May 2022

Nonparametric Bivariate Distribution Estimation under Right Censoring using Poisson Polynomials

Luchen Liu
liu85@uwindsor.ca

Nonparametric Bivariate Distribution Estimation under Right Censoring using Poisson Polynomials

by

Luchen Liu

A Major Research Paper
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

Nonparametric Bivariate Distribution Estimation under Right Censoring using Poisson Polynomials

by

Luchen Liu

APPROVED BY:

M. Hlynka
Department of Mathematics and Statistics

M. Belalia, Advisor
Department of Mathematics and Statistics

Apr 29, 2022

Author's Declaration of Originality

I hereby certify that I am the sole author of this major paper and that no part of this major paper has been published or submitted for publication.

I certify that, to the best of my knowledge, my major paper does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my major paper, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my major paper and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my major paper, including any final revisions, as approved by my major paper committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

In this major paper, a nonparametric estimator of the joint cumulative distribution function using Poisson probability under right censoring was proposed and discussed. In particular, the joint cumulative distribution of two non-negative random variables X and Y was of interest in this work, where X was assumed to be complete, while Y was subjected to right censoring. The asymptotic properties of the proposed estimator (called hereafter Poisson polynomial estimator) were established, including independent and identically distributed representation, expectation and variance and asymptotic normality.

Furthermore, two real datasets were analyzed to assess the Poisson polynomial estimator: Loss-ALAE joint distribution estimation and age's impact on survival in patients with colon cancer. Also, the Stute empirical estimator and empirical Bernstein estimator were introduced as a comparison. Evidence reveals that the Poisson estimator gives a more precise value along with the best effect on smoothness among all three estimators and provides an alternative solution for future simulation and prediction. However, more calculations are required to obtain a precise solution with an acceptable error as the Poisson probability takes values from zero to infinity.

Acknowledgments

First, I would like to appreciate my supervisor, Dr. Belalia, for his support in my master's life. My study and major paper research went smoothly under his supervision. His academic knowledge and personalities of kindness, carefulness, and patience helped me facilitate the study in statistics. Also, I am thankful to all graduate committee members. In particular, I would like to thank Dr. Hlynka and Dr. Hussein for reading my major paper and offering suggestions to become better in the future.

I am much appreciated for Mrs. Kate Hargreaves' support and graduate administration work. Without her help, I could not finish the major paper in an appropriate manner.

Most importantly, I would like to express my appreciation to my family and friends. Their understanding and accompanying are essential for my success. I sincerely thank them for all efforts and support in my master's life.

Contents

List of Figures

Chapter 1

Introduction

In multivariate analysis, the joint cumulative distribution function plays a vital role. However, this function is not available in practice. Consequently, one needs to find a statistical approach to be estimate it from underlying data sets.

To begin our discussion, let recall the mathematical definition of joint cumulative distribution function in the bivariate cases.

Definition 1.1 (Joint Cumulative Distribution Function)**.** *Suppose X and Y are two real-valued random variables. The joint cumulative distribution function (cdf)* $\mathrm{F}(x, y)$ *is the probability that X will take a value less than or equal to x and Y will take a value less than or equal to y, i.e.,*

$$\mathrm{F}(x, y) = \mathbb{P}\left(X \leq x, Y \leq y\right). \tag{1.1}$$

Data completeness is crucial for a researcher to make a high quality estimation of a joint cdf. In reality, collected data is categorized as complete and incomplete data. Data completeness refers to the comprehensiveness or wholeness of the data. There should be no gaps or missing information for data to be truly complete.

For the complete data, the means of estimating joint cdf have been explored intensively over the decades. In summary, these methods can be classified as two categories: (i) parametric approach, when the underlying cdf F_θ is assumed to be belonging to a theoretical distribution family indexed by a parameter $\theta \in \mathbb{R}^q$. (ii) The non-parametric

approaches have been utilized more frequently as they require fewer assumptions and can avoid the potential bias from inappropriate parametric models. In nonparametric estimation, the multivariate empirical cumulative distribution function (ecdf) is the most widely used non-parametric tool. Given a sample $(X_1, \ldots, X_n)$ drawn from F, the ecdf is defined as follow:

$$\mathrm{F}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i \leq x, Y_i \leq y\right). \tag{1.2}$$

However, since the multivariate empirical function is a discontinuous estimator with derivative equal to zero almost everywhere, researchers have dedicated to smoothing empirical function by kernels, such as methods developed by Silverman (1986), Wand and Jones (1995), and Hanif et al. (2018).

Even though a high demanding in many fields like survival analysis and actuarial science, literature dealing with incomplete data set is still scarce. Practitioners can only rely on the incomplete data set to make the decision. For example, the observation consists of a non-negative random vector (X, Y), where X is complete and Y is subjected to right censoring (incomplete), one of a major characteristics in survival analysis. This may become an issue when the starting or ending events are not precisely observed. Right censoring is the most common type, which refers to situations that the final endpoint is only known to exceed a particular value. Mathematically, we can say that one can only observe a random vector (X, T, δ) instead of (X, Y), where T is the minimum between the true variable Y and a censoring variable. We denote the censoring variable by C with cumulative distribution function G, and $\delta = \mathbb{I}_{\{Y<C\}}$ with $\mathbb{I}_A$ as the indicator function of the set A.

An example from Moore (2015) is introduced below as an illustration of right censoring. Figure 1.1 presents the data of six patients from a hypothetical clinical trial. They were followed over a 2.5 year accrual period ($2000/1/1 - 2002/6/30$) and 4.5 years of additional follow-up time lasted until 2007/12/31. The $\times$'s denote deaths while the open circles denote censoring events. The data were meant to be analyzed on 2007/12/31, but three patients (Patients 1, 3 and 4) were still alive. Also shown in this example is the ultimate fate of these three survival patients whose conditions would not have been known at the

time of analysis. Thus, for these three patients, we have incomplete information about their survival time. For example, we know that Patient 1 survived at least 7 years, but as of the end of 2007 it would not have been known how long the patient would ultimately live. Therefore, these data of survival time were subject to right censoring.

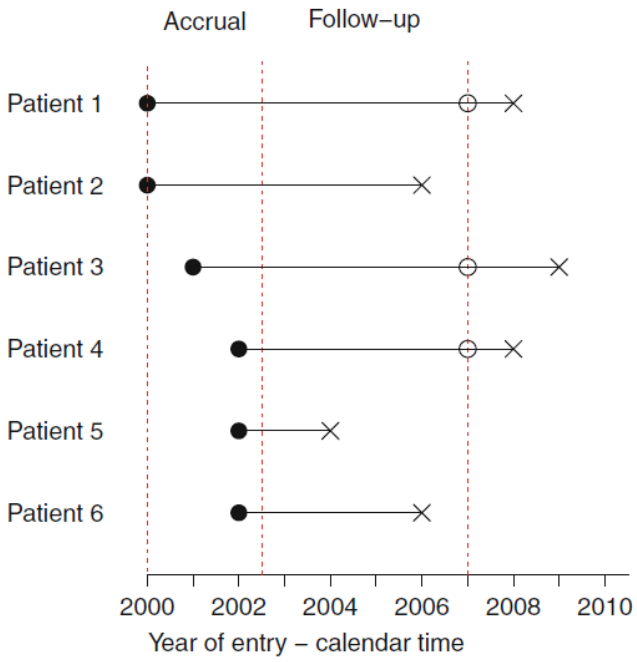


Figure 1.1: Clinical trial accrual and follow-up periods

To overcome the limitations above, researches have focused on nonparametric methods using Bernstein polynomials. Initially, Sergei Bernstein's idea was derived from the demonstration of Weierstrass approximation theorem, which is stated below.

