H}_n(Z_{i1}-, Z_{i2}-) = n^{-1}\delta_{i1}\delta_{i2}/\bar{H}_n(Z_{i1}-, Z_{i2}-)$. Then Eq.(2.3), with $x_1 = Z_{i1},\ x_2 = Z_{i2},\ 1 \leq i \leq n$, may be rewritten in matrix notation as,

$$\mathbf{Ap}{=}\mathbf{ABAp},\ \sum_{i=1}^n p_i = 1, \tag{2.4}$$

where $\mathbf{A} = ((a_{ij})),\ \mathbf{p} = (p_1, \ldots, p_n),\ \mathbf{B} = \text{diag}\ (b_1, \ldots, b_n)$. Now order $(Z_{i1}, Z_{i2}),\ 1 \leq i \leq n$, in the increasing order of the first coordinate, i.e., as $(Z_{[i:n1]}, Z_{[i:n2]}),\ 1 \leq i \leq n$, where

$Z_{1:n1} \leq \cdots \leq Z_{n:n1}$ and $Z_{[i:n2]}$, $1 \leq i \leq n$, are the corresponding concomitant. Then, with any suitable convention for breaking ties, $\mathbf{A}$ becomes a non-singular, *upper-triangular* matrix, i.e.,

$$a_{ij} = \begin{cases} 0 & \text{if } j < i \\ 1 & \text{if } j = i \\ 1 \text{ or } 0 & \text{if } j > i \end{cases}$$

Note that for univariate ordered data, $a_{ij} = 1$ for all $j \geq i$. Thus $\mathbf{A}$ now becomes invertible, and Eq.(2.4) becomes

$$\mathbf{p}{=}\mathbf{BAp},\ \sum_{i=1}^{n} p_i = 1. \tag{2.5}$$

Finally, we have $\bar{F}_{ei} = \sum_{j=1}^{n} a_{ij} p_j$ or in matrix notation

$$\bar{\mathbf{F}}_e = \mathbf{Ap} \tag{2.6}$$

The results of the estimation $F(x_1, x_2)$ or equivalently, its survivor function $\bar{F}(x_1, x_2) = P\{X_1 > x_1, X_2 > x_2\}$ based on the observed data, i.e., the bi-variate version of the Kaplan-Meier estimator, is as follows.

1) Observation bivariate data $(\mathbf{X}_1, \mathbf{X}_2)$ has the density of $f(x_1, x_2)$.

To estimate $\bar{F}(x_1, x_2) = P\{X_1 > x_1, X_2 > x_2\}$ based on a sample $(X_{i1}, X_{i2}), i = 1, 2, \ldots, n$ we use

$$\hat{\bar{F}}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} I\{X_{i1} > x_1, X_{i2} > x_2\}.$$

2) Censoring data $(\mathbf{Y}_1, \mathbf{Y}_2)$ has the density $g(y_1, y_2)$, where $(Y_1, Y_2), (X_1, X_2)$ are independent.

3) For $(Y_{i1}, Y_{i2}),\ i = 1, 2, \ldots, n$, data matrix : $(\delta_{i1}, \delta_{i2}, z_{i1}, z_{i2}),\ i = 1, 2, \ldots, n$, where

$$\mathbf{z}_{i1} = X_{i1} \wedge Y_{i1}, \qquad \delta_{i1} = I(X_{i1} \leq Y_{i1})$$

$$\mathbf{z}_{i2} = X_{i2} \wedge Y_{i2}, \qquad \delta_{i2} = I(X_{i2} \leq Y_{i2})$$

Let

$$X_1 = (X_{i1},\ 1 \leq i \leq n), \quad X_2 = (X_{i2},\ 1 \leq i \leq n),$$

$$Y_1 = (Y_{i1},\ 1 \leq i \leq n), \quad Y_2 = (Y_{i2},\ 1 \leq i \leq n),$$

$$\delta_1 = (X_1 \le Y_1) + 0, \qquad \delta_2 = (X_2 \le Y_2) + 0.$$

Arrange matrix $(\delta_{i1}, \delta_{i2}, z_{i1}, z_{i2}, i = 1, 2, \ldots, n)$ according to increasing order of $(z_{i1}, 1 \le i \le n)$, then the matrix change to $(\delta_{[i1]}, \delta_{[i2]}, z_{[i1]}, z_{[i2]}, i = 1, 2, \ldots, n)$.

We define $\mathbf{A} = ((a_{ij}))$, where $a_{ij} = \begin{cases} 1 & \text{if } Z_j 1 \ge Z_{i1},\ Z_{j2} \ge Z_{i2} \\ 0 & \text{otherwise;} \end{cases}$ $,1 \le i \le n,\ 1 \le j \le n.$
$\mathbf{B} = \text{diag }(b_1, \ldots, b_n)$, where $b_i = \frac{\delta_{i1}\delta_{i2}}{\sum_{j=1}^n a_{ij}}$
We may assume that the estimator gives $p_i \ge 0$ to the observation $(Z_{i1}, Z_{i2}),\ \ 1 \le i \le n$, so that $\bar{F}_n(Z_{i1}-, Z_{i2}-) = \sum_{j=1}^n a_{ij} p_j$,where $\mathbf{p} = (p_1, \ldots, p_n)$
So we solve the following *eigenvector* problem in $\mathbf{p}$:

$$\begin{cases} \mathbf{BAp{=}p} \\ \sum_{i=1}^n p_i = 1 \end{cases} \qquad (1.1)$$

Rewrite (1.1) as:

$$\left[\begin{array}{l} (\mathbf{I} - \mathbf{BA})\mathbf{p} = \mathbf{0} \\ \mathbf{1}^T \mathbf{p} = 1, \end{array} \right. \qquad \textit{where } \mathbf{1}^T = (1, 1, \ldots, 1)$$

so, matrix equation as below:

$$\begin{bmatrix} \mathbf{I} - \mathbf{BA} \\ \mathbf{1}^T \end{bmatrix}_{(n+1)\times n} \mathbf{p}_{n\times 1} = \begin{bmatrix} \mathbf{0}_{n\times 1} \\ \mathbf{1}_{1\times 1} \end{bmatrix}. \qquad (1.2)$$

Case(1.1): Unique solution if $b_i = 1$ for only one i, $b_i < 1$ for all other i.
In this case 1 is an eigenvalue of $\mathbf{BA}$ of multiplicity one. Hence the matrix equation Eq.(1.2) gives a unique solution $\mathbf{p}$.

Case (1.2): Multiple solution if $b_i = 1$ for more than one i, i.e. $b_{i_1} = \cdots = b_{i_k} = 1$.
In this case 1 is an eigenvalue of $\mathbf{BA}$ of multiplicity $k > 1$. Hence the matrix equation Eq.(1.2) gives k linearly independent solutions. We enforce a unique solution by letting $p_{i_1} = \cdots = p_{i_k}$ in the matrix equation (1.2), then solve out $\mathbf{p}$.

Case(1.3): No solution if $b_i < 1$ for every i.
In this case 1 is not an eigenvalue of $\mathbf{BA}$. However, we obtain a pseudo-solution as follows. Add a dummy variable p_{n+1}, with $b_{n+1} = 1$. ignore $\hat{p}_{n+1}$, $\sum_{i=1}^n p_i < 1$.

i.e.

$$\begin{cases} a_{i(n+1)} = 1, & 1 \le i \le n \\ a_{(n+1)j} = 0, & 1 \le j \le n \\ a_{(n+1)(n+1)} = 1 \end{cases}$$

Then, we change the matrix equation (1.2) to :

We switch to

$$\mathbf{A}' = \begin{bmatrix} (a_{ij}) & \mathbf{1}_{i(n+1)} \\ \mathbf{0}_{(n+1)j} & \mathbf{1}_{(n+1)(n+1)} \end{bmatrix}.$$

$$\mathbf{B}' = \text{diag } (b_1, \ldots, b_n, 1_{(n+1)})$$

, where $b_i = \frac{\delta_{1i}\delta_{2i}}{\sum_{j=1}^{n} a_{ij}}, 1 \le i \le n$

$$\mathbf{p}' = (p_1, \ldots, p_n, p_{n+1})$$

$$\begin{bmatrix} \mathbf{I} - \mathbf{B}'\mathbf{A}' \\ \mathbf{1}^T \end{bmatrix}_{(n+2)\times(n+1)} \mathbf{p}'_{(n+1)\times 1} = \begin{bmatrix} \mathbf{0}_{(n+1)\times 1} \\ \mathbf{1}_{1\times 1} \end{bmatrix}. \quad (1.2)'$$

We solve the adjusted matrix equation to get the solution of $\mathbf{p}'$.

Based on the above three cases, we have $\mathbf{p}$. Then we can calculate $\bar{F}_e(x_1, x_2) = \sum_{i=1}^{n} p_i I(Z_{1i} \ge x_1, Z_{2i} \ge x_2)$, where $p_1 + p_2 + \cdots + p_n = 1$

2) Mean squared error $\mathbf{MSE} = E(\bar{F}_e - \bar{F})^2$

For N repetitions, $MSE = \frac{1}{N}\sum_{i=1}^{N}(\bar{F}_e^i - \bar{F})^2$, where $\bar{F}(a_1, a_2) = P(X_1 > a_1, X_2 > a2)$

Two method to calculate $\bar{F}(a_1, a_2) = P(X_1 > a_1, X_2 > a2)$:

(2.1) **Exact method of calculation survivor function**:

$$\bar{F}(a_1, a_2) = P(X_1 > a_1, x_2 > a_2) = \int_{a_1}^{\infty}\int_{a_2}^{\infty}(f(x_1, x_2)dx_1dx_2$$

or

(2.1′) **Approximate method of calculation survivor function using the empirical**

method:

$$\tilde{F}(a_1, a_2) = \frac{1}{n}\sum_{i=1}^{n} I(X_{i1} > a_1, X_{i2} > a_2)$$

2.2 Simulation results from the following distributions

Simulation (3.1) Let observation (X_1, X_2) has the distribution $f(x_1, x_2)$, and censoring data has the distribution $g(y_1, y_2)$.

$$f(x_1, x_2) = \begin{cases} 6(1-x_2) & 0 \le x_1 \le x_2 \le 1 \\ 0 & elsewhere \end{cases} \qquad (3-1-1)$$

$$g(y_1, y_2) = \begin{cases} 1 & 0 \le y_1 \le 1, 2y_2 \le y_1 \\ 0 & elsewhere \end{cases} \qquad (3-1-2)$$

From (3-1-1),

$$f(x_1) = \int_{x_1}^{1} 6(1-x_2)dx_2 = \begin{cases} 3(1-x_1)^2 & 0 \le x_1 \le x_2 \le 1 \\ 0 & elsewhere \end{cases}$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)} = \frac{6(1-x_2)}{3(1-x_1)^2} = \begin{cases} \frac{2(1-x_2)}{(1-x_1)^2} & 0 \le x_1 \le x_2 \le 1 \\ 0 & elsewhere \end{cases}$$

Step 1) Generate pairs of (U_1, U_2): U_1 and U_2 are i.i.d. and uniform(0,1).
Step 2) Generate X_1:

$$F(X_1) = \int_0^{X_1} 3(1-t)^2 dt = 1 - (1-X_1)^3$$

Let $F(X_1) = U_1$ so,

$$X_1 = 1 - [1 - U_1]^{1/3}$$

Sept 3) Generate X_2:

$$F(X_2|X_1) = \begin{cases} \int_{X_1}^{X_2} \frac{2(1-t)}{(1-X_1)^2} dt & 0 \le x_1 \le x_2 \le 1 \\ 0 & elsewhere \end{cases} = 1 - \frac{(X_2-1)^2}{(X_1-1)^2}, \quad 0 \le X_1 \le 1$$

Let $F(X_2|X_1) = U_2$, so

$$U_2 = 1 - \frac{(X_2-1)^2}{(X_1-1)^2}$$

$$X_2 = 1 - (1-U_2)^{1/2}(1-X_1)$$

From (3-1-2),

$$g(y_1) = \int_0^{1/2y_1} dy_2 = \begin{cases} 1/2y_1 & 0 \le y_1 \le 2 \\ 0 & elsewhere \end{cases}$$

$$G(y_1) = \int_0^{y_1} 1/2ydy = 1/4y_1^2$$

Let $U_3 = G(y_1)$, so $Y_1 = 2U_3^{1/2}$

$$g(y_2|y_1) = \frac{g(y_1, y_2)}{g(y_1)} = \begin{cases} \frac{2}{y_1} & 0 \le y_1 \le 1, 2y_2 \le y_1 \\ 0 & elsewhere \end{cases}$$

$$G(y_2|y_1) = \int_0^{y_2} \frac{2}{y_1} dy$$

Let $U_4 = G(y_2|y_1)$, so $Y_2 = \frac{1}{2}U_4 Y_1$

Simulation result(3.1): n is sample size, N is repetition times. Test Results(N=200 samples, each of size n=100)

Table 2.1: Estimation results (3.1): $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\bar{F}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.395	0.6358	0.0436	0.307	0.6686	0	0.0248
(0.1,0.3)	0.395	0.6358	0.0436	0.307	0.6686	0	0.0248
(0.1,0.2)	0.5250	0.7001	0.0314	0.2993	0.69695	0	0.0238
(0.2,0.1)	0.3315	0.5101	0.0614	0.2949	0.67925	0	0.0259
(0.3,0.1)	0.1652	0.3461	0.1393	0.2962	0.67885	0	0.0250
(0.5,0.2)	0.0265	0.1277	0.2373	0.2978	0.67675	0	0.0255