Theorem 1.1 (Weierstrass Approximation Theorem(1885))**.** *Let* $f : [a, b] \to \mathbb{R}$. *Then there is a sequence of polynomials* $g_m(x)$ *that converges uniformly to* $f(x)$ *on* $[a, b]$. *i.e., Given* $\epsilon > 0$, *there exists a sequence of polynomials* $g_m(x)$ *such that*

$$|f(x) - g_m(x)| < \epsilon, \qquad \forall x \in [a, b].$$

Therefore, to approximate the function $f(x)$ on the closed and bounded interval $[0, 1]$, Bernstein (1912) proposed an alternative probabilistic method to Theorem 1.1. His ap-

proach was based on the so called Bernstein polynomials, defined below

Definition 1.2 (Bernstein polynomials)**.** *For $m \in \mathbb{N}$ and $0 \leq k \leq m$, the Bernstein polynomials $P_{m,k}$ of degree m are defined as*

$$P_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}, \quad k = 0, 1, 2, \ldots, m.$$

for $x \in [0, 1]$.

Based on the above definition, one can rewrite the Weierstrass approximation theorem in the form of Bernstein polynomials.

Theorem 1.2 (Bernstein Theorem)**.** *Let $f : [0, 1] \to \mathbb{R}$ be a continuous real-functions. The Bernstein polynomials of order m associate to f are give by :*

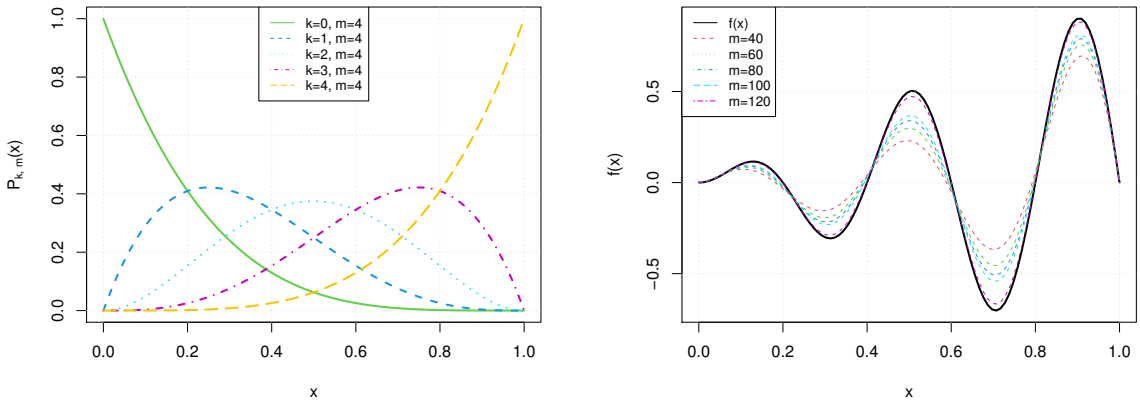
$$\forall m \in \mathbb{N},\ \forall x \in [0, 1],\ B_m(f)(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) \binom{m}{k} x^k \left(1-x\right)^{m-k}.$$

Then, we have

$$\lim_{m\to\infty} \|f - B_m(f)\|_\infty = \lim_{m\to\infty} \sup_{x\in[0,1]} |f(x) - B_m(f)(x)| = 0.$$

In particular, any continuous function on $[0, 1]$ is the uniform limit of a sequence of Bernstein polynomials.

As an illustration of the usefulness of the above theorem, Bernstein polynomials and Bernstein approximation of the function $f(x) = x \sin(5\pi x)$ are depicted in Figures 1.2(a) and 1.2(b), receptively. One can see that the approximation becomes more as the polynomial degree m increases.



(a) Bernstein polynomials basis

(b) Bernstein Approximation

Figure 1.2: (a) Bernstein polynomials, (b) Approximation of function $f(x) = x\sin(5\pi x)$ using Bernstein polynomials of degree $m = 40, 60, 80, 100, 500$.

Extending the univariate Bernstein polynomials approximation to the bivariate case, the distribution function $\mathrm{F}(x, y)$, being continuous on $[0, 1]^2$ can be approximated by Bernstein polynomials in the following way:

$$\mathrm{F}_m(x, y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} \mathrm{F}\left(\frac{k}{m}, \frac{\ell}{m}\right) \mathrm{P}_{k,m}(x)\mathrm{P}_{\ell,m}(y), \quad x, y \in [0, 1]^2 \tag{1.3}$$

where integer m is the smoothing parameter and $P_{j,m}(z) = \binom{m}{j} z^j(1-z)^{m-j}$ for $z \in [0, 1]$, is binomial probability.

It has been proved by Babu and Chaubey (2006) that F_m converges uniformly to F in $[0, 1]^2$ as m goes to infinity. For complete data, the Bernstein estimator of the joint cdf F is obtained by replacing $\mathrm{F}\left(\frac{k}{m}, \frac{\ell}{m}\right)$ by $\mathrm{F}_n\left(\frac{k}{m}, \frac{\ell}{m}\right)$ in (1.3), where F_n is the empirical distribution function defined in (1.2)

Furthermore, Babu and Chaubey (2006) proposed the Bernstein estimator of a distribution function F on the cube $[0, 1]$, which is described as

$$\widehat{\mathrm{F}}^{u}_{m,n}(x, y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} \mathrm{F}_n\left(\frac{k}{m}, \frac{\ell}{m}\right) \mathrm{P}_{k,m}(x)\mathrm{P}_{\ell,m}(y), \quad x, y \in [0, 1] \tag{1.4}$$

Note that $\hat{\mathrm{F}}_{m,n}$ is a polynomial in x and y, hence it has all derivatives. Moreover it was shown in Babu and Chaubey (2006) that $\hat{\mathrm{F}}_{m,n}$ is a proper distribution function. The asymptotic properties of $\hat{\mathrm{F}}_{m,n}$ were studied in Belalia (2016). This nonparametric estimation approach has solved the problem of multivariate empirical function that makes derivative impossible. However, it is only utilized for complete data. Later, more researches were conducted to involve incomplete data observation using Bernstein polynomials. First of all, based on a sample $\{(X_i, T_i, \delta_i)\}_{i=1}^n$, Stute (1993) has proposed the empirical estimator of the joint cdf in dealing with right censored data, denoted by $\widehat{\mathrm{F}}_n$:

$$\widehat{\mathrm{F}}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} W_{in}\mathbb{I}_{\{X_i\leq x, T_i\leq y\}}, \tag{1.5}$$

where $W_{in} = \frac{\delta_i}{1-G_n(T_i^-)}$. Particularly, we can notice that it is a step function with derivative equal to zero almost everywhere. In order to build a smooth version of $\widehat{\mathrm{F}}_n$ using Bernstein polynomials, Dib et al. (2021) proposed an empirical Bernstein estimator using binomial polynomials. The estimator is defined as follows:

$$\widehat{\mathrm{F}}^b_{m,n}(x,y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m}\widehat{\mathrm{F}}_n\left(\frac{k}{m},\frac{\tau\ell}{m}\right)\mathrm{P}_{k,m}(x)\mathrm{P}^{\tau}_{\ell,m}(y), \tag{1.6}$$

where $P^{\tau}_{\ell,m}(y) = P_{\ell,m}(y/\tau)$ and τ is defined such that $1-L(\tau) > 0$. Here L is the cumulative function of T. In practice, τ can be replaced by $T_{(n)}$, i.e., the maximum of the sample $\{T_i\}_{i=1}^n$.

Furthermore, Dib et al. (2021) also pointed out that Poisson distribution function can be utilized as a type of substitution to the binomial distribution function in (1.6). Specifically, we can estimate the bivariate joint cumulative distribution function for incomplete data by

$$\widehat{\mathrm{F}}_{m,n}(x,y) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\widehat{\mathrm{F}}_n\left(\frac{k}{m},\frac{\tau\ell}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}^{\tau}_{\ell,m}(y). \tag{1.7}$$

where $\text{Pois}_{k,m}(z)$ is the Poisson probability mass function with parameter mz, namely:

$$\text{Pois}_{k,m}(z) = \exp(-mz)\frac{(mz)^k}{k!}, \qquad k = 0, 1, 2, ...$$

and $\text{Pois}^{\tau}_{k,m}(y) = \text{Pois}_{\ell,m}(y/\tau)$.