Simulation (3.2):

Observation data has distribution as

$$f(x_1, x_2) = \begin{cases} 8x_1x_2 & 0 < x_1 < x_2 < 1 \\ 0 & elsewhere \end{cases} \qquad (3-2-1)$$

Censoring data distribuion is

$$g(y_1, y_2) = \begin{cases} 3y_1 & 0 < y_2 < y_1 < 1 \\ 0 & elsewhere \end{cases} \qquad (3-2-2)$$

From (3-2-1),

$$f(x_1) = \begin{cases} 4x_1(1-x_1^2) & 0 < x_1 < 1 \\ 0 & elsewhere \end{cases}$$

$$U_1 = F(x_1) = 2x_1^2 - x_1^4$$
$$X_1 = \sqrt{1 - \sqrt{1 - U_1}}$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)} = \begin{cases} \frac{2x_2}{1-x_1^2} & 0 < x_1 < x_2 < 1 \\ 0 & elsewhere \end{cases}$$
$$F(x_2|x_1) = \frac{1}{1-x_1^2}(x_2^2 - x_1^2) = U_2$$

so,

$$X_2 = (X_1^2 + U_2(1 - X_1^2))^{1/2}$$

From (3-2-2),

$$g(y_1) = \int_0^{y_1} 3y_1 dy_2 = \begin{cases} 3y_1^2 & 0 < y_1 < 1 \\ 0 & elsewhere \end{cases}$$
$$G(y_1) = \int_0^{y_1} 3t^2 dt = y_1^3 = U_1$$
$$Y_1 = U_1^{1/3}$$

$$g(y_2|y_1) = \frac{g(y_1, y_2)}{g(y_1)} = \begin{cases} y_1^{-1} & 0 < xy_2 < y_1 < 1 \\ 0 & elsewhere \end{cases}$$
$$G(y_2|y_1) = \int_0^{y_2} = \begin{cases} \frac{y_2}{y_1} & 0 < y_2 < y_1 < 1 \\ 0 & elsewhere \end{cases}$$

So $U_2 = \frac{y_2}{y_1}$,

$$Y_2 = Y_1 U_2 = U_1^{1/3} U_2$$
$$(Y_1, Y_2) = (U_3^{1/3}, U_3^{1/3} U_4)$$

Simulation result(3.2):(N=200 samples, each of size n=100)

Table 2.2: Estimation results (3.2) : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\bar{F}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.97375	0.5554	0.219482	0.08365	0.68235	0	0.234
(0.1,0.2)	0.97805	0.5889765	0.088067	0.085	0.68555	0	0.22945
(0.2,0.1)	0.9235	0.4935362	0.086302	0.08235	0.6868	0	0.23085
(0.3,0.1)	0.82785	0.4054272	0.092361	0.0843	0.69	0	0.2257
(0.5,0.2)	0.56075	0.1837466	0.148653	0.08695	0.68145	0	0.2316
(0.5,0.2)	0.5681	0.1650683	0.161627	0.08315	0.6849	0	0.23195

Simulation (3.3)

(X_1, X_2) is survival data. (Y_1, Y_2) is censoring data.

COPULA MODEL: $F(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}$

CLAYTON'S COPULA MODEL: For X:

$$\begin{aligned}\bar{F}(x_1, x_2) &= P\{X_1 > x_1, X_2 > x_2\} \\ &= \frac{1}{\left[\frac{1}{[\bar{F}_1(x_1)]^\theta} + \frac{1}{[\bar{F}_2(x_2)]^\theta} - 1\right]^{\frac{1}{\theta}}}\end{aligned} \tag{2.1}$$

$\theta > 0$, $\bar{F}_1, \bar{F}_2$ are survival marginal function

(3-3-4): Take $\theta = 4$,

$$\bar{F}_1(x_1) = e^{-x_1},\ \bar{F}_2(x_2) = e^{-x_2}$$

$$F_1(x_1) = \int_0^{x_1} e^{-t} dt = 1 - e^{-X_1}$$

Let $F_1(x_1) = U_1$, then $X_1 = -\ln(1 - U_1)$

$$\bar{F}(x_1, x_2) = \frac{1}{[e^{4x_1} + e^{4x_2} - 1]^{1/4}}$$

$$f(x_1, x_2) = \frac{d}{dx_2 dx_1}[\bar{F}(x_1, x_2)] = \frac{5e^{4x_1}e^{4x_2}}{[e^{4x_1} + e^{4x_2} - 1]^{9/4}}$$

$$f_1(x_1) = -e^{-x_1}$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{5e^{5x_1}e^{4x_2}}{[e^{4x_1} + e^{4x_2} - 1]^{9/4}}$$

$$\begin{aligned} F(x_2|x_1) &= \int_0^{x_2} \frac{5e^{5x_1}e^{4t}}{[e^{4x_1} + e^{4t} - 1]^{9/4}} dt \\ &= 1 - e^{5X_1}[e^{4X_1} + e^{4X_2} - 1]^{-5/4} \end{aligned} \tag{2.2}$$

Let $F(x_2|x_1) = U_2$,

Then

$$\begin{aligned} U_2 &= 1 - e^{5X_1}[e^{4X_1} + e^{4X_2} - 1]^{-5/4} \\ X_2 &= \frac{1}{4}\ln[1 - e^{4X_1} + [(1 - U_2)e^{-5X_1}]^{-4/5}] \end{aligned}$$

Randomly generate: U_1 and U_2 is uniform distribution (0,1)

$$\bar{F}_1(x_1) = e^{-x_1}, \qquad \bar{F}_2(x_2) = e^{-x_2}$$

$$\Longrightarrow X_1 = -\ln(1 - U_1)$$

$$X_2 = \frac{1}{4}\ln[1 - e^{4X_1} + [(1 - U_2)e^{-5X_1}]^{-\frac{4}{5}}]$$

(3-3-4-a-i) $Y : Y_1 \sim EXP(200),\ Y_2 = \infty$

Table 2.3: Estimation results 3-3-4-a-i : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000, 0.000)	0.9996	1.0000	0.0013	0.9954	0	0.0046	0
(0.375, 0.000)	0.6843	0.6859	0.0003	0.9950	0	0.0050	0
(0.750, 0.000)	0.4692	0.4712	0.0001	0.9953	0	0.0047	0
(1.125, 0.000)	0.3216	0.3236	0.0000	0.9949	0	0.0051	0
(1.500, 0.000)	0.2209	0.2227	0.0000	0.9948	0	0.0052	0
(0.000, 0.375)	0.6833	0.6844	0.0003	0.9949	0	0.0051	0
(0.375, 0.375)	0.5917	0.5933	0.0001	0.9952	0	0.0048	0
(1.500, 0.375)	0.2215	0.2232	0.0000	0.9950	0	0.0050	0
(0.375, 0.750)	0.4518	0.4538	0.0000	0.9947	0	0.0053	0
(0.750, 0.750)	0.3934	0.3953	0.0000	0.9951	0	0.0049	0
(1.500, 0.750)	0.2181	0.2199	0.0000	0.9948	0	0.0052	0
(0.000, 1.125)	0.3224	0.3242	0.0000	0.9947	0	0.0053	0
(0.375, 1.125)	0.3198	0.3216	0.0000	0.9948	0	0.0052	0
(0.750, 1.125)	0.3074	0.3092	0.0000	0.9947	0	0.0053	0
(1.125, 1.125)	0.2693	0.2711	0.0000	0.9952	0	0.0048	0
(1.500, 1.125)	0.2117	0.2134	0.0000	0.9953	0	0.0047	0
(0.000, 1.500)	0.2252	0.2270	0.0000	0.9953	0	0.0047	0
(0.375, 1.500)	0.2200	0.2216	0.0000	0.9952	0	0.0048	0
(0.750, 1.500)	0.2215	0.2230	0.0000	0.9953	0	0.0047	0
(1.125, 1.500)	0.2098	0.2115	0.0000	0.9949	0	0.0051	0
(1.500, 1.500)	0.1853	0.1870	0.0001	0.9949	0	0.0051	0

(3-3-4-a-ii): $Y:\ Y_1 \sim EXP(200),\ Y_2 = Y_1$

Table 2.4: Estimation results 3-3-4-a-ii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000,0.000)	0.9997	1.0000	0.0001	0.9944	0.0007	0.0006	0.0043
(0.375,0.000)	0.6864	0.6880	0.0000	0.9944	0.0006	0.0007	0.0043
(0.750,0.000)	0.4684	0.4703	0.0000	0.9941	0.0006	0.0006	0.0047
(1.125,0.000)	0.3251	0.3269	0.0000	0.9948	0.0005	0.0007	0.0040
(1.500,0.000)	0.2224	0.2241	0.0000	0.9949	0.0000	0.0006	0.0039
(0.000,0.375)	0.6864	0.6880	0.0000	0.9941	0.0007	0.0007	0.0046
(0.375,0.375)	0.5946	0.5963	0.0000	0.9943	0.0007	0.0008	0.0043
(0.750,0.375)	0.4508	0.4526	0.0000	0.9943	0.0008	0.0007	0.0042
(1.125,0.375)	0.3178	0.3197	0.0000	0.9942	0.0007	0.0006	0.0045
(1.500,0.375)	0.2212	0.2231	0.0000	0.9940	0.0008	0.0007	0.0046
(0.000,0.750)	0.4698	0.4715	0.0000	0.9942	0.0007	0.0008	0.0043
(0.375,0.750)	0.4486	0.4505	0.0000	0.9946	0.0007	0.0005	0.0042
(0.750,0.750)	0.3978	0.3996	0.0000	0.9946	0.0008	0.0007	0.0046
(1.125,0.750)	0.3078	0.3096	0.0000	0.9940	0.0007	0.0008	0.0042
(1.500,0.750)	0.2195	0.2213	0.0000	0.9941	0.0005	0.0007	0.0048
(0.000,1.125)	0.3229	0.3247	0.0000	0.9944	0.0005	0.0007	0.0045
(0.375,1.125)	0.3227	0.3248	0.0000	0.9944	0.0008	0.0006	0.0042
(0.750,1.125)	0.3098	0.3120	0.0000	0.9936	0.0008	0.0007	0.0049
(1.125,1.125)	0.2715	0.2734	0.0000	0.9946	0.0006	0.0006	0.0041
(1.500,1.125)	0.2102	0.2119	0.0000	0.9940	0.0009	0.0007	0.0044
(0.000,1.500)	0.2221	0.2238	0.0000	0.9939	0.0006	0.0008	0.0047
(0.375,1.500)	0.2229	0.2248	0.0000	0.9940	0.0008	0.0006	0.0045
(0.750,1.500)	0.2194	0.2211	0.0000	0.9947	0.0006	0.0006	0.0041
(1.125,1.500)	0.2119	0.2138	0.0000	0.9943	0.0008	0.0007	0.0042
(1.500,1.500)	0.1871	0.1887	0.0000	0.9945	0.0007	0.0007	0.0041

(3-3-4-a-iii): $Y:\ Y_1 \sim EXP(200),\ Y_1,\ Y_2$ are i.i.d.

Table 2.5: Estimation results 3-3-4-a-iii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$MSE\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000, 0.000)	0.9997784	1.0000	1.306597e-03	0.9905	0.0047	0.0048	0.000
(0.375, 0.000)	0.6859239	0.6883	3.249541e-04	0.9905	0.0047	0.0048	0.000
(0.750, 0.000)	0.4715727	0.4753	3.193285e-05	0.9895	0.0055	0.0050	0.000
(1.125, 0.000)	0.3212743	0.3251	9.160115e-06	0.9896	0.0055	0.0048	0.000
(1.500, 0.000)	0.2181039	0.2213	8.069634e-05	0.9901	0.0051	0.0047	0.000
(0.000, 0.375)	0.6803728	0.6831	3.135022e-04	0.9900	0.0051	0.0049	0.000
(0.375, 0.375)	0.5937038	0.5968	1.613447e-04	0.9900	0.0049	0.0051	0.000
(0.750, 0.375)	0.4506014	0.4541	1.971480e-05	0.9902	0.0046	0.0051	0.000
(1.125, 0.375)	0.3160567	0.3196	1.107430e-05	0.9898	0.0055	0.0047	0.000
(1.500, 0.375)	0.2200086	0.2235	7.873276e-05	0.9899	0.0048	0.0058	0.000
(0.000, 0.750)	0.4671715	0.4707	2.912554e-05	0.9901	0.0048	0.0051	0.000
(0.375, 0.750)	0.4501753	0.4538	1.949694e-05	0.9901	0.0047	0.0052	0.000
(0.750, 0.750)	0.3996358	0.4030	2.242911e-06	0.9902	0.0052	0.0045	0.000
(1.125, 0.750)	0.3079297	0.3113	1.441737e-05	0.9903	0.0049	0.0048	0.000
(1.500, 0.750)	0.2146828	0.2180	8.428403e-05	0.9900	0.0049	0.0050	0.001
(0.000, 1.125)	0.3178020	0.3216	1.041381e-05	0.9898	0.0049	0.0052	0.001
(0.375, 1.125)	0.3190518	0.3229	9.953295e-06	0.9902	0.0049	0.0049	0.000
(0.750, 1.125)	0.3002254	0.3037	1.799310e-05	0.9904	0.0049	0.0047	0.000
(1.125, 1.125)	0.2707833	0.2743	3.530341e-05	0.9901	0.0050	0.0048	0.000
(1.500, 1.125)	0.2099726	0.2130	8.935121e-05	0.9900	0.0048	0.0051	0.000
(0.000, 1.500)	0.2195676	0.2229	7.918518e-05	0.9902	0.0050	0.0047	0.000
(0.375, 1.500)	0.2185869	0.2218	8.019608e-05	0.9899	0.0051	0.0050	0.000
(0.750, 1.500)	0.2156402	0.2190	8.327215e-05	0.9900	0.0048	0.0052	0.000
(1.125, 1.500)	0.2084467	0.2119	9.102444e-05	0.9900	0.0052	0.0048	0.000
(1.500, 1.500)	0.1842161	0.1875	1.196754e-04	0.9897	0.0053	0.0050	0.001