In statistics, a discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$, if it has a probability mass function given by:

$$f(k, \lambda) = \mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, ...$$

where k is the number of occurrences, e is Euler's number ($e = 2.71828$). Figure 1.3 presents the Poisson probability mass function with parameters $\lambda = 1$, $\lambda = 4$, and $\lambda = 10$, respectively.

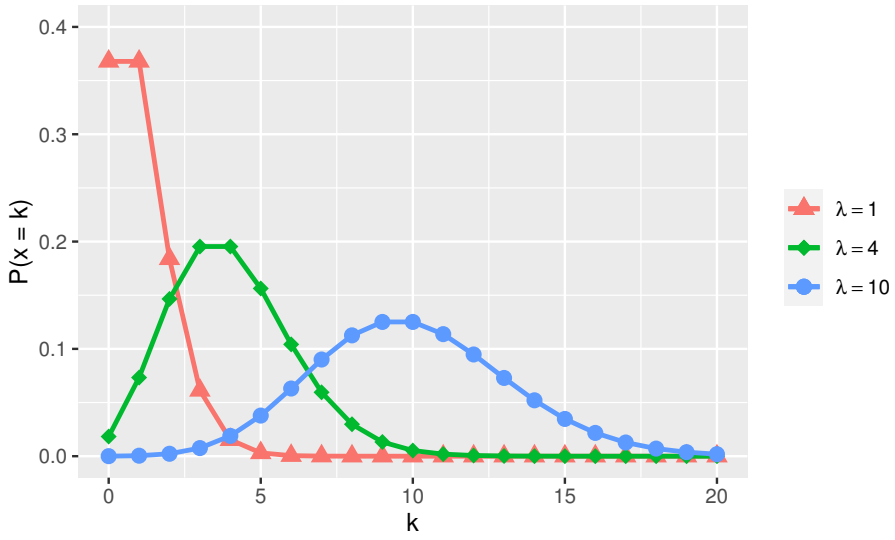


Figure 1.3: Poisson probability mass function with parameters $\lambda = 1, 4, 10$

The focus of this major paper is to study and discuss the Poisson polynomial estimator (1.7). In Chapter 2, the asymptotic properties of the Poisson polynomial estimator (1.7) were established with detailed theoretical proof, including independent and identically distributed representation, asymptotic bias, variance of the Poisson polynomial estimator,

and asymptotic normality. In Chapter 3, two real data analysis were carried out to assess the Poisson estimator: Loss-ALAE joint distribution estimation and age's impact on survival in patients with colon cancer. Furthermore, the results from the previous Stute empirical estimator (1.5) and empirical Bernstein estimator (1.6) were included and served as comparisons.

Chapter 2

Properties of the Poisson Polynomial Estimator

In this chapter, four major properties were established with mathematical proof, which included independent and identically distributed representation in Section 2.1, asymptotic bias and variance of the Poisson polynomial estimator in Section 2.2, and asymptotic normality in Section 2.3. The theoretical research laid a robust foundation and support for future application and real data analysis.

2.1 Independent and Identically Distributed Representation

First of all, the property of independent and identically distributed (i.i.d.) representation is presented. The i.i.d. conclusion helps the further calculation in asymptotic bias and variance, and the proposal of asymptotic normality.

Proposition 1. *Assume that* F *and* G *are continuous. Then, we have the following asymptotic i.i.d. representation of* $\widehat{\mathrm{F}}_{m,n}$ *in* (1.7)

$$\widehat{\mathrm{F}}_{m,n}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\eta_i(x,y) + O_{a.s.}\left(\frac{\log n}{n}\right) \tag{2.1}$$

where

$$\eta_i(x,y) = W_i \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} Pois_{k,m}(x) Pois_{\ell,m}(y/\tau)$$

with $W_i = \frac{\delta_i}{1-G(T_i)}$.

Proof of Proposition 1. Recall the Stute empirical estimator that

$$\widehat{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} W_{in} \mathbb{I}_{\{X_i \le x, T_i \le y\}}. \tag{2.2}$$

It can be referred to Major and Rejto. (1988) that $W_{in} = W_i + R_{i,n}$, where $W_i = \frac{\delta_i}{1-G(T_i)}$ and $R_{i,n} = O_{a.s.}(n^{-1}\log n)$. Further, one can verify that $\mathbb{E}\left(R_{i,n}\right) = O\left(n^{-1}\right)$. Therefore,

$$\begin{aligned}
\widehat{\mathrm{F}}_{m,n}(x,y) &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \widehat{\mathrm{F}}_n\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau) \\
&= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \frac{1}{n}\sum_{i=1}^{n} W_{in} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau) \\
&= \frac{1}{n}\sum_{i=1}^{n} W_{in} \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau) \\
&= \frac{1}{n}\sum_{i=1}^{n} W_i \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau) + O_{a.s.}(n^{-1}\log n) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{1-G(T_i)} \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau) + O_{a.s.}(n^{-1}\log n).
\end{aligned}$$

To sum up, it can be obtained that

$$\widehat{\mathrm{F}}_{m,n}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \eta_i(x,y) + O_{a.s.}\left(\frac{\log n}{n}\right) \tag{2.3}$$

where

$$\eta_i(x,y) = W_i \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau).$$

□

2.2 The Asymptotic Bias and Variance of the Poisson Polynomial Estimator

Based on the result from Proposition 1, expressions of the asymptotic bias and variance of the Poisson polynomial estimator were established and stated in Proposition 2 and 3 respectively.

Assumption 1. *All the partial derivatives up to second order, denoted by* $\mathrm{F}_x = \frac{\partial \mathrm{F}(x,y)}{\partial x}$, $\mathrm{F}_y = \frac{\partial \mathrm{F}(x,y)}{\partial y}$, $\mathrm{F}_{xx} = \frac{\partial^2 \mathrm{F}(x,y)}{\partial x^2}$, $\mathrm{F}_{yy} = \frac{\partial^2 \mathrm{F}(x,y)}{\partial y^2}$ *and* $\mathrm{F}_{xy} = \frac{\partial^2 \mathrm{F}(x,y)}{\partial x \partial y}$ *are continuous on* $[0, +\infty)^2$.

Proposition 2. *Under Assumption 1, for any* $x > 0$ *and* $y > 0$, *we have*

$$\mathbb{E}\left[\widehat{\mathrm{F}}_{m,n}(x,y)\right] = \mathrm{F}(x,y) + \frac{1}{2m}\mathrm{F}_{xx}(x,y)x + \frac{1}{2m}\mathrm{F}_{yy}(x,y)\tau y + o\left(m^{-1}\right) + O\left(n^{-1}\right). \quad (2.4)$$

Proof of Proposition 2. From the result in Section 2.1, we have

$$\widehat{\mathrm{F}}_{m,n}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \eta_i(x,y) + R_{i,n}. \quad (2.5)$$

Thus,

$$\mathbb{E}\left[\widehat{\mathrm{F}}_{m,n}(x,y)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\eta_i(x,y)\right] + \mathbb{E}(R_{i,n}). \quad (2.6)$$

Since it has been proved that the n samples are identically and independently distributed in Section 2.1 and $\mathbb{E}(R_{i,n}) = O(n^{-1})$, we have

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathrm{F}}_{m,n}(x,y)\right] &= \frac{1}{n} \times n \times \mathbb{E}\left[\eta_i(x,y)\right] + \mathbb{E}(R_{i,n}) \\ &= \mathbb{E}\left[\eta_i(x,y)\right] + O(n^{-1}). \end{aligned} \quad (2.7)$$

Now calculate $\mathbb{E}\left[\eta_i(x,y)\right]$:

$$\begin{aligned}\mathbb{E}\left[\eta_i(x,y)\right] &= \mathbb{E}\left[W_i \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}} \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\right] \\ &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathbb{E}\left(W_i \mathbb{I}_{\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\}}\right) \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\end{aligned}$$

$$= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \mathrm{F}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) \mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau). \tag{2.8}$$