Test Result 3-3-4-b:

The distribution of (X_1, X_2: COPULA MODEL $theta = 4$

(3-3-4-b-i) The distribution: $g(y_1, y_2)$

$Y_1 = 2exp(-2y_1)$

$Y_2 = \infty$

Test Results(N=200 samples, each of size n=100)

Table 2.6: Estimation results 3-3-4-b-i : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,02)	0.7794	0.5209	0.0293	0.3273	0	0.6727	0
(0.1,03)	0.7187	0.4225	0.0366	0.3308	0	0.6693	0
(0.2,0.1)	0.7803	0.4834	0.0325	0.3364	0	0.6637	0
(0.3,0.1)	0.7102	0.3544	0.0480	0.3394	0	0.6606	0
(0.5,0.2)	0.5794	0.1899	0.1192	0.3371	0	0.6629	0

(3-3-4-b-ii):

The distribution of X1, X2: COPULA MODEL $theta = 4$

The distribution: $g(y_1, y_2)$: $Y_1 = 2exp(-2y_1)$, $Y_1 = Y_2$

Test Results(N=200 samples, each of size n=100)

Table 2.7: Estimation results 3-3-4-b-ii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.2)	0.78155	0.399234	0.036083	0.27	0.0662	0.066	0.5975
(0.1,03)	0.71755	0.3212684	0.058701	0.26875	0.0662	0.065	0.60005
(0.2,0.1)	0.782	0.4138579	0.037208	0.2676	0.0663	0.0666	0.5995
(0.3,0.1)	0.71465	0.3169887	0.059355	0.2661	0.066	0.0666	0.6013
(0.5,0.2)	0.5859	0.1552863	0.137779	0.2628	0.06795	0.0671	0.60215

(3-3-4-b-iii)The distribution: $g(y_1, y_2)$

Y_1, Y_2 are i.i.d. $Y_1 = 2exp(-2y_1)$ and $Y_2 = 2exp(-2y_2)$

Test Results(N=200 samples, each of size n=100)

Table 2.8: Estimation results 3-3-4-b-iii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.7114	0.2065745	0.132304	0.17955	0.1604	0.15385	0.5062
(0.1,0.2)	0.77865	0.2886584	0.098576	0.17625	0.1534	0.15525	0.5151
(0.2,0.1)	0.77805	0.2983941	0.098035	0.1807	0.1534	0.1557	0.5102
(0.3,0.1)	0.71645	0.217031	0.124461	0.1826	0.15225	0.15195	0.5132
(0.5,0.2)	0.58505	0.07306155	0.202845	0.1797	0.1535	0.1531	0.5137

Test Results 3-3-4-c:

The distribution of X1, X2: COPULA MODEL $\theta = 4$

(i) The distribution: $g(y_1, y_2)$

$Y_1 \sim exp(2):\ 0.5exp(-0.5y_1)$

$Y_2 = infinite$

Test Results(N=200 samples, each of size n=100)

Table 2.9: Estimation results 3-3-4-c-i : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.7141	0.5362535	0.017474	0.4996	0	0.5004	0
(0.1,0.2)	0.77905	0.6194021	0.024148	0.4949	0	0.5051	0
(0.2,0.1)	0.7795	0.6132116	0.022747	0.5008	0	0.4992	0
(0.3,0.1)	0.7149	0.5271148	0.013131	0.49915	0	0.50085	0
(0.5,0.2)	0.5819	0.3468149	0.040801	0.4972	0	0.5028	0

(ii) The distribution: $g(y_1, y_2)$

$Y1 \sim exp(2) : 0.5exp(-0.5y_1)$

$Y_1 = Y_2$

Test Results(N=200 samples, each of size n=100)

Table 2.10: Estimation results 3-3-4-c-ii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.7206	0.4912762	0.012835	0.4434	0.05685	0.05775	0.442
(0.1,0.2)	0.78335	0.5887734	0.016168	0.44075	0.0588	0.05775	0.4427
(0.2,0.1)	0.7782	0.5803228	0.015992	0.44615	0.05635	0.05655	0.44095
(0.3,0.1)	0.717	0.4889063	0.012192	0.43995	0.0561	0.05585	0.4481
(0.5,0.2)	0.3264068	0.5859	0.043037	0.4482	0.05515	0.0579	0.43875
(0.5,0.2)	0.58815	0.3380997	0.03888	0.44	0.0583	0.0578	0.4439

(iii) The distribution $g(y_1, y_2)$:

Y_1, Y_2 is i.i.d., and $exp(2) : 0.5exp(-0.5y_1)$

Table 2.11: Estimation results 3-3-4-c-iii (N=200 samples, each of size n=100): $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.1,0.3)	0.717	0.3993221	0.044479	0.3184	0.1805	0.1747	0.3264
(0.1,0.2)	0.7718	0.4830862	0.035084	0.328	0.17935	0.1793	0.31335
(0.2,0.1)	0.7757	0.4906433	0.03434	0.3263	0.17925	0.17475	0.3197
(0.3,0.1)	0.71685	0.3915652	0.047766	0.3188	0.1813	0.1759	0.324
(0.5,0.2)	0.58175	0.2121534	0.110404	0.3241	0.1755	0.1796	0.3208

(3-3-6): Take $\theta = 6$,

$$\bar{F}_1(x_1) = e^{-x_1}, \ \bar{F}_2(x_2) = e^{-x_2}$$

$$F_1(x_1) = \int_0^{x_1} e^{-t} dt = 1 - e^{-X_1}$$

Let $F_1(x_1) = U_1$, then $X_1 = -\ln(1 - U_1)$

$$\bar{F}(x_1, x_2) = \frac{1}{[e^{6x_1} + e^{6x_2} - 1]^{1/6}}$$

$$f(x_1, x_2) = \frac{d}{dx_2 dx_1}[\bar{F}(x_1, x_2)] = \frac{5e^{5x_1}e^{4x_2}}{[e^{6x_1} + e^{6x_2} - 1]^{13/6}}$$

$$f_1(x_1) = -e^{-x_1}$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{7e^{7x_1}e^{6x_2}}{[e^{6x_1} + e^{6x_2} - 1]^{13/6}}$$

$$\begin{aligned} F(x_2|x_1) &= \int_0^{x_2} \frac{7e^{7x_1}e^{6t}}{[e^{6x_1} + e^{6t} - 1]^{13/6}} dt \\ &= 1 - e^{7X_1}[e^{6X_1} + e^{6X_2} - 1]^{-7/6} \end{aligned} \tag{2.3}$$

Let $F(x_2|x_1) = U_2$,
Then $U_2 = 1 - e^{7X_1}[e^{6X_1} + e^{6X_2} - 1]^{-7/6}$

$$X_2 = \frac{1}{6}\ln[1 - e^{6X_1} + [(1 - U_2)e^{-7X_1}]^{-6/7}]$$

Randomly generate: U_1 and U_2 are uniform distribution $(0, 1)$

$$\bar{F}_1(x_1) = e^{-x_1}, \ \bar{F}_2(x_2) = e^{-x_2}$$

$$\Longrightarrow X_1 = -\ln(1 - U_1)$$

$$X_2 = \frac{1}{6}\ln[1 - e^{6X_1} + [(1 - U_2)e^{-7X_1}]^{-\frac{6}{7}}]$$

(3-3-6-i) $Y : Y_1 \sim EXP(200),\ Y_2 = \infty$

Table 2.12: Estimation results 3-3-6-i : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000, 0.000)	0.9996	1.0000	0.0012	0.9949	0	0.0051	0
(0.375, 0.000)	0.6851	0.6867	0.0003	0.9952	0	0.0048	0
(0.750, 0.000)	0.4710	0.4727	0.0000	0.9952	0	0.0048	0
(1.125, 0.000)	0.3188	0.3205	0.0000	0.9951	0	0.0049	0
(1.500, 0.000)	0.2227	0.2246	0.0000	0.9946	0	0.0054	0
(0.000, 0.375)	0.6854	0.6870	0.0003	0.9952	0	0.0048	0
(0.375, 0.375)	0.6158	0.6175	0.0002	0.9951	0	0.0049	0
(0.750, 0.375)	0.4638	0.4657	0.0000	0.9952	0	0.0048	0
(1.125, 0.375)	0.3207	0.3227	0.0000	0.9951	0	0.0049	0
(1.500, 0.375)	0.2225	0.2241	0.0000	0.9951	0	0.0049	0
(0.000, 0.750)	0.4688	0.4708	0.0000	0.9949	0	0.0051	0
(0.375 ,0.750)	0.4614	0.4634	0.0000	0.9947	0	0.0051	0
(0.750, 0.750)	0.4189	0.4210	0.0000	0.9947	0	0.0053	0
(1.125, 0.750)	0.3158	0.3176	0.0000	0.9949	0	0.0051	0
(1.500, 0.750)	0.2198	0.2218	0.0000	0.9947	0	0.0053	0
(0.000, 1.125)	0.3249	0.3268	0.0000	0.9949	0	0.0051	0
(0.375, 1.125)	0.3231	0.3250	0.0000	0.9952	0	0.0048	0
(0.750, 1.125)	0.3185	0.3205	0.0000	0.9949	0	0.0051	0
(1.125, 1.125)	0.2879	0.2895	0.0000	0.9956	0	0.0044	0
(1.500, 1.125)	0.2151	0.2169	0.0000	0.9951	0	0.0049	0
(0.000, 1.500)	0.2236	0.2254	0.0000	0.9946	0	0.0054	0
(0.375 ,1.500)	0.2218	0.2236	0.0000	0.9949	0	0.0051	0
(0.750, 1.500)	0.2195	0.2211	0.0000	0.9951	0	0.0049	0
(1.125, 1.500)	0.2196	0.2213	0.0000	0.9950	0	0.0050	0
(1.500, 1.500)	0.1965	0.1980	0.0001	0.9953	0	0.0047	0

(3-3-6-ii) $Y:\ Y_1 \sim EXP(200),\ Y_2 = Y_1$

Table 2.13: Estimation results 3-3-6-ii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{Fe}$	$\hat{\bar{F}}$	$mse\bar{Fe}$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000, 0.000)	0.9996	1.0000	1.280748e-03	0.9944	0.0004	0.0005	0.0047
(0.375, 0.000)	0.6844	0.6859	3.092975e-04	0.9942	0.0006	0.0006	0.0046
(0.750, 0.000)	0.4700	0.4719	2.714036e-05	0.9946	0.0004	0.0004	0.0046
(1.125, 0.000)	0.3241	0.3261	1.033198e-05	0.9945	0.0004	0.0004	0.0046
(1.500, 0.000)	0.2208	0.2226	8.427519e-05	0.9947	0.0004	0.0004	0.0044
(0.000, 0.375)	0.6854	0.6869	3.111915e-04	0.9943	0.0006	0.0006	0.0046
(0.375, 0.375)	0.6188	0.6205	1.904258e-04	0.9945	0.0005	0.0005	0.0046
(0.750, 0.375)	0.4640	0.4659	2.365601e-05	0.9946	0.0005	0.0004	0.0045
(1.125, 0.375)	0.3228	0.3246	1.083043e-05	0.9946	0.0005	0.0004	0.0045
(1.500, 0.375)	0.2251	0.2267	7.977515e-05	0.9946	0.0005	0.0005	0.0045
(0.000, 0.750)	0.4727	0.4746	2.880326e-05	0.9949	0.0004	0.0005	0.0042
(0.375, 0.750)	0.4632	0.4650	2.318379e-05	0.9944	0.0005	0.0004	0.0047
(0.750, 0.750)	0.4174	0.4194	4.722666e-06	0.9946	0.0003	0.0004	0.0047
(1.125, 0.750)	0.3149	0.3170	1.403197e-05	0.9944	0.0005	0.0005	0.0047
(1.500, 0.750)	0.2191	0.2211	8.578078e-05	0.9945	0.0006	0.0006	0.0043
(0.000, 1.125)	0.3222	0.3241	1.104290e-05	0.9947	0.0004	0.0005	0.0044
(0.375, 1.125)	0.3235	0.3254	1.056248e-05	0.9947	0.0004	0.0005	0.0044
(0.750, 1.125)	0.3193	0.3211	1.220348e-05	0.9947	0.0004	0.0006	0.0044
(1.125, 1.125)	0.2874	0.2892	2.841380e-05	0.9950	0.0005	0.0003	0.0043
(1.500, 1.125)	0.2144	0.2161	9.115753e-05	0.9945	0.0006	0.0005	0.0045
(0.000, 1.500)	0.2237	0.2252	8.114177e-05	0.9948	0.0004	0.0005	0.0044
(0.375, 1.500)	0.2208	0.2226	8.423188e-05	0.9946	0.0005	0.0004	0.0045
(0.750, 1.500)	0.2197	0.2215	8.544754e-05	0.9944	0.0007	0.0005	0.0044
(1.125, 1.500)	0.2204	0.2222	1.115536e-04	0.9946	0.0005	0.0005	0.0044
(1.500, 1.500)	0.1968	0.1984	1.115536e-04	0.9946	0.0005	0.0005	0.0044

(3-3-6-iii) $Y: \; Y_1 \sim EXP(200), \; Y_1, \; Y_2$ are i.i.d. $n = 1000, N = 25,$

Table 2.14: Estimation results 3-3-6-iii : $\bar{F}(x_1, x_2)$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	$\delta(1,1)$	$\delta(1,0)$	$\delta(0,1)$	$\delta(0,0)$
(0.000, 0.000)	0.9989	1.000	0.3858	0.987	0.007	0.006	0.000
(0.375, 0.000)	0.6680	0.670	0.0843	0.995	0.002	0.003	0.000
(0.750, 0.000)	0.4834	0.489	0.0112	0.985	0.009	0.006	0.000
(1.125, 0.000)	0.3134	0.316	0.0041	0.987	0.008	0.005	0.000
(1.500, 0.000)	0.2490	0.253	0.0166	0.993	0.005	0.002	0.000
(0.000, 0.375)	0.6811	0.684	0.0920	0.987	0.005	0.008	0.000
(0.375, 0.375)	0.5988	0.600	0.0488	0.991	0.002	0.007	0.000
(0.750, 0.375)	0.4475	0.453	0.0049	0.989	0.008	0.003	0.000
(1.125, 0.375)	0.3486	0.353	0.0009	0.990	0.006	0.004	0.000
(1.500, 0.375)	0.2072	0.208	0.0291	0.997	0.002	0.001	0.000
(0.000, 0.750)	0.4628	0.464	0.0072	0.996	0.001	0.003	0.000
(0.375, 0.750)	0.4519	0.457	0.0055	0.990	0.002	0.008	0.000
(0.750, 0.750)	0.3722	0.376	0.0000	0.988	0.007	0.005	0.000
(1.125, 0.750)	0.3039	0.306	0.0055	0.991	0.006	0.003	0.000
(1.500, 0.750)	0.2119	0.216	0.0275	0.982	0.004	0.013	0.001
(0.000, 1.125)	0.3216	0.322	0.0032	0.992	0.002	0.005	0.001
(0.375, 1.125)	0.2953	0.299	0.0068	0.990	0.003	0.007	0.000
(0.750, 1.125)	0.3111	0.314	0.0044	0.994	0.004	0.002	0.000
(1.125, 1.125)	0.2701	0.274	0.0116	0.992	0.004	0.004	0.000
(1.500, 1.125)	0.2083	0.210	0.0287	0.993	0.005	0.002	0.000
(0.000, 1.500)	0.2435	0.246	0.0180	0.987	0.008	0.005	0.000
(0.375, 1.500)	0.2441	0.249	0.0179	0.991	0.004	0.005	0.000
(0.750, 1.500)	0.2488	0.250	0.0166	0.991	0.007	0.002	0.000
(1.125, 1.500)	0.2304	0.232	0.0217	0.994	0.002	0.004	0.000
(1.500, 1.500)	0.1985	0.203	0.0321	0.988	0.006	0.005	0.001

2.3 Calculation for the real data

3-4) **The real data simulation results**:

1. twins

Table 2.15: 3-4 Estimates of distributions for Twins, based on Hanley and Parnes, Dabrowska and Sen-Stute method. In multiples of 1/60

Point or set	Hanley and parnes	Dabrowska	Sen and Stute
(1,4)	0	-1	8.57
(1, 5)	0	-1	10.29
(1, 8)	12	14	41.14
(3, 4)	16	16	0
(6+, 5)	16	-	0
(7+, 5)	-	16	0
(7+, 7+)	16	16	0

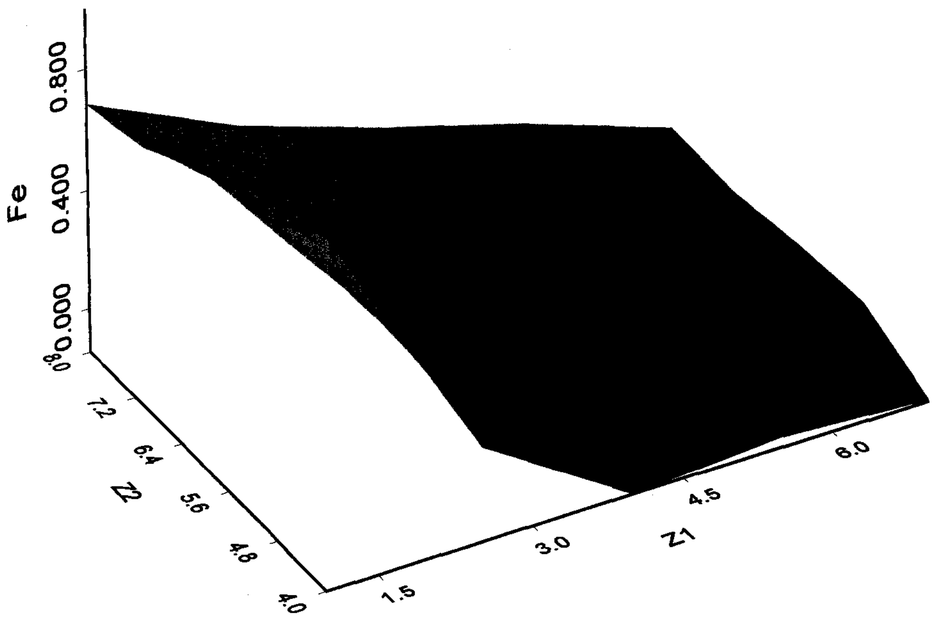

Figure 2.1: 3-4-1 Twins

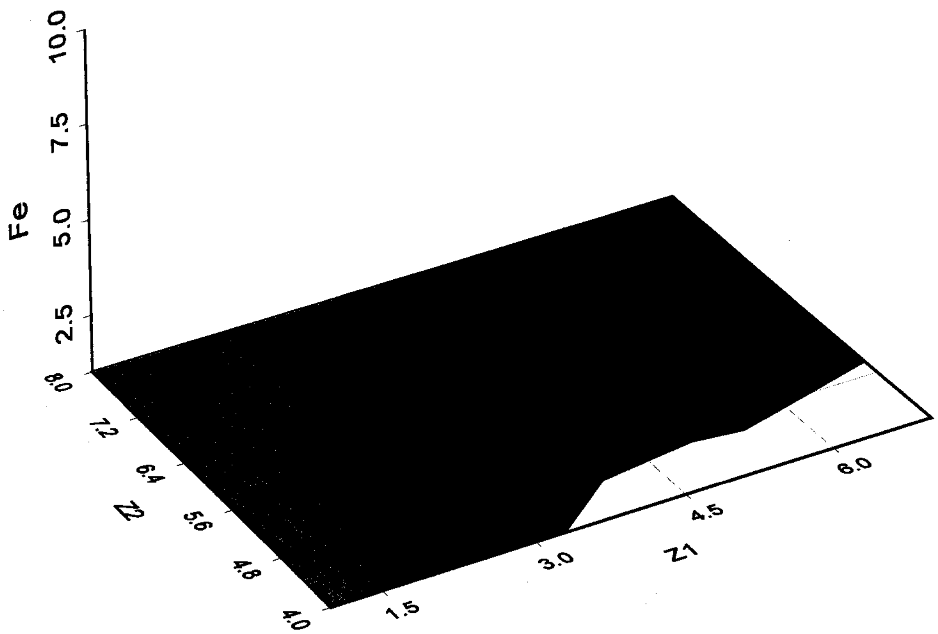

Figure 2.2: 3-4-2 Twins-contour

Comment for comparison of different estimator:

Table 3-4 shows the estimated non-zero probabilities under the Hanley and Parnes method, the Dabrowska method and A. Sen's method of Figure 3-4. They agree for the point(3.4) and the set (7+,7+). The Dabrowska method requires the marginal data, and therefore it, compared to ths Hanley and Parness estimate, moves mass from (1,4) and (1,5) and gives it to (1,8). As (1,4) and (1,5) have zero mass initially, their mass under the Dabrowska method becomes negative. For the single cesoring (6+,5), Hanley and Parnes just gives the mass to this interval, but the Dabrowska method has no mass in the marginal distribution for the interval from 6 to 7 and therefore lesds to the same mass's being concentrated on the smaller univariate interval, as such, is not assigned a probability mass. But, we can, as the sets are nested,calculate that Dabrowska's method gives a total mass of 16/60 in the interval (6+, 5) and Hanley and Parnes method has a probability between 0 and 16/60 for the interval (7+, 5). But for A. Sen method, all the mass points are nonnegative.

3-5-1) Pairs of Kidney

Table 2.16: 3-5-1. Estimation of survivor function for infection in kidney catheters. Data of McGilchrist and Aisbett (1991)

Observation	Estimation of Survivor function
(8, 16)	0.71878
(22, 28)	0.46100
(30, 12)	0.52161
(7, 9)	0.81002
(53, 196)	0.15748
(7, 333)	0.04724
(96, 38)	0.15678
(536, 25+)	0.00000
(185, 177)	0.05590
(22+, 159+)	0.21180
(152, 562)	0.02415
(3, 66)	0.34346
(12, 40)	0.41406
(132, 156)	0.10380
(2, 25)	0.76811
(27, 58)	0.27156
(152, 30)	0.16470
(119, 8)	0.32379
(6+, 78)	0.33758
(23, 13+)	0.53608

continue

Observation	Estimation of Survivor function
(447, 318)	0.02402
(24, 245)	0.07061
(511, 30)	0.03904
(15,154)	0.29020
(141,8+)	0.16470
(149+,70+)	0.10434
(17, 4+)	0.63847
(292, 114)	0.04831
(15, 108+)	0.31449
(402, 24+)	0.06306
(39, 46+)	0.23740
(113+, 201)	0.04817
(34, 30)	0.37017
(130, 26)	0.29776
(+5 , 43)	0.40071
(190, 5+)	0.08735
(54+,16+)	0.32645
(63, 8+)	0.35248