We define (2.8) as $\mathrm{F}_m^{\tau}(x,y)$. The fact that F is a twice differentiable function on $[0,+\infty)^2$, and using Taylor expansion, we get

$$\begin{aligned}\mathrm{F}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) = \mathrm{F}(x,y) &+ \mathrm{F}_x(x,y)\left(\frac{k}{m}-x\right) + \mathrm{F}_y(x,y)\left(\frac{\tau\ell}{m}-y\right) + \frac{1}{2}\mathrm{F}_{xx}(x,y)\left(\frac{k}{m}-x\right)^2 \\ &+ \frac{1}{2}\mathrm{F}_{yy}(x,y)\left(\frac{\tau\ell}{m}-y\right)^2 + \mathrm{F}_{xy}(x,y)\left(\frac{k}{m}-x\right)\left(\frac{\tau\ell}{m}-y\right) \\ &+ o\left(\left(\frac{k}{m}-x\right)^2 + \left(\frac{\tau\ell}{m}-y\right)^2\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\eta_i(x,y)\right] = \mathrm{F}(x,y) &+ \mathrm{F}_x(x,y)\sum_{k=0}^{\infty}\left(\frac{k}{m}-x\right)\mathrm{Pois}_{k,m}(x) + \mathrm{F}_y(x,y)\sum_{\ell=0}^{\infty}\left(\frac{\tau\ell}{m}-y\right)\mathrm{Pois}_{\ell,m}(y/\tau) \\ &+ \frac{1}{2}\mathrm{F}_{xx}(x,y)\sum_{k=0}^{\infty}\left(\frac{k}{m}-x\right)^2\mathrm{Pois}_{k,m}(x) + \frac{1}{2}\mathrm{F}_{yy}(x,y)\sum_{\ell=0}^{\infty}\left(\frac{\tau\ell}{m}-y\right)^2\mathrm{Pois}_{\ell,m}(y/\tau) \\ &+ o\left(\left(\frac{k}{m}-x\right)^2 + \left(\frac{\tau\ell}{m}-y\right)^2\right).\end{aligned} \tag{2.9}$$

If we define U to be a random variable having the Poisson distribution with parameter mx, that is,

$$\mathbb{P}\left(U=k\right) = \frac{(mx)^k}{k!}\exp(-mx).$$

Then the calculation is illustrated as follows,

$$\begin{aligned}\mathbb{E}\left(\frac{U}{m}-x\right) &= \sum_{k=0}^{\infty}\left(\frac{k}{m}-x\right)\mathrm{Pois}_{k,m}(x) \\ &= \sum_{k=0}^{\infty}\left(\frac{k}{m}-x\right)\exp(-mx)\frac{(mx)^k}{k!} \\ &= \sum_{k=0}^{\infty}\frac{k-mx}{m}\exp(-mx)\frac{(mx)^k}{k!} \\ &= \frac{1}{m}\left(\sum_{k=0}^{\infty}k\exp(-mx)\frac{(mx)^k}{k!} - mx\sum_{k=0}^{\infty}\exp(-mx)\frac{(mx)^k}{k!}\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m}(mx - mx) \\
&= 0.
\end{aligned}
$$

Also,

$$
\begin{aligned}
\mathbb{E}\left(\frac{U}{m} - x\right)^2 &= \sum_{k=0}^{\infty}\left(\frac{k}{m} - x\right)^2 \exp(-mx)\frac{(mx)^k}{k!} \\
&= \sum_{k=0}^{\infty}\frac{k^2 + m^2x^2 - 2kmx}{m^2}\exp(-mx)\frac{(mx)^k}{k!} \\
&= \frac{1}{m^2}\sum_{k=0}^{\infty}\left(k^2\exp(-mx)\frac{(mx)^k}{k!} + m^2x^2\exp(-mx)\frac{(mx)^k}{k!} - 2kmx\exp(-mx)\frac{(mx)^k}{k!}\right) \\
&= \frac{1}{m^2}\left((mx)^2 + mx + m^2x^2 - 2m^2x^2\right) \\
&= \frac{x}{m}.
\end{aligned}
$$

Similarly, if we define V as a random variable that follows the Poisson distribution with parameter $m(y/\tau)$, one can derive the following result

$$
\begin{aligned}
\mathbb{E}\left(\frac{\tau V}{m} - y\right) &= \sum_{\ell=0}^{\infty}\left(\frac{\tau\ell}{m} - y\right)\mathrm{Pois}_{\ell,m}(y/\tau) \\
&= \sum_{\ell=0}^{\infty}\frac{\tau\ell - my}{m}\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!} \\
&= \frac{1}{m}\left(\sum_{\ell=0}^{\infty}(\tau\ell)\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!} - (my)\sum_{\ell=0}^{\infty}\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!}\right) \\
&= \frac{1}{m}\left(\tau\frac{my}{\tau} - my\right) \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left(\frac{\tau V}{m} - y\right)^2 &= \sum_{\ell=0}^{\infty}\left(\frac{\tau\ell}{m} - y\right)^2\mathrm{Pois}_{\ell,m}(y/\tau) \\
&= \sum_{\ell=0}^{\infty}\frac{(\tau\ell)^2 + m^2y^2 - 2\tau\ell my}{m^2}\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{m^2}\sum_{\ell=0}^{\infty}(\tau\ell)^2\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!} + \frac{1}{m^2}\sum_{\ell=0}^{\infty}m^2y^2\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!} \\
&\quad - \frac{1}{m^2}\sum_{\ell=0}^{\infty}2\tau\ell my\exp(-my/\tau)\frac{(my/\tau)^\ell}{\ell!} \\
&= \frac{1}{m^2}\left[\tau^2\left[(my/\tau)^2 + my/\tau\right] + m^2y^2 - 2m^2y^2\right] \\
&= \frac{\tau y}{m}.
\end{aligned}
$$

Therefore,

$$
\mathbb{E}\left[\eta_i(x,y)\right] = \mathrm{F}(x,y) + \frac{1}{2}\mathrm{F}_{xx}(x,y)\frac{x}{m} + \frac{1}{2}\mathrm{F}_{yy}(x,y)\frac{\tau y}{m} + o(m^{-1}). \tag{2.10}
$$

Finally,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{F}}_{m,n}(x,y)\right] &= \mathbb{E}\left[\eta_i(x,y)\right] + O(n^{-1}) \\
&= \mathrm{F}(x,y) + \frac{1}{2}\mathrm{F}_{xx}(x,y)\frac{x}{m} + \frac{1}{2}\mathrm{F}_{yy}(x,y)\frac{\tau y}{m} + o(m^{-1}) + O(n^{-1}),
\end{aligned} \tag{2.11}
$$

which completes the proof. □

To find the asymptotic variance, we introduce the following quantities:

- $I_m(z) = \sum_{k=0}^{\infty}\left|\frac{k}{m} - z\right|\mathrm{Pois}^2_{k,m}(z)$,
- and for $j = 0, 1, 2$,

$$
R_{j,m}(x) = m^{-j}\sum_{0\le k<\ell}\sum(k - mx)^j\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{\ell,m}(x).
$$

Then, the asymptotic variance is provided in the following proposition

Proposition 3. *Under Assumption 1 and suppose that $m \to +\infty$ such that $nm^{1/2} \to +\infty$ as n tends to $+\infty$. Then, for any $x > 0$ and $y > 0$, we have,*

$$
\mathbb{V}\mathrm{ar}\left[\widehat{\mathrm{F}}_{m,n}(x,y)\right] = n^{-1}\sigma^2(x,y) + n^{-1}V(x,y) + o\left(n^{-1}m^{-1/2}\right) \tag{2.12}
$$

where

$$\sigma^2(x,y) = H(x,y) - \left(\mathrm{F}_m^{\tau}(x,y)\right)^2, \quad H(x,y) = \int_0^x \int_0^y \frac{dF(t,s)}{1-G(s)}$$

and

$$V(x,y) = H_x(x,y)\left(O\left(I_m(x)\right) + 2R_{1,m}(x)\right) + \tau H_y(x,y)\left(O\left(I_m(y/\tau)\right) + 2R_{1,m}(y/\tau)\right).$$