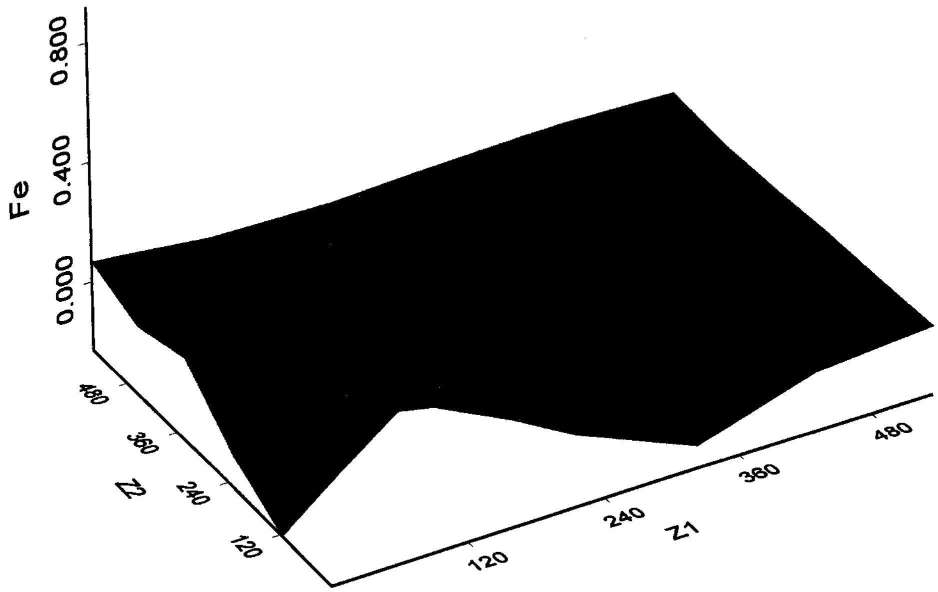

Figure 2.3: 3-5-1-1 Kidney-drap

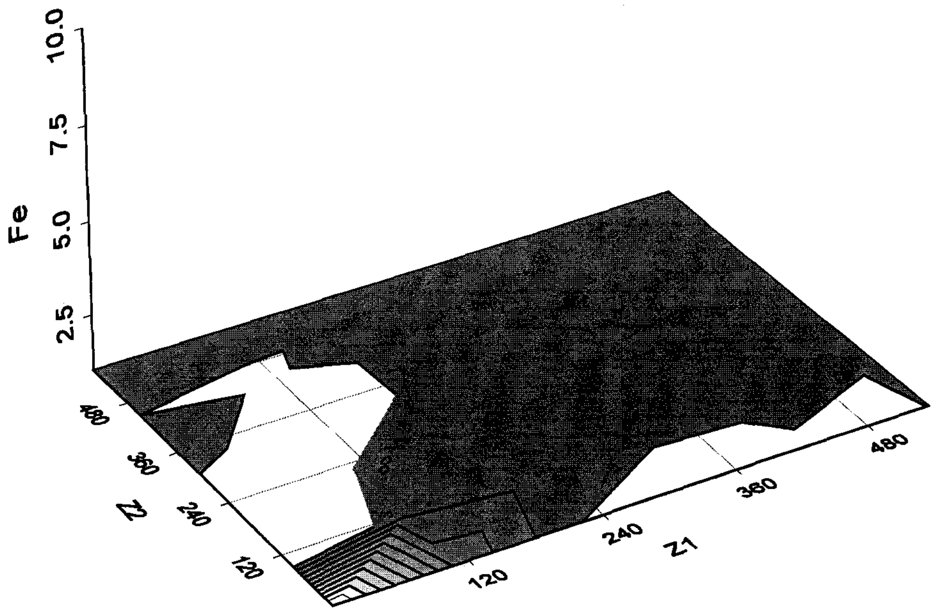

Figure 2.4: 3-5-1-2 Kidney-contour

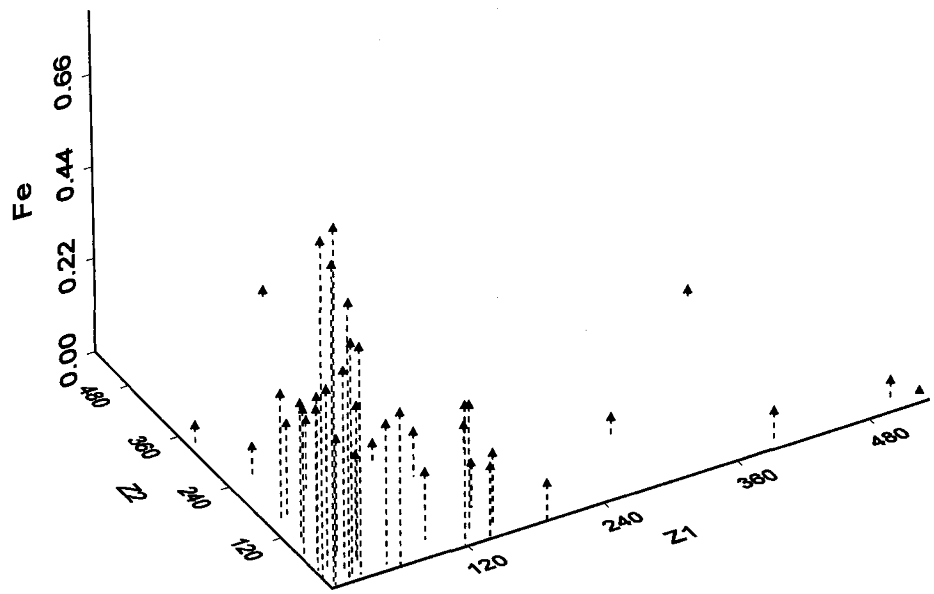

Figure 2.5: 3-5-1-3.Kidney-scater

3-5-2) **Estimation of survivor function of male's kidney**

Table 2.17: 3-5-2. Estimation of survivor function for infection in kidney catheters of male. Data of McGilchrist and Aisbett (1991)

Observation	Estimation of Survivor function
(8 , 16)	0.54287
(22, 28)	0.22858
(30, 12)	0.22858
(7 , 9)	0.77145
(152 ,562)	0.11429
(12, 40)	0.31429
(2 , 25)	0.54287
(15 ,154)	0.22858
(17 , 4+)	0.34287
(63 ,8+)	0.11429

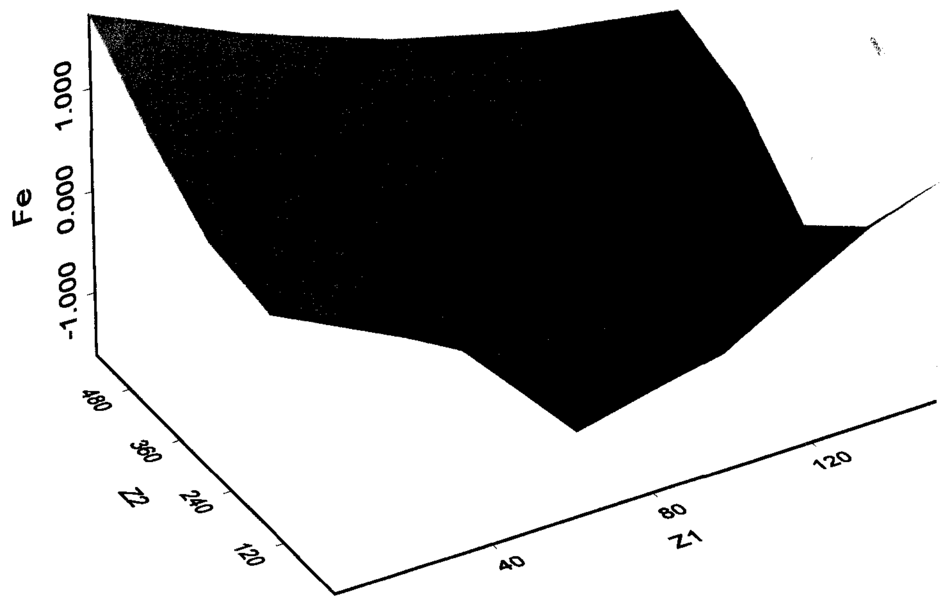

Figure 2.6: 3-5-2-1.Kidney-male-drap

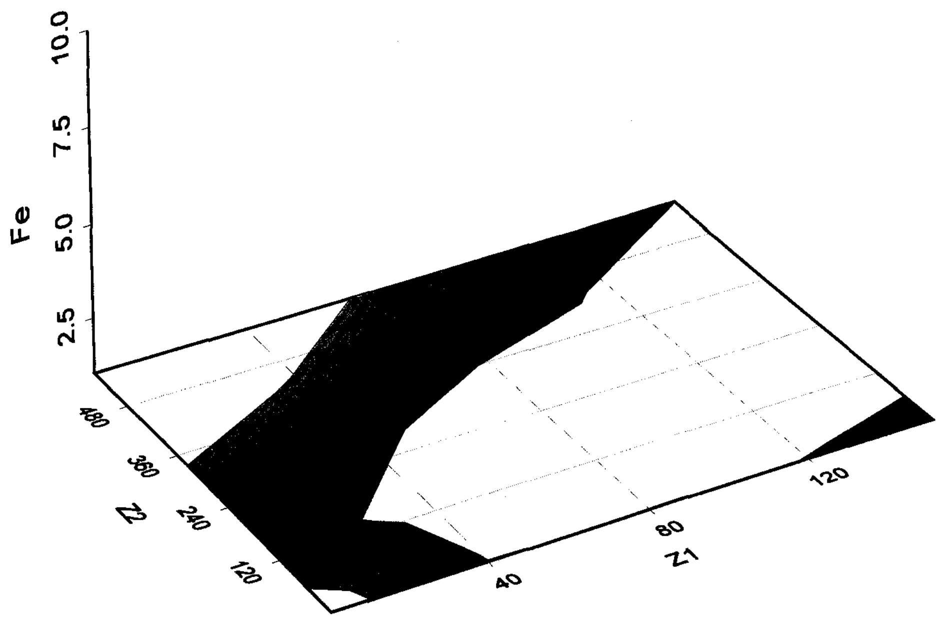

Figure 2.7: 3-5-2-2.Kidney-male-contour

3-5-3) **Estimation of survivor function of female's kidney**

Table 2.18: 3-5-3. Estimation of survivor function for infection in kidney catheters of female. Data of McGilchrist and Aisbett (1991)

Observation	Estimation of Survivor function
(53, 196)	0.06250
(7, 333)	0.08523
(96, 38)	0.85227
(536, 25+)	0.00000
(185 , 177)	0.00000
(22+ , 159+)	0.06250
(13 , 66)	0.06250
(132 , 156)	0.00000
(27 , 58)	0.06250
(152 , 30)	0.00000
(119 , 8)	0.00000
(6+ , 78)	0.14773
(23 , 13+)	0.91477
(447, 318)	0.00000
(24 , 245)	0.00000
(511 , 30)	0.00000
(141 , 8+)	0.00000
(149+, 70+)	0.00000
(292, 114)	0.00000

continue

Observation	Estimation of Survivor function
(15 , 108+)	0.06250
(402, 24+)	0.00000
(39 , 46+)	0.06250
(113+, 201)	0.00000
(34 , 30)	0.91477
(130 , 26)	0.00000
(5+ , 43)	0.14773
(190 , 5+)	0.00000
(54+ , 16+)	0.85227

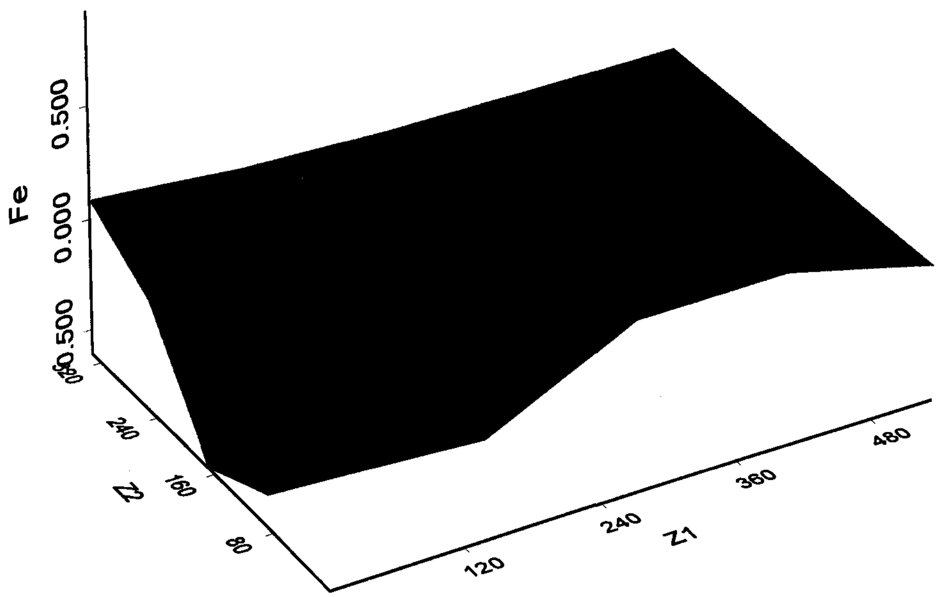

Figure 2.8: 3-5-3-1.Kidney-female-drap

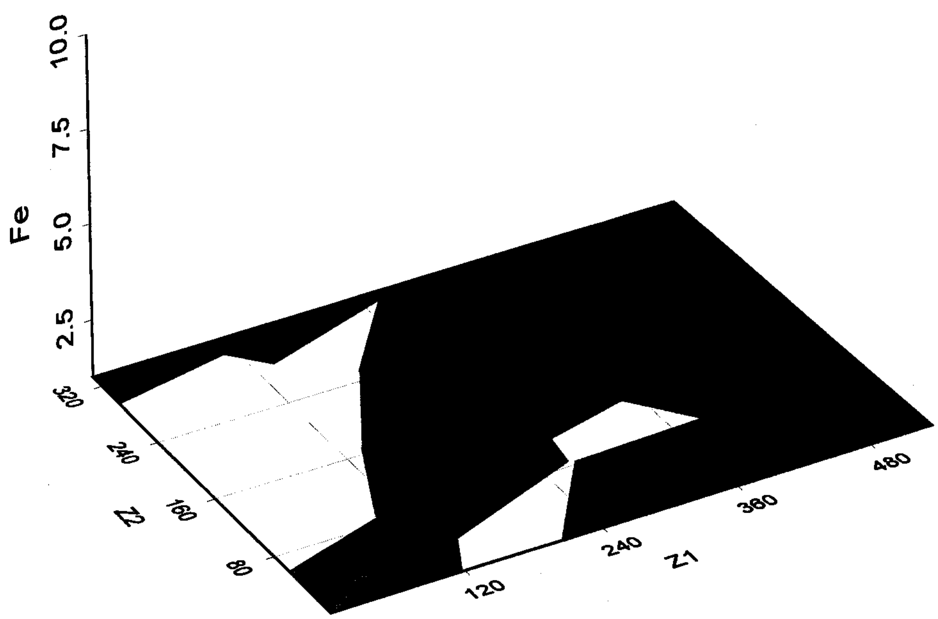

Figure 2.9: 3-5-3-2.Kidney-female-contour

Chapter 3

Estimation of Variance of the Sen-Stute Estimator

3.1 Definition: Influence function

For a given distribution F in $\mathbb{R}^d$ and an $\epsilon > 0$, the version of F contaminated by an ε amount of an arbitrary distribution G is denoted by $F(\epsilon, G) = (1-\epsilon)F + \epsilon G$. The maximum bias of a given location functional T under an ϵ amount of contamination at F is defined as [Hampel, Ronchetti, Rousseeuw and Stahel (1986)]

$$B(\varepsilon; T, F) = sup_G \|T(F(\varepsilon, G)) - T(F)\|$$

, where (and hereafter) $\|.\|$ stands for Euclidean norm.

1. The influence function (IF) of T at a given point $x \in \mathbb{R}^d$ R for a given F is defined as

$$IF(x; T, F) = lim_{\epsilon \longrightarrow 0^+}(T(F(\epsilon, \delta_x)) - T(F))/\epsilon$$

, where δ_x is the point-mass probability measure at $x \in \mathbb{R}^d$.

Influence function of the estimator

Note that Equations (2.3)–(2.6) and their solutions are completely *dimension-free*, i.e., is valid for $\Delta_i := (\delta_{1i}, \ldots, \delta_{mi})$, $Z_i = (Z_{1i}, \ldots, Z_{mi})$ for $m \geq 1$, with the definitions $\delta_i =$

$\prod_{j=1}^{m} \delta_{ji}$ and $a_{ik} = \mathbf{1}\{Z_k \geq Z_i\}$ where the inequality is defined in the coordinate-wise sense. Hence in this section we shall use scalar notation also for vector variables, with the above interpretation.