Proof of Proposition 3. First, it can be written that

$$\mathbb{E}\left[\eta_i(x,y)^2\right] = \sum_{k=0}^{\infty}\sum_{k'=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0}^{\infty}\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\le\frac{k}{m},T_i\le\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\le\frac{k'}{m},T_i\le\frac{\tau\ell'}{m}\right\}}\right]\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x) \times\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau).$$

Let's start by calculating $\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\le\frac{k}{m},T_i\le\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\le\frac{k'}{m},T_i\le\frac{\tau\ell'}{m}\right\}}\right]$.

$$\begin{aligned}\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\le\frac{k}{m},T_i\le\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\le\frac{k'}{m},T_i\le\frac{\tau\ell'}{m}\right\}}\right] &= \mathbb{E}\left[\frac{\delta_i}{\left(1-G(T_i)\right)^2}\mathbb{I}_{\left\{X_i\le\frac{k\wedge k'}{m},T_i\le\tau\frac{\ell\wedge\ell'}{m}\right\}}\right]\\ &= \mathbb{E}\left[\frac{\mathbb{I}_{\{Y_i\le C_i\}}}{\left(1-G(Y_i)\right)^2}\mathbb{I}_{\left\{X_i\le\frac{k\wedge k'}{m},Y_i\le\tau\frac{\ell\wedge\ell'}{m}\right\}}\right]\end{aligned}$$

With the property of conditional expectation, we have

$$\mathbb{E}(X) = \mathbb{E}\left(\mathbb{E}(X|Y)\right) = \mathbb{E}_Y\left(\mathbb{E}_X(X|Y)\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\le\frac{k}{m},T_i\le\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\le\frac{k'}{m},T_i\le\frac{\tau\ell'}{m}\right\}}\right] &= \mathbb{E}_{X_i,Y_i}\left[\mathbb{E}_{C_i}\left(\frac{\mathbb{I}_{\{Y_i\le C_i\}}}{\left(1-G(Y_i)\right)^2}\mathbb{I}_{\left\{X_i\le\frac{k\wedge k'}{m},Y_i\le\tau\frac{\ell\wedge\ell'}{m}\right\}}\middle| X_i,Y_i\right)\right]\\ &= \mathbb{E}_{X_i,Y_i}\left[\frac{1}{\left(1-G(Y_i)\right)^2}\mathbb{I}_{\left\{X_i\le\frac{k\wedge k'}{m},Y_i\le\tau\frac{\ell\wedge\ell'}{m}\right\}}\mathbb{E}\left(\mathbb{I}_{\{Y_i\le C_i\}}|X_i,Y_i\right)\right]\\ &= \mathbb{E}\left[\frac{\mathbb{I}_{\left\{X_i\le\frac{k\wedge k'}{m},Y_i\le\tau\frac{\ell\wedge\ell'}{m}\right\}}}{1-G(Y_i)}\right]\\ &= \int_0^{\frac{k\wedge k'}{m}}\int_0^{\tau\frac{\ell\wedge\ell'}{m}}\frac{dF(t,s)}{1-G(s)}.\end{aligned}$$

We define this result as $H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)$. Then

$$
\begin{aligned}
&\mathbb{E}\left[\eta_i(x,y)^2\right] \\
=&\sum_{k=0}^{\infty}\sum_{k'=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0}^{\infty} H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau) \\
=&\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} H\left(\frac{k}{m}, \tau\frac{\ell}{m}\right)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau) \\
&\quad+\sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau) \\
&\quad+\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} H\left(\frac{k}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau) \\
&\quad+\sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty} H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau) \\
=&I_1+I_2+I_3+I_4. \qquad (2.13)
\end{aligned}
$$

Now apply Taylor expansion of $H\left(\frac{k}{m}, \frac{\tau\ell}{m}\right)$ around (x,y), it can be obtained that

$$H\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) = H(x,y) + \left(\frac{k}{m}-x\right)H_x(x,y) + \left(\frac{\tau\ell}{m}-y\right)H_y(x,y) + r_1 \qquad (2.14)$$

where $r_1 = o\left(\left(\frac{k}{m}-x\right)^2 + \left(\frac{\tau\ell}{m}-y\right)^2\right)$. If we define $S_m(z) = \sum_{k=0}^{\infty}\mathrm{Pois}^2_{k,m}(z)$, it can be obtained that

$$
\begin{aligned}
I_1 &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} H\left(\frac{k}{m}, \frac{\tau\ell}{m}\right)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau) \\
&= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} H(x,y)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau) \\
&\quad+\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\left(\left(\frac{k}{m}-x\right)H_x(x,y) + \left(\frac{\tau\ell}{m}-y\right)H_y(x,y) + r_1\right)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau) \\
&= H(x,y)S_m(x)S_m(y/\tau) + H_x(x,y)S_m(y/\tau)O\left(I_m(x)\right) + \tau H_y(x,y)S_m(x)O\left(I_m(y/\tau)\right) + R_1,
\end{aligned}
$$

where $R_1 = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} r_1\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau)$.

Similarly, we utilize Taylor expansion of $H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)$ around (x,y), which implies for I_2

$$\begin{aligned}
I_2 &= \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau)\\
&= \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,l'\neq\ell}^{\infty} H(x,y)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau)\\
&\quad+\sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty}\left(\left(\frac{k\wedge k'}{m}-x\right)H_x(x,y)+\left(\tau\frac{\ell\wedge\ell'}{m}-y\right)H_y(x,y)+r_2\right)\\
&\qquad\qquad\times\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{l,m}(y/\tau)\mathrm{Pois}_{l',m}(y/\tau)\\
&= H(x,y)\left(1-S_m(x)\right)\left(1-S_m(y/\tau)\right)+2R_{1,m}(x)H_x(x,y)\left(1-S_m(y/\tau)\right)\\
&\quad+2\tau H_y(x,y)\left(1-S_m(x)\right)R_{1,m}(y/\tau)+R_2,
\end{aligned}$$

where $r_2 = o\left(\left(\frac{k\wedge k'}{m}-x\right)^2+\left(\tau\frac{\ell\wedge\ell'}{m}-y\right)^2\right)$ and

$$R_2 = \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} r_2\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau).$$

For I_3, we have

$$\begin{aligned}
I_3 &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} H\left(\frac{k}{m}, \tau\frac{\ell\wedge\ell'}{m}\right)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau)\\
&= \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} H(x,y)\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau)\\
&\quad+\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty}\left(\left(\frac{k}{m}-x\right)H_x(x,y)+\left(\tau\frac{\ell\wedge\ell'}{m}-y\right)H_y(x,y)+r_3\right)\\
&\qquad\qquad\times\mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau)\\
&= H(x,y)S_m(x)\left(1-S_m(y/\tau)\right)+H_x(x,y)\left(1-S_m(y/\tau)\right)O\left(I_m(x)\right)\\
&\quad+2\tau H_y(x,y)S_m(x)R_{1,m}(y/\tau)+R_3,
\end{aligned}$$

where $r_3 = o\left(\left(\frac{k}{m} - x\right)^2 + \left(\tau\frac{\ell\wedge\ell'}{m} - y\right)^2\right)$ and

$$R_3 = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{\ell'=0,\ell'\neq\ell}^{\infty} r_3 \mathrm{Pois}^2_{k,m}(x)\mathrm{Pois}_{\ell,m}(y/\tau)\mathrm{Pois}_{\ell',m}(y/\tau).$$

For I_4, we have

$$\begin{aligned}
I_4 &= \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty} H\left(\frac{k\wedge k'}{m}, \tau\frac{\ell}{m}\right)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau)\\
&= \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty} H(x,y)\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau)\\
&\quad + \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty}\left(\left(\frac{k\wedge k'}{m} - x\right)H_x(x,y) + \left(\frac{\tau\ell}{m} - y\right)H_y(x,y) + r_4\right)\\
&\qquad\qquad \times\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau)\\
&= H(x,y)\left(1 - S_m(x)\right)S_m(y/\tau) + 2R_{1,m}(x)H_x(x,y)S_m(y/\tau)\\
&\quad + \tau H_y(x,y)\left(1 - S_m(x)\right)O\left(I_m(y/\tau)\right) + R_4,
\end{aligned}$$

where $r_4 = o\left(\left(\frac{k\wedge k'}{m} - x\right)^2 + \left(\tau\frac{\ell}{m} - y\right)^2\right)$ and

$$R_4 = \sum_{k=0}^{\infty}\sum_{k'=0,k'\neq k}^{\infty}\sum_{\ell=0}^{\infty} r_4\mathrm{Pois}_{k,m}(x)\mathrm{Pois}_{k',m}(x)\mathrm{Pois}^2_{\ell,m}(y/\tau).$$

Finally, by combining the above results, we have

$$\begin{aligned}
&\mathbb{E}\left[\eta_i(x,y)^2\right]\\
=&I_1 + I_2 + I_3 + I_4\\
=&H(x,y)S_m(x)S_m(y/\tau) + H_x(x,y)S_m(y/\tau)O\left(I_m(x)\right) + \tau H_y(x,y)S_m(x)O\left(I_m(y/\tau)\right) + R_1\\
&\quad + H(x,y)\left(1 - S_m(x)\right)\left(1 - S_m(y/\tau)\right) + 2R_{1,m}(x)H_x(x,y)\left(1 - S_m(y/\tau)\right)\\
&\qquad\qquad + 2\tau H_y(x,y)\left(1 - S_m(x)\right)R_{1,m}(y/\tau) + R_2\\
&\quad + H(x,y)S_m(x)\left(1 - S_m(y/\tau)\right) + H_x(x,y)\left(1 - S_m(y/\tau)\right)O\left(I_m(x)\right)
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + 2\tau H_y(x,y) S_m(x) R_{1,m}(y/\tau) + R_3 \\
&+ H(x,y)\left(1 - S_m(x)\right) S_m(y/\tau) + 2R_{1,m}(x) H_x(x,y) S_m(y/\tau) \\
&\qquad\qquad + \tau H_y(x,y)\left(1 - S_m(x)\right) O\left(I_m(y/\tau)\right) + R_4 \\
=& H(x,y) + H_x(x,y)\left(O\left(I_m(x)\right) + 2R_{1,m}(x)\right) + \tau H_y(x,y)\left(O\left(I_m(y/\tau)\right) + 2R_{1,m}(y/\tau)\right) \\
&+ o\left(\left(\frac{k}{m} - x\right)^2 + \left(\frac{\tau l}{m} - y\right)^2\right).
\end{aligned}
$$

In conclusion,

$$
\begin{aligned}
\mathbb{E}\left[\eta_i(x,y)^2\right] = & H(x,y) + H_x(x,y)\left(O\left(I_m(x)\right) + 2R_{1,m}(x)\right) \\
& + \tau H_y(x,y)\left(O\left(I_m(y/\tau)\right) + 2R_{1,m}(y/\tau)\right) + o(m^{-1/2}).
\end{aligned} \tag{2.15}
$$

Second, calculate the variance of $\eta_i(x,y)$.

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\eta_i(x,y)\right] &= \mathbb{E}\left[\eta_i(x,y)^2\right] - \left[\mathbb{E}\left(\eta_i(x,y)\right)\right]^2 \\
&= H(x,y) + H_x(x,y)\left(O\left(I_m(x)\right) + 2R_{1,m}(x)\right) \\
&\quad + \tau H_y(x,y)\left(O\left(I_m(y/\tau)\right) + 2R_{1,m}(y/\tau)\right) + o(m^{-1/2}) - \left(F_m^\tau(x,y)\right)^2 \\
&= H(x,y) - \left(F_m^\tau(x,y)\right)^2 + V(x,y) + o(m^{-1/2}),
\end{aligned} \tag{2.16}
$$

where

$$
\begin{aligned}
V(x,y) = & H_x(x,y)\left(O\left(I_m(x)\right) + 2R_{1,m}(x)\right) \\
& + \tau H_y(x,y)\left(O\left(I_m(y/\tau)\right) + 2R_{1,m}(y/\tau)\right).
\end{aligned} \tag{2.17}
$$

Since the property that n examples are independent and identically distributed has been

proved in Proposition 1, the variance of $\widehat{F}_{m,n}(x,y)$ is

$$\begin{aligned}\mathbb{V}\text{ar}[\widehat{F}_{m,n}(x,y)] &= \frac{1}{n^2} \times n \times \mathbb{V}\text{ar}\left[\eta_i(x,y)\right] \\ &= n^{-1}\left[H(x,y) - (F_m^{\tau}(x,y))^2\right] + n^{-1}V(x,y) + o\left(n^{-1}m^{-1/2}\right) \\ &= n^{-1}\sigma^2(x,y) + n^{-1}V(x,y) + o\left(n^{-1}m^{-1/2}\right),\end{aligned}$$

which completes the proof. □

2.3 Asymptotic Normality

Theorem 2.1. *Under Assumption 1, when both m and n tend to infinity, we have*

$$n^{1/2}\left(\widehat{F}_{m,n}(x,y) - (F_m^{\tau}(x,y))\right) \xrightarrow{d.} N\left(0, \sigma^2(x,y) + V(x,y)\right). \tag{2.18}$$

Proof of Theorem 2.1. Note that from (2.1),

$$\widehat{F}_{m,n}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\eta_i(x,y) + O_{a.s.}\left(\frac{\log n}{n}\right).$$

In addition, we know that $F_m^{\tau}(x,y) = \mathbb{E}\left[\eta_i(x,y)\right]$. Therefore, by applying Lindeberg–Lévy Central Limit Theorem, as both m and n tend to infinity,

$$n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i(x,y) - \mathbb{E}\left(\eta_i(x,y)\right)\right) \xrightarrow{d.} N\left(0, \mathbb{V}\text{ar}\left(\eta_i\left(x,y\right)\right)\right), \tag{2.19}$$

which equivalently is

$$n^{1/2}\left(\widehat{F}_{m,n}(x,y) - F_m^{\tau}(x,y)\right) \xrightarrow{d.} N\left(0, \sigma^2(x,y) + V(x,y)\right).$$

□

Chapter 3

Application and Real Data Analysis

In this Chapter, we reported on two cases of application and real data analysis of the proposed Poisson polynomial estimator. The first real data analysis illustrated the data set from insurance company indemnity claims in **R**. The previous methods of step empirical function in Stute (1993) and empirical Bernstein estimator in Dib et al. (2021) were introduced to compare smoothness effect with the Poisson polynomial estimator.

The second example stated a study of age's impact on survival in patients with colon cancer from 929 participants. Purpose of the study was to determine the relationship between time and age and calculate their joint cdf. In this case, time is subject to right censoring while age of the patient is complete.

3.1 Example I: Relationship of Loss and Allocated Loss Adjustment Expenses (ALAE) in Insurance Claims

We put the proposed Poisson polynomial estimator (1.7) into practice to find the relationship between Loss and ALAE in insurance claims. This data set was analyzed by Denuit and Keilegom. (2006), Frees and Valdez. (1998), and more recently by Gribkova

and Lopez. (2015). There are 1500 observations which include allocated loss adjustment expenses (ALAE, X) and indemnity payment (Loss, Y). In practice, the indemnity payment has a corresponding limit, i.e., the insurance policy limit. If the amount reaches over the limit, Loss will be recorded as the limit amount. Therefore, the indemnity payment is subject to right censoring and it is incomplete. Meanwhile, ALAE are costs attributed to the processing of a specific insurance claim. These costs may include payments to third parties for activities like investigating claims, acting as loss adjusters, or as legal counsel for the insurer. In this case, ALAE were recorded as complete data. Table 3.1 shows the first and last three rows of the Loss-ALAE data set below.