Now to derive the *influence functions* for the estimators $\bar{F}_n(x)$ and $\int \varphi dF_n$ for a given $\varphi(\cdot)$, let P denote the distribution of (δ, Z) and P_n the empirical distribution of (δ_i, Z_i), $1 \leq i \leq n$. Also, let $T_x(P) := \bar{F}(x-)$, $T_\varphi(P) := \int \varphi dF$, and let $T_x(P_n)$, $T_\varphi(P_n)$ be their estimators, respectively, obtained via Eq.(2.5)–(2.6). Thus we rewrite Eq. (2.2) and (2.3) as the eigenvalue problems

$$\begin{aligned} T_x(P) &= \int \mathbf{1}\{t \geq x\} T_t(P) \frac{dH^{11}(t)}{\bar{H}(t-)}, \\ T_x(P_n) &= \int \mathbf{1}\{t \geq x\} T_t(P_n) \frac{dH_n^{11}(t)}{\bar{H}_n(t-)}, \end{aligned} \tag{3.1}$$

with the initial conditions $T_0(P) = 1,\ T_0(P_n) = 1$.

Note also that, for a function $\varphi(\cdot)$ satisfying $\varphi(x) = 0$ if $x \notin [0, \tau]$ for some τ with $\bar{H}(\tau) > 0$,

$$\begin{aligned} T_\varphi(P) &= \int \varphi(t) T_t(P) \frac{dH^{11}(t)}{\bar{H}(t-)}, \\ T_\varphi(P_n) &= \int \varphi(t) T_t(P_n) \frac{dH_n^{11}(t)}{\bar{H}_n(t-)}. \end{aligned} \tag{3.2}$$

3.2 The estimator of variance of bivariate survivor function estimator

The influence function $L_x(P_n)$ of $T_x(P_n) = \bar{F}_n(x-)$ derived by Sen and Stute(2007) as below, where $H_n^{11}(t), G(t)$ is as defined in Chapter 2:

$$\begin{aligned} & L_x(P_n) - \int \int \{t \geq x\} L_t(P_n) \frac{dF(t)}{\bar{F}(t-)} \\ &= \int \mathbf{1}\{t \geq x\} \left[\frac{dH_n^{11}(t)}{\bar{G}(t-)} - \bar{H}_n(t-) \frac{dF(t)}{\bar{H}(t-)} \right] \end{aligned} \tag{3.3}$$

$$\tilde{\mathbf{p}} = \tilde{\mathbf{B}} \tilde{\mathbf{A}} \tilde{\mathbf{p}}$$

$$\mathbf{p} = (p_1, \ldots, p_n), \ \sum_{i=1}^{n} p_i = 1, \tag{2.5}$$

$$\bar{F}_n(x_1, x_2) = \sum_{i=1}^{n} p_i I(z_{1i} > x_1, z_{2i} > x_2)$$

$$1) \ \bar{F}_n(x_1-, x_2-) \longrightarrow \bar{F}(x_1-, x_2-) = P(X_1 > x_1, X_2 > x_2)$$

$$(\bar{F}_n \ - \ \bar{F}) \ \sim \ N\left(0, \ \frac{v(x_1, x_2)}{n}\right)$$

$$\frac{v(x_1, x_2)}{n} = E[L_x^2(P_n)], \quad E[L_x(P_n)] = 0, \ x = (x_1, x_2)$$

where $L_x(P_n)$ is given by

$$L_x(P_n) \ = \ a_x(P_n) - \ \bar{F}(x) a_0(P_n)$$

and

$$a_x(P_n) \ = \ \mathfrak{z}_n(x) + \sum_{r=1}^{\infty} \int \cdots \int I(x \le y_1 \le \ldots \le y_n) \mathfrak{z}_n(y_r) \frac{F(y_1)}{\bar{F}(y_1)} \cdots \frac{F(y_r)}{\bar{F}(y_r)}$$

$$a_0(P_n) \ = \ \mathfrak{z}_n(0) + \sum_{r=1}^{\infty} \int \cdots \int I(0 \le y_1 \le \ldots \le y_r) \mathfrak{z}_n(y_r) \frac{F(y_1)}{\bar{F}(y_1)} \cdots \frac{F(y_r)}{\bar{F}(y_r)}$$

$$\mathfrak{z}_n(x) \ = \ \int I(t \ge x) \bar{F}(t) \left[\frac{dH_n^{11}(t)}{\bar{H}(t)} \ - \ \bar{H}_n(t) \frac{dH_n(t)}{\bar{H}^2(t)} \right]$$

Thus

$$
\begin{aligned}
L_x(p_n) = \mathfrak{z}_n(x) \; - \; \bar{F}(x)\mathfrak{z}_n(0) \\
+ \sum_{i=1}^{n} \int \ldots \int \left[I(x \leq y_1 \leq \ldots \leq y_r) - \; \bar{F}(x) I(0 \leq y_1 \leq \ldots \leq y_r) \right] \\
\mathfrak{z}_n(y_r) \frac{dF(y_1)}{\bar{F}(y_1)} \cdots \frac{dF(y_r)}{\bar{F}(y_r)}
\end{aligned}
\tag{3.4}
$$

and

$$
\mathfrak{z}_x(x) \; - \; \bar{F}(x)\mathfrak{z}_n(0) \; = \; \int \left[I(t \geq x) \; - \; \bar{F}(x) \right] \bar{F}(x) \left[\frac{dH_n^{11}(t)}{\bar{H}(t)} \; - \; \bar{H}_n(t) \frac{dH^{11}(t)}{\bar{H}^2(t)} \right]
$$

$$
\begin{aligned}
H^{11}(t) = E(H_n^{11}(t)) = \; & P(z_{1i} \leq t_1, \; z_{2i} \leq t_2, \; \delta_{1i} = 1, \; \delta_{2i} = 1) \\
= & P(X_{1i} \; \leq t_1, \; X_{2i} \; \leq t_2, \; X_{1i} \leq Y_{1i}, \;\; X_{2i} \; \leq Y_{2i}) \\
= & \int_0^{t_1} \int_0^{t_2} \bar{G}(w_1, \; w_2) dF(w_1, \; w_2)
\end{aligned}
\tag{3.5}
$$

$$
\bar{H}_n(t) = \; \frac{1}{n} \sum_{i=1}^{n} I(z_{1i} > \; t_1, \; z_{2i} \; > \; t_2 \;)
\tag{3.6}
$$

$$
\begin{aligned}
\bar{H}(t) = \; & E(\bar{H}_n(t)) \; = \; P((z_{1i} > \; t_1, \; z_{2i} \; > \; t_2 \;) \\
& = P(X_{1i} \; > \; t_1, \; X_{2i} \; > t_2, \; Y_{1i} \; > t_1, \;\; Y_{2i} \; > t_2) \\
& = \bar{F}(t_1, \; t_2 \;) \bar{G}(t_1, \; t_2)
\end{aligned}
\tag{3.7}
$$

Also, for any function $\varphi(w_1, \; w_2)$,

$$
\int \varphi(w_1, \; w_2) \; dH^{11}(w_1, \; w_2) \equiv \; \int \varphi(w_1, \; w_2) \; \bar{G}(w_1, \; w_2) dF(w_1, \; w_2)
$$

$$
\int \varphi(w_1, \; w_2) \; dH_n^{11}(w_1, \; w_2) = \frac{1}{n} \sum_{i=1}^{n} \delta_{1i} \delta_{2i} \varphi(z_{1i}, \; z_{2i})
$$

Hence

$$
\begin{aligned}
\mathfrak{z}_n(x) = & \int I(t \geq x)\bar{F}(t)\left[\frac{dH_n^{11}(t)}{\bar{H}(t)} - \bar{H}_n(t)\frac{dH^{11}(t)}{\bar{H}^2(t)}\right] \\
& = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} I(z_{1i} \geq x_1,\ z_{2i} \geq x_2)\frac{\bar{F}(z_{1i},\ z_{2i})}{\bar{H}(z_{1i},\ z_{2i})} \\
& - \int \bar{H}_n(t) I(t \geq x)\bar{F}(t)\frac{\bar{G}(t_1,\ t_2)dF(t_1,\ t_2)}{\bar{H}^2(t_1,\ t_2)}
\end{aligned}
\tag{3.8}
$$

One-dimension case:

Data:

$$
(\delta_i,\ z_i), 1 \leq\ i\ \leq n,\ \delta_i\ =\ I(X_i\ \leq\ Y_i),\ z_i\ =\ X_i \wedge\ Y_i
$$

$$
\begin{aligned}
L_x(p_n) &= -\bar{F}(x)\int I(t < x)\left[\frac{dH_n^{11}(t)}{\bar{H}(t)} - \bar{H}(t)\frac{dH^{11}(t)}{\bar{H}(t)}\right] \\
&= -\bar{F}(x)\left[\frac{1}{n}\sum_{i=1}^{n} \delta_{1i} I(z_i \leq x)\frac{1}{\bar{H}(z_i)} - \int\right] \bar{H}_n(t) I(t < x)\frac{dH^{11}(t)}{\bar{H}(t)}
\end{aligned}
\tag{3.9}
$$

$$
E(L_x^2(p_n)) = \ \frac{\bar{F}^2(x)}{n}E\left[\frac{\delta_i I(z_i \leq x)}{\bar{H}(z_i)}\ \ -\ \ \int I(z_i > t)I(x > t)\frac{H^{11}(t)}{\bar{H}(t)}\right]^2 = \frac{v(x)}{n}
$$

$v(x)$ is estimated by "Greenwoods's Formula"

$$
\hat{v}(x)\ \ =\ \ \bar{F}_n^2(x)\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i I(z_i < x)}{\bar{H}_n^2(z_i)} - \frac{2}{n}(\dots.) + \frac{1}{n}\sum_i (\ )^2\right]
$$

$$
b_i\ \ =\ \ \frac{\delta_i}{n\bar{H}_n(z_i)}\ \ =\frac{H^{11}}{\bar{H}}
$$

1) To estimate $var(\bar{F}_n(X)) = E(L_x^2(P_n))$

2) $L_x(P_n) \;=\; a_x(P_n) - \bar{F}(x)a_0(P_n)$

$E(L_x^2(P_n)) \;=\;$ different, except in 1-dimension

1-dimension:

$$\begin{aligned}
L_x(P_n) &= -\bar{F}(x)\int I(t\le x)\left[\frac{dH_n^{11}(t)}{\bar{H}(t)} - \bar{H}_n(t)\frac{dF(t)}{\bar{H}(t)\bar{F}(t)}\right] \\
&= -\bar{F}(x)\int I(t\le x)\left[\frac{dH_n^{11}(t)}{\bar{H}(t)} - \bar{H}_n(t)\frac{H^{11}(t)}{\bar{H}^2(t)}\right] \\
&= -\bar{F}(x)\left[\frac{1}{n}\sum_{i=1}^{n}\delta_{1i}I(z_i\le x)\frac{1}{\bar{H}(z_i)} - \int\left(\frac{1}{n}\sum_{i=1}^{n}I(z_i>x)\right)I(t\le x)\frac{H^{11}(t)}{\bar{H}^2(t)}\right] \\
&= -\frac{\bar{F}^2(x)}{n}\sum_{i=1}^{n}\left[\frac{\delta_i I(z_i\le x)}{\bar{H}(z_i)} - \int I(z_i>t,\ x\ge t)\frac{dH^{11}(t)}{\bar{H}^2(t)}\right]
\end{aligned} \tag{3.10}$$

Let $\varphi(\delta_i,\ z_i) = \frac{\delta_i I(z_i\le x)}{\bar{H}(z_i)} - \int I(z_i>t,\ x\ge t)\frac{dH^{11}(t)}{\bar{H}^2(t)}$

$$E(L_x^2(P_n)) \;=\; \frac{\bar{F}^2(x)}{n}E\varphi_x^2(\delta_i,\ z_i)$$

Now $E\varphi_x^2(\delta_i,\ z_i) \;=\; E\left[\left(\frac{\delta_i I(z_i\le x)}{\bar{H}(z_i)}\right)\right] = E\left[\frac{\delta_i I(z_i\le x)}{\bar{H}^2(z_i)}\right]$

Hence $E\varphi_x^2(\delta_i,\ z_i)$ is estimated by $\frac{1}{n}\sum_{j=1}^{n}\frac{\delta_i I(z_j\le x)}{\bar{H}_n^2(z_j)}$

$$
\begin{aligned}
v\hat{a}r(\bar{F}_N(X)) &= \frac{1}{n}\left[\bar{F}_n^2(x)\frac{1}{n}\sum_{j=1}^{n}\frac{\delta_i I(z_j \le x)}{\bar{H}_n^2(z_j)}\right] \\
&= \frac{1}{n}\left[\bar{F}_n^2(x)\sum_{j=1}^{n}\frac{\delta_{[j]} I(z_{(j)} \le x)}{(n-j+1)^2}\right],
\end{aligned} \tag{3.11}
$$

$$
\begin{array}{cccc}
z_{(1)} \le & z_{(2)} \le & \ldots \le & z_{(n)} \\
\downarrow & \downarrow & & \downarrow \\
\delta_{[1]} & \delta_{[2]} & & \delta[n]
\end{array}
$$

$\sum_{j=1}^{n}\frac{\delta_{[j]} I(z_{(j)} \le x)}{(n-j+1)^2}$ is Greenwood's Formula