Table 3.1: Loss-ALAE data set

ID	Loss	ALAE	Limit	Censored
1	10	3806	500000	0
2	24	5658	1000000	0
3	45	321	1000000	0
⋮	⋮	⋮	⋮	⋮
1498	1000000	43966	1000000	1
1499	1000000	135653	1000000	1
1500	2173595	134743	2500000	0

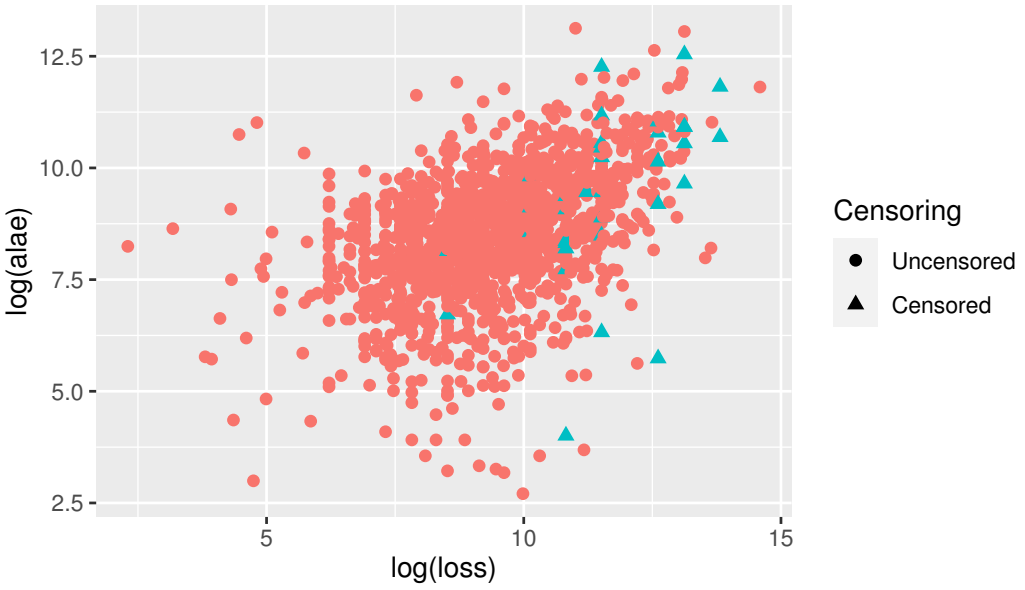
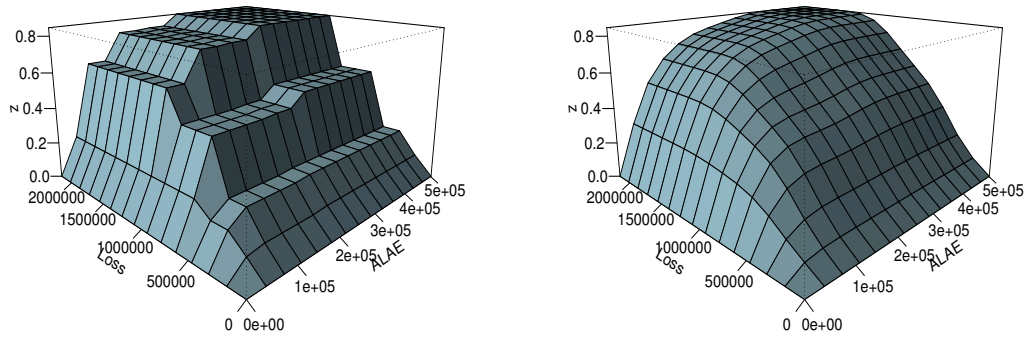


Figure 3.1: Plot of ALAE and Loss in a logarithmic scale

First, the logarithm of ALAE and Loss was utilized to present the their relationship. From Figure 3.1, it is obvious that there exists a strong relationship between these variables. This evidence suggests a joint cumulative distribution function analysis of them.

Furthermore, the Stute empirical estimator and empirical Bernstein estimator were used to estimate the joint cdf of ALAE and Loss. See Figure 3.2 below. It can be seen clearly that there is a significant smoothness improvement from the original Stute empirical function (3.2(a)) to the empirical Bernstein estimator (3.2(b)).



(a) Stute empirical function

(b) Empirical Bernstein estimator

Figure 3.2: Joint cdf estimation by Stute empirical function and empirical Bernstein estimator

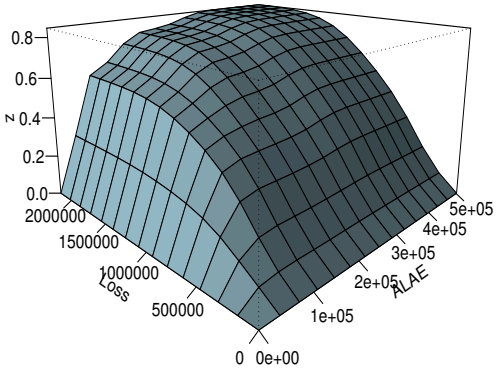


Figure 3.3: Joint cdf estimation by Poisson polynomial estimator

Finally, the proposed Poisson polynomial estimator was applied to this data set as an estimation of the joint cdf. It can be viewed clearly from Figure 3.3 that the Poisson polynomial estimation has the best smoothness effect compared with the Stute estimator

and empirical Bernstein estimator. If we consider the marginal distribution of ALAE or Loss, one can conclude that there is a strong relationship. This matches the preliminary finding in Figure 3.1.

In conclusion, the Poisson polynomial estimator has significantly improved the smoothness of joint cdf estimation. Compare with Stute empirical estimator and empirical Bernstein estimator, it provides a more specific result regarding simulation and potential prediction. However, it requires more calculation than the previous two estimators as the Poisson probability takes value from zero to infinity. In order to obtain a more precise result with acceptable errors, more calculations are needed when using the Poisson polynomial estimator.

3.2 Example II: Age's Impact on Survival in Patients with Colon Cancer

In this section, the Poisson polynomial estimator was applied to find the influence of age on a group of patients with colon cancer. This data set was originally described in Laurie et al. (1989). The main report was found in Moertel CG (1990) and Moertel CG (1991). These data were from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals. 5-FU is a moderately toxic chemotherapy agent. There are two records for every participant, one for recurrence and one for death. Since death is the event of interest in this case, we only select data regarding death. Table 3.2 presents the first and last three rows of the data set.

Table 3.2: Chemotherapy for Stage B/C colon cancer

ID	Study	rx	Sex	Age	Obstruct	Perfor	Adhere	Nodes	Status	Differ	Extent	Surg	Node4	Time	Etype
1	1	Lev+5FU	1	43	0	0	0	5	1	2	3	0	1	1521	2
2	1	Lev+5FU	1	63	0	0	0	1	0	2	3	0	0	3087	2
3	1	Obs	0	71	0	0	1	7	1	2	2	0	1	963	2
⋮	⋮	⋮	⋮	⋮											
927	1	Lev	1	76	0	0	1	1	1	3	3	0	0	1018	2
928	1	Lev+5FU	0	48	1	0	0	4	0	2	3	1	1	2072	2
929	1	Lev	0	66	1	0	0	1	0	2	3	0	0	1820	2

Variable description

ID id.
Study 1 for all patients.
rx Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU.
Sex 0 = female; 1 = male.
Age in years.
Obstruct obstruction of colon by tumour.
Perfor perforation of colon.
Adhere adherence to nearby organs.
Nodes number of lymph nodes with detectable cancer.
Status censoring status (0 = censored, 1 = event)
Differ differentiation of tumour (1 = well, 2 = moderate, 3 = poor).
Extent Extent of local spread (1 = submucosa, 2 = muscle, 3 = serosa, 4 = contiguous).
Surg time from surgery to registration (0 = short, 1 = long).
Node4 more than 4 positive lymph nodes.
Time Survival time in days.
Etype event type: 1 = recurrence, 2 = death.