3) In general (2-dimension or more), by Eq. (17)+(18)

$$
\begin{cases}
L_x(P_n) - \int I(t \ge x) L_t(P_n)\frac{dF(t)}{\bar{F}(t)} = \mathfrak{z}_n(x), \\
L_0(P_n) = 0
\end{cases}
$$

$\Rightarrow$

$$
\begin{aligned}
L_x(P_n)L_y(P_n) &- \int I(t \ge x)L_y(P_n)Lt(P_n)\frac{dF(t)}{\bar{F}(t)} - \int I(s \ge y)L_x(P_n)L_s(P_n)\frac{dF(t)}{\bar{F}(s)} \\
&+ \int\int I(t \ge x)I(s \ge y)L_t(p_n)L_s(P_n)\frac{dF(t)}{\bar{F}(t)}\frac{dF(s)}{\bar{F}(s)} \\
&= \mathfrak{z}_n(x)\mathfrak{z}_n(y)
\end{aligned} \tag{3.12}
$$

Let $M(x,y) = E[L_x(P_n)L_y(P_n)]$, so that $M(x,x) = E[L_x^2(P_n)]$

Thus we have,

$$
\begin{aligned}
M(x,y) &- \int I(t \ge x)M(y,t)\frac{dF(t)}{\bar{F}(t)} - \int I(s \ge y)M(x,s)\frac{dF(s)}{\bar{F}(s)} \\
&+ \int\int I(t \ge x)I(s \ge y)M(t,s)\frac{dF(t)}{\bar{F}(t)}\frac{dF(s)}{\bar{F}(s)} \\
&= E[\mathfrak{z}_n(x)\mathfrak{z}_n(y)]
\end{aligned} \tag{3.13}
$$

(1)

$$
\begin{cases}
L_x(P_n) - \int I(t \ge x)L_t(P_n)\frac{dF(t)}{\bar{F}(t)} = \mathfrak{z}_n(x), \; L_0(P_n) = 0 \\
L_y(P_n) - \int I(t' \ge y)L_{t'}(P_n)\frac{dF(t')}{\bar{F}(t')} = \mathfrak{z}_n(y), \; L_0(P_n) = 0
\end{cases}
$$

Let $v(x,y) = E(L_x(P_n)L_y(P_n))$, $v(x,x) = E(L_x^2(P_n))$

$$\begin{cases} v(x,y) \ - \ \int I(t \geq x) v(t,y) \frac{dF(t)}{\bar{F}(t)} \ - \ \int I(t' \geq y) v(x,t') \frac{dF(t')}{\bar{F}(t')} \\ + \ \int \int I(t \geq x) I(t' \geq y) v(t,t') \frac{dF(t)}{\bar{F}(t)} \frac{dF(t')}{\bar{F}(t')} \ = \ E(\mathfrak{z}_n(x) \mathfrak{z}_n(y)), \\ v(0,y) \ = \ v(x,0) \ = \ 0. \end{cases}$$

(2) Sample Version:

Take $x \ = \ z_i, \ y \ = \ z_j$, and let $\hat{v}(z_i, \ z_j) \ \ v_{ij}$

Then

$$\begin{aligned} v_{ij} - & \int I(t \ \geq \ z_i) v(t, z_j) \frac{aH_n^{11}(t)}{\bar{H}_n(t)} \\ - & \int I(t' \geq z_j) v(z_i, t') \frac{aH_n^{11}(t')}{\bar{H}_n(t')} \\ + & \int I(t \ \geq \ z_i) I(t' \geq z_j) v(t,t')) \frac{aH_n^{11}(t)}{\bar{H}_n(t)} \frac{aH_n^{11}(t')}{\bar{H}_n(t')} \\ = & \ \hat{E}(\mathfrak{z}_n(z_i) \mathfrak{z}_n(z_j)) \end{aligned} \tag{3.14}$$

Initial conditions:

$$\left[\begin{array}{l} v(0,y) \ = \ 0 \\ \Leftrightarrow \begin{cases} - \ \int v(t,y) \frac{dF(t)}{\bar{F}(t)} \\ + \ \int \int I(t' \geq y) v(t,t') \frac{dF(t)}{\bar{F}(t)} \frac{dF(t')}{\bar{F}(t')} \ = \ E[\mathfrak{z}_n(0) \mathfrak{z}_n(y))] \end{cases} \\ \\ v(x,0) \ = \ 0 \\ \Leftrightarrow \begin{cases} - \ \int v(x,t') \frac{dF(t')}{\bar{F}(t')} \\ + \ \int \int I(t \geq x) v(t,t') \frac{dF(t)}{\bar{F}(t)} \frac{dF(t')}{\bar{F}(t')} \ = \ E[\mathfrak{z}_n(x) \mathfrak{z}_n(0))] \end{cases} \end{array} \right.$$

Sample version of initial conditions:

$$\left[\begin{array}{l} z_i \ = \ 0 \Rightarrow \ v_{il} \ = \ 0, \ a_{ik} \ = \ 1 : \\ \quad - \ \sum_{k=1}^n b_k v_{kj} \ + \ \sum_{k=1}^n \sum_{l=1}^n b_k a_{jl} b_l v_{kl} \ = \ \mathfrak{z}_{0j}, \ 1 \leq j \leq n \\ \\ z_j \ = \ 0 \ \Rightarrow v_{kj} \ = \ 0, \ a_{jl} \ = \ 1 : \\ \quad - \sum_{l=1}^n b_l v_{il} \ + \ \sum_{k=1}^n \sum_{l=1}^n a_{ik} b_k b_l v_{kl} \ + \ \mathfrak{z}_{i0}, \ 1 \leq j \leq n \end{array} \right.$$

$\mathfrak{z}_{0j} \ = \ $ put $a_{ik} \ = \ 1$ in Eq.(4)

$\mathfrak{z}_{i0} \;=\;$ put $a_{jk} = 1,\; a_{jl} = 1$ in Eq.(4)

$$\begin{aligned} & v_{ij} - \sum_{k=1}^{n} I(z_k \geq z_i) v_{kj} b_k \;-\; \sum_{l=1}^{n} I(z_l \geq z_j) v_{il} b_l \\ & + \sum_{k=1}^{n}\sum_{l=1}^{n} I(z_k \geq z_i) I(z_l \geq z_j) v_{kl} b_k b_l \\ & = \hat{E}(\mathfrak{z}_n(z_i)\mathfrak{z}_n(z_j)) \end{aligned} \tag{3.15}$$

Recall $a_{ij} \;=\; I(z_j \geq \; z_i)$,

$$\begin{aligned} & v_{ij} - \sum_{k=1}^{n} a_{ik} b_k v_{kj} - v_{ij} - \sum_{k=1}^{n} a_{ik} b_k v_{kj} - \sum_{l=1}^{n} a_{jl} b_l v_{il} + \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik} b_k a_{jl} b_l v_{kl} \\ & = \hat{E}(\mathfrak{z}_n(z_i)\mathfrak{z}_n(z_j)) \\ & = \mathfrak{z}_{ij} \end{aligned} \tag{3.16}$$

Let $\mathbf{V} \;=\; ((v_{ij}))_{n\times n},\; \hat{z} \;=\; ((\mathfrak{z}_{ij}))_{n\times n}$. Then

$$\mathbf{V} \;-\; \mathbf{ABV} \;-\; \mathbf{VBA}^T \;+\; \mathbf{ABVBA}^T \;=\; \hat{\mathbf{z}} \quad \rightarrow (**)$$

where $\mathbf{A} \;=\; ((a_{ij}))$, $\mathbf{B} \;=\; diag(b_1, \ldots,\; b_n) \;=\; ((b_j\delta_{ij}))$,
$\delta_{ij} \;=\; 1$ if $i \;=\; j$, $\delta_{ij} \;=\; 0$ if $i \neq \; j$: "Kronecker Delta"

(3)

$$\begin{cases} z_n(x) \;=\; \int I(t \;\geq x) \;\left[\frac{dH_n^{11}(t)}{\bar{G}(t)} \;-\; \bar{H}_n(t)\frac{dF(t)}{\bar{H}(t)}\right] \\ z_n(y) \;=\; \int I(t' \;\geq y) \;\left[\frac{dH_n^{11}(t')}{\bar{G}(t')} \;-\; \bar{H}_n(t')\frac{dF(t')}{\bar{H}(t')}\right] \end{cases}$$

$$\mathfrak{z}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\bar{G}(t_i)} I(z_i \geq \; x) \;-\; \frac{1}{n}\sum_{i=1}^{n} \int I(z_i \geq t) I(t \;\geq x) \frac{dF(t)}{\bar{H}(t)}$$

$$\mathfrak{z}_n(y) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\bar{G}(t_i)} I(z_i \geq \; y) \;-\; \frac{1}{n}\sum_{i=1}^{n} \int I(z_i \geq t') I(t' \;\geq y) \frac{dF(t')}{\bar{H}(t')}$$

$$
\begin{aligned}
E(\mathfrak{z}_n(x)\mathfrak{z}_n(y)) = & \frac{1}{n}[E\left(\frac{\delta_i^1}{\bar{G}^2(z_i)}I(z_i \geq x, \quad z_i \geq y)\right) \\
& - \int E\left(\frac{\delta_i}{\bar{G}(z_i)}I(z_i \geq t, \ z_i \geq y)\right) I(t \geq x)\frac{dF(t)}{\bar{H}(t)} \\
& - \int E\left(\frac{\delta_i}{\bar{G}(z_i)}I(z_i \geq t, \ z_i \geq y)\right) I(t' \geq y)\frac{dF(t')}{\bar{H}(t')} \\
& + \int\int E\left(I(z_i \geq t, \ z_i \geq t')\right) I(t \geq x)I(t' \geq y)\frac{dF(t)}{\bar{H}(t)}\frac{dF(t)}{\bar{H}(t')}] \\
= & \frac{1}{n}[\int I(t \geq x, t)\frac{dF(t)}{\bar{G}(t)} \quad - \quad \int \bar{F}(t \vee y)I(t \geq x)\frac{dF(t)}{\bar{H}(t)} \\
& - \int \bar{F}(x \ \vee t')I(t' \geq y)\frac{dF(t')}{\bar{H}(t')} \\
& + \int\int \bar{H}(t \vee t')I(t \geq x)I(t' \geq y)\frac{dF(t)}{\bar{H}(t)}\frac{dF(t')}{\bar{H}(t')}]
\end{aligned}
\tag{3.17}
$$

Where $t \vee y \ = \ (t_1, t_2) \ \vee (y_1, y_2) \ = \ (t_1 \vee y_1, t_2 \vee y_2)$

Hence

$$
\begin{aligned}
E(\mathfrak{z}_n(x)\mathfrak{z}_n(y)) = & \frac{1}{n}[\int I(t \geq x)I(t \geq y)\frac{dF(t)}{\bar{G}(t)} \\
& + \int\int I(t \ngeq t')I(t' \ngeq t)\bar{H}(t \vee t')I(t \geq x)I(t' \geq y)\frac{dF(t)}{\bar{H}(t)}\frac{dF(t')}{\bar{H}(t')}] \\
= & \frac{1}{n}[\int I(t \geq x)I(t \geq y)F^{-2}(t)\frac{dH^{11}(t)}{\bar{H}^2(t)} \\
& + \int\int I(t \ngeq t')I(t' \ngeq t)\bar{H}(t \vee t')\bar{F}(t \vee t')\bar{F}(t)\bar{F}(t')I(t \geq x)I(t' \geq y) \\
& \frac{dH^{11}(t)}{\bar{H}^2(t)}\frac{dH^{11}(t')}{\bar{H}^2(t')}]
\end{aligned}
\tag{3.18}
$$

(4) Sample Version of $E(\mathfrak{z}_n(x)\mathfrak{z}_n(y))$:

$$
\begin{aligned}
\mathfrak{z}_{ij} = & \hat{E}\left(\mathfrak{z}_n(z_i)\mathfrak{z}_n(z_j)\right) \\
= & \frac{1}{n}[n\sum_{k=1}^{n} I(z_k \geq z_i)I(z_k \geq z_j)\bar{F}_n^2(z_k)\frac{\delta_k}{\left(\sum_r a_{kr}\right)^2} + n^2\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{k\neq l}(1-a_{lk})(1-a_{kl}) \\
& \bar{H}_n(z_k \vee z_l)\bar{F}_n(z_k)\bar{F}_n(z_l)I(z_k \geq z_i)I(z_l \geq z_j)]\frac{\delta_k}{\left(\sum_r a_{kr}\right)^2}\frac{\delta_l}{\left(\sum_s a_{ls}\right)^2}
\end{aligned}
\tag{3.19}
$$