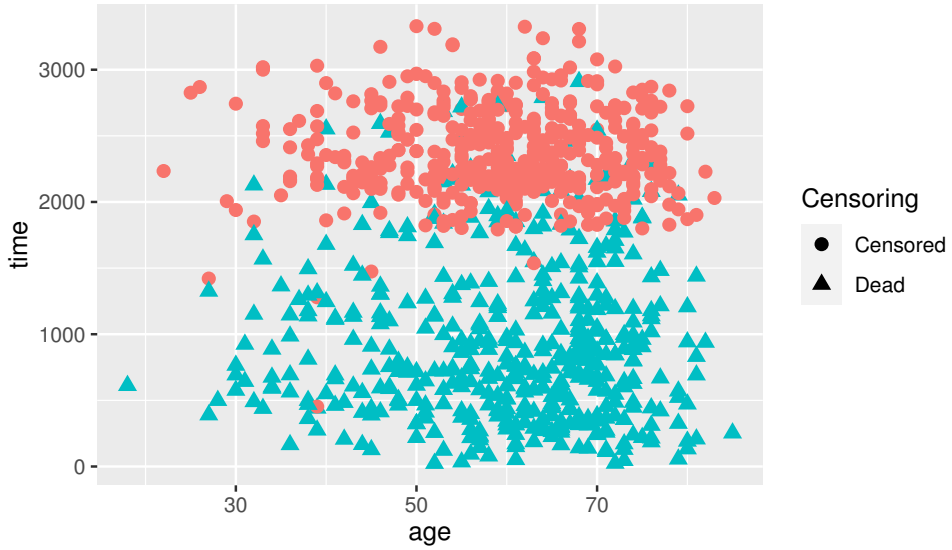


Figure 3.4: Plot of age and survival time

First, Cox PH model and Akaike Information Criterion (AIC) were used to determine which covariates play an important role in the overall survival. Complete analysis with the **R** code is attached to Appendix A. From the preliminary result, one can find that under

AIC step selection, model with rx, age, obstruct, nodes, differ, extent, surg, and node4 has the smallest value (5415.89), that is, the listed 8 vairables have significant effect on the survival probability. Therefore, we consider the relationship between age (complete) and survival time (right-censored) and make the research of their joint distribution function. Figure 3.4 illustrates the scatter plot of age and the corresponding survival time of all participants.

From the scatter plot, one can observe that most cases with the event (death) occurred within 1500 days while most cases with censoring occurred from 1500 to 3000 days. Then we can easily obtain the joint distribution function estimation using Stute's empirical function. See Figure 3.5 below. It is significant that when age is more than 40, one can notice an significantly growing probability as the survival time increases.

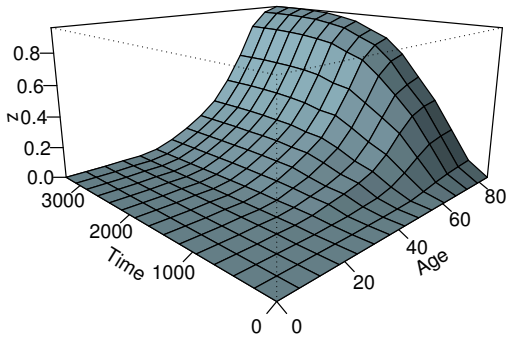
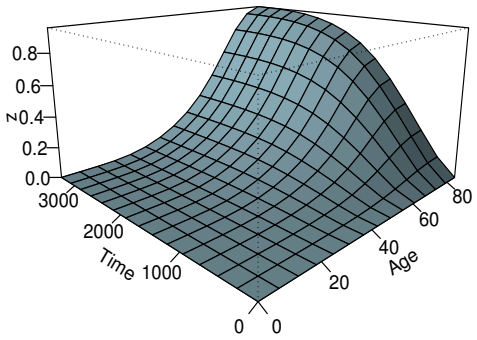
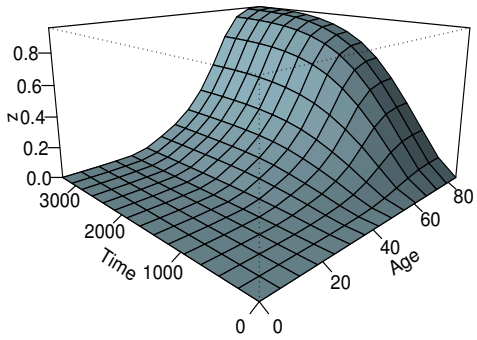


Figure 3.5: Joint cdf estimation by Stute empirical function



(a) Empirical Bernstein estimator



(b) Poisson polynomial estimator

Figure 3.6: Joint cdf estimation by empirical Bernstein estimator and Poisson polynomial estimator

Finally, the empirical Bernstein estimator and the Poisson polynomial estimator were applied to the data set. See Figure 3.6 above.

Compared with empirical Bernstein estimator (3.6(a)), It can be seen that the Poisson polynomial estimator (3.6(b)) generates a closer cdf to the original one by Stute empirical function. It is more obvious when age takes the value greater than 70. Meanwhile, it is evident that different ages have contrasting impact on the joint cdf. With the joint cdf, we are able to derive the estimated marginal distribution function of age given a certain time. Also, the conclusion can be made if there exists an age impact on survival probability under different circumstances.

Furthermore, we should point out that because of the fact that Poisson probability takes value from zero to positive infinity, more calculations are required when a researcher sets the acceptable error closer to the real value. An alternative method is to modify the value of smoothing parameter m to obtain a better result.

Chapter 4

Conclusion

In this major paper, we have discussed the nonparametric bivariate distribution estimation under right censoring using Poisson polynomials. First of all, two difficulties were addressed: estimation of joint distribution function and data incompleteness. To deal with them, the previous study of empirical Bernstein estimator by Dib et al. (2021) utilized the Bernstein polynomials with binomial distribution to smooth the empirical joint cdf estimation by Stute (1993). While this major research paper proposed an estimator with Poisson probability to smooth it. We illustrated the Poisson polynomial estimator (1.7) of the joint cdf in the case of a non-negative random vector (X, Y) where X was assumed as complete while Y was subject to right censoring.

The asymptotic properties of the Poisson polynomial estimator were established in Chapter 2, which included i.i.d. representation, asymptotic bias and variance, and asymptotic normality. These properties support the feasibility of practical implementation.

Finally, two examples of real data applications of the Poisson estimator were stated. Also, we presented Poisson polynomial estimator's comparison with the proposed Stute estimator and empirical Bernstein estimator. One can observe that the Poisson polynomial estimator presents a better smoothness on joint distribution function estimation. Thus, it provides an alternative solution for real data analysis and simulation with continuous derivatives. However, the Poisson polynomial estimator requires more calculations than

the previous two estimators as the Poisson probability takes value from zero to infinity. In order to obtain a more precise result with acceptable errors, more calculations will be needed when using the Poisson polynomial estimator.

Bibliography

Babu, G. J. and Y. P. Chaubey (2006). Smooth estimation of a distribution and density function on a hyper-cube using Bernstein polynomials for dependent random vectors. *Statistics and Probability Letters 76*, 959–969.

Belalia, M. (2016). On the asymptotic properties of the bernstein estimator of the multivariate distribution function. *Statistics & Probability Letters 110*(C), 249–256.

Bernstein, S. (1912). Démonstration du théorème de weierstrass fondée sur le calcul des probabilités. *Communications of the Kharkov Mathematical Society 13*, 1–2.

Denuit, M., O. P. and I. V. Keilegom. (2006). Bivariate archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice 13*, 5–32.

Dib, K., T. Bouezmarni, M. Belalia, and A. Kitouni (2021). Nonparametric bivariate distribution estimation using bernstein polynomials under right censoring. *Communications in Statistics - Theory and Methods 50*(23), 5574–5584.

Frees, E. and