Where

$$a_{ik} = I(z_k \geq z_i),\ a_{jk} = I(z_k \geq z_j),\ a_{ik} = I(z_k \geq z_i),\ a_{jl} = I(z_l \geq z_j),$$

$$\bar{F}_n(a_1, a_2) = \sum p_i I(z_{i1} \geq a_1, z_{i2} \geq a_2),$$

$$\bar{F}_n(z_k) = \sum p_i I(z_{i1} \geq z_{k1}, z_{i2} \geq z_{k2}),$$

$$\bar{H}_n(a_1, a_2) = \frac{1}{n} \sum_i I(z_{i1} \geq a_1, z_{i2} \geq a_2),$$

$$z_k \vee z_l = (z_{k1} \vee z_{l1}, z_{k2} \vee z_{l2}) = (a_1, a_2).$$

(5) To solve $(*)$, write $\mathbf{V} = ((v_{ij}))$
in vector form, ie., $\underline{\mathbf{V}} = (v_{11}, v_{12}, \ldots, v_{nn})_{n^2 \times 1}$

$$\begin{cases} (\mathbf{I} - \mathbf{AB})\mathbf{V}(\mathbf{I} - \mathbf{BA}^T) = \hat{z} \\ \mathbf{1}^T\mathbf{BV}(I - \mathbf{BA}^T) = \underline{\mathfrak{z}}_0^T \end{cases}$$

where $\mathfrak{z}_0 = \underline{\mathfrak{z}}_0^T\mathbf{x}$, $\mathbf{v} = \mathbf{V}(\mathbf{I} - \mathbf{BA}^T)\mathbf{x}$, $\hat{\mathfrak{z}} = z\mathbf{x}$, $\mathbf{1} = (1, 1, \ldots, 1)^T_{n\times 1}$,
Take $\mathbf{x} \in \{(1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1)\}$, $\mathbf{B} = diag(b_1, \ldots, b_n)$, $\mathbf{1B} = (b_1, \ldots, b_n)^T$
Let $P = (\mathbf{I} - \mathbf{AB})$, $P^T = (\mathbf{I} - \mathbf{BA}^T)$
Then,

$$\begin{cases} \mathbf{PVP}^T = \hat{z} \\ \mathbf{b}^T\mathbf{VP}^T = \underline{\mathfrak{z}}_0^T \end{cases}$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{P} \\ \mathbf{b}^T \end{bmatrix}_{(n+1)\times n}, \qquad \hat{\mathbf{Z}}_0 = \begin{bmatrix} \hat{\mathbf{z}} \\ \hat{\mathfrak{z}}^T \end{bmatrix}_{(n+1)\times n}$$

$$\boxed{\mathbf{QVP}^T = \hat{\mathbf{z}}_0}$$

If $\mathbf{Q}$, $\mathbf{P}$were non-singular,then $\mathbf{V} = \mathbf{Q}^{-1}\hat{\mathbf{z}}_0\mathbf{P}^{-1}$

But $\mathbf{B}$,$\mathbf{P}$ are singular because $\mathbf{P}\underline{\mathbf{F}} = (\mathbf{I} - \mathbf{AB})\underline{\mathbf{F}} = 0$.

Hence use G-inverse:

$$\mathbf{V} = \mathbf{Q}^{-}\hat{\mathbf{z}}_0\mathbf{P}^{-}$$

3.3 Simulation for the variance estimator of the bivariate survivor function

1) **The real data simulation results:**

(1). twins

Table 3.1: 3-4 Estimation result of variance of survivor function of Twins, based on Sen method.

Observation	Estimation of Survivor function	variance
(1,4)	1.00000	3.367604e-001
(1, 5)	0.85714	2.216862e-001
(1, 8)	0.68571	-1.502201e-001
(3, 4)	0	-9.850951e-033
(6+, 5)	0	0
(7+, 5)	0	0
(7+, 7+)	0	0

(2). kidney of male

Table 3.2: 3-4 Estimation result of variance of survivor function of pairs of kidney of male

Observation	Estimation of Survivor function	variance
(8 , 16)	0.54287	0.082
(22, 28)	0.22858	0.167
(30, 12)	0.22858	0.081
(7 , 9)	0.77145	0.027
(152,562)	0.11429	0.004
(12, 40)	0.31429	0.0453
(2, 25)	0.54287	0.008
(15 ,154)	0.22858	0.011
(17 , 4+)	0.34287	0
(63 ,8+)	0.11429	0

(3). kidney of female

Table 3.3: Estimation results kidney of female : $Variance$

(a_1, a_2)	$\bar{Fe}$	**variance**
(53, 196)	0.06250	4.851232e-001
(7, 333)	0.08523	4.205529e-001
(96, 38)	0.85227	-3.738503e-001
(536, 25+)	0.00000	-2.611333e-031
(185, 177)	0.00000	-7.456825e-032
(22+, 159+)	0.06250	-7.939887e-032
(13, 66)	0.06250	-6.293170e-031
(132,156)	0.00000	-1.212809e-031
(27, 58)	0.06250	-1.907828e-033
(152, 30)	0.00000	2.419727e-032
(119, 8)	0.00000	-1.336574e-031
(6+, 78)	0.14773	7.223717e-032
(23, 13+)	0.91477	-1.466093e-031
(447, 318)	0.00000	-1.935449e-031
(24, 245)	0.00000	1.571959e-031
(511, 30)	0.00000	-6.241149e-032
(141, 8+)	0.00000	-9.684226e-034

continue

Table 3.4: continue: Estimation results of kidney of female: $Variance$

(a_1, a_2)	$\bar{Fe}$	**variance**
(149+,70+)	0.00000	6.439935e-034
(292, 114)	0.00000	-1.332293e-032
(15, 108+)	0.06250	8.651571e-033
(402, 24+)	0.00000	-2.396189e-032
(39, 46+)	0.06250	-6.249315e-035
(113+, 201)	0.00000	-8.476003e-032
(34, 30)	0.91477	1.484698e-033
(130, 26)	0.00000	-1.475209e-032
(5+, 43)	0.14773	0
(190, 5+)	0.00000	0
(54+, 16+)	0.85227	0

(4).The distribution of X1, X2: COPULA MODEL $\theta = 4$ (ii) Y : $Y_1, Y_2 \sim EXP(200)$, $Y_1 = Y_2$

Table 3.5: Estimation results 3-3-4-a-ii : $Variance$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	**variance**
(0.375, 0.000)	0.6864129	0.6879889	3.244625e-04	0.0188161695
(0.750, 0.000)	0.4684367	0.4703000	2.946298e-05	0.01146865052
(1.125, 0.000)	0.3250774	0.3268889	8.116079e-06	0.01515010932
(1.500, 0.000)	0.2224239	0.2241333	7.701070e-05	0.01515010932
(0.000, 0.375)	0.6864163	0.6879556	3.244696e-04	0.00440535970
(0.375, 0.375)	0.5945767	0.5963444	1.615612e-04	0.00834550537
(0.750, 0.375)	0.4507879	0.4526111	1.943949e-05	0.0024081586
(1.125, 0.375)	0.3178298	0.3197000	1.067533e-05	0.00405958277
(1.500, 0.375)	0.2212427	0.2231333	7.821228e-05	0.0002535174
(0.000, 0.750)	0.4697573	0.4715222	3.029647e-05	0.0002535174
(0.375, 0.750)	0.4486247	0.4505333	1.835378e-05	0.0002535174
(0.750, 0.750)	0.3977959	0.3995556	1.821240e-06	-0.0003887308
(1.125, 0.750)	0.3077842	0.3095889	1.480168e-05	-0.0006089111
(1.500, 0.750)	0.2195278	0.2213222	7.997339e-05	0.0002996643
(0.000, 1.125)	0.3228700	0.3246889	8.858455e-06	-0.0006269530
(0.375, 1.125)	0.3227294	0.3247556	8.906865e-06	-0.0001208961
(0.750, 1.125)	0.3098059	0.3120444	1.391717e-05	-0.0002095059
(1.125, 1.125)	0.2714946	0.2734111	3.531304e-05	0.00052193561
(1.500, 1.125)	0.2102351	0.2119111	8.985703e-05	0.0012557191
(0.000, 1.500)	0.2221120	0.2237778	7.732708e-05	0.003220793
(0.375, 1.500)	0.2228548	0.2247889	7.657471e-05	0.0029365223
(0.750, 1.500)	0.2193650	0.2210889	8.014152e-05	3.642526e-003
(1.125, 1.500)	0.2118530	0.2137556	8.809491E-05	0.00254675388
(1.500, 1.500)	0.1871494	0.1887333	1.169026e-04	0.0041182479

(5).The distribution of X1, X2: COPULA MODEL $\theta = 4$ (3-3-4-a-iii): Y : $Y_1 \sim EXP(200)$, Y_1, Y_2 are i.i.d.

Table 3.6: Estimation results 3-3-4-a-iii : $Variance$

(a_1, a_2)	$\bar{F}e$	$\hat{\bar{F}}$	$mse\bar{F}e$	**variance**
(0.000, 0.000)	0.9997784	1.0000000	1.306597e-03	0.0189335023
(0.375, 0.000)	0.6859239	0.6883111	3.249541e-04	0.00096608637
(0.750, 0.000)	0.4715727	0.4752556	3.193285e-05	0.0195283675
(1.500, 0.000)	0.3212743	0.3250556	9.160115e-06	0.01862049929
(0.000, 0.375)	0.2181039	0.2213000	8.069634e-05	0.00899143641
(0.375, 0.375)	0.6803728	0.6830667	3.135022e-04	0.00635332909
(0.750, 0.375)	0.5937038	0.5968444	1.613447e-04	0.0019954300
(1.125, 0.375)	0.3160567	0.3196000	1.107430e-05	0.00704323382
(1.500, 0.375)	0.2200086	0.2234667	7.873276e-05	0.00062036830
(0.000, 0.750)	0.4671715	0.4707333	2.912554e-05	0.0063661602
(0.375, 0.750)	0.4501753	0.4537889	1.949694e-05	-0.0003154168
(0.750, 0.750)	0.3996358	0.4029778	2.242911e-06	0.0008761438
(1.125, 0.750)	0.3079297	0.3113444	1.441737e-05	0.00127697827
(1.500, 0.750)	0.2146828	0.2179889	8.428403e-05	-0.0004741107
(0.000, 1.125)	0.3178020	0.3215556	1.041381e-05	-0.0007804777
(0.375, 1.125)	0.3190518	0.3228889	9.953295e-06	-0.0000768104
(0.750, 1.125)	0.3002254	0.3036778	1.799310e-05	-0.0004292545
(1.125, 1.125)	0.2707833	0.2743111	3.530341e-05	-0.0005504304
(1.500, 1.125)	0.2099726	0.2129667	8.935121e-05	0.0009924766
(0.000, 1.500)	0.2195676	0.2229111	7.918518e-05	0.00009573828
(0.375, 1.500)	0.2185869	0.2218000	8.019608e-05	0.0009924766
(0.750, 1.500)	0.2156402	0.2190222	8.327215e-05	0.0019814793
(1.125, 1.500)	0.2084467	0.2119111	9.102444e-05	0.0029622123
(1.500, 1.500)	0.1842161	0.1875111	1.196754e-04	0.0046999643

Chapter 4

Conclusion

My thesis work simulated the bi-variate Kaplan Meier estimator derived by Sen and Stute (2007) by using different joint distribution of (X_1, X_2) and real data.

From all simulation results, the estimator of bivariate survivor function (Sen and Stute(2007)) is efficient to estimate the survivor function. It gives nonnegative masses. Using this estimator is easily graph the trend of the survivor function which is very useful in applied field. Comparing with the other estimators, we have the best estimator.

The variance estimator has a good form. But the variance has negative values for some points.

Chapter 5

Further study

1. *Confidence interval.* In the variance estimator, we have some negative variance which needs to be corrected. We then plan to use the corrected variance estimator to compute confidence intervals for survival as well as interval probabilities, i.e.,

$$Pr(X_1 \in (a_1, b_1), X_2 \in (a_2, b_2)) = \bar{F}(b_1, b_2) - \bar{F}(a_1, b_2) - \bar{F}(b_1, a_2) + \bar{F}(a_1, a_2).$$

2. *Model selection.* The bivariate survivor function estimator $\bar{F}_{\mathrm{e}}(\cdot)$ could be used for goodness-of-fit tests and other model-checks by comparing it to a given parameterized family of survivor functions $\{\bar{F}_\theta(\cdot), \theta \in \Theta\}$, such as a copula model. However, we need to develop appropriate methods.

3. *Regression.* The bivariate point-masses $(p_1, \ldots, p_n)$ obtained in Chapter 2 could be used for regression estimation. For instance, to estimate a linear regression model of the form $E(X_2|X_1) = \beta_0 + \beta_1 X_1$, we will have to solve $\min_{b_0, b_1} \sum_{i=1}^{n} p_i[Z_{2i} - b_0 - b_1 Z_{1i}]^2$. Performance of the resulting estimators will then have to be studied.

Bibliography

[1] Collett, D. *Modelling Survival Data In Medical Research*, (Second Edition) *Chapman & Hall/CRC*, (1993).

[2] Dabrowska, D.M.(1988). *Kaplan-Meier Estimate on the Plane. The Annals of Statistics.* Vol. 16, No. 4, pp. 1475-1489.

[3] Hougaard, P. *Analysis of multivariate survival data.* Springer,New York (2000).

[4] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

[5] Prentice, R.L., Moodie, F.Z. and Wu, J.(2004). *Hazard-based nonparametric survivor function estimation. J.R.Statist. Soc.B*, 66, Part 2, pp.305-319.

[6] Sen, A. and Stute, W.(2007) *A bi-variate Kaplan-Meier estimator small via an integral eqwtion.* Technical Report # 03/07, Dept. of Math. and Stat.,Concordia University

[7] Smith, P.J. *Analysis Of Failure And Survival Data. Chapman & Hall/CRC*, (2002).

[8] van der Laan, M. J.(1996). *Efficient estimtion inthe bivariate censoring model and repairing NPMLE. The Annals of Statistics*, Vol. 24, No. 2. pp. 596-627.

[9] Zuo, Y., Cui, H. and Young, D. (2004). *Influence Function And Maximum Bias Of Projection Depth Based Estimators. The Annals of Statistics*, Vol. 32, No. 1, 189 - 218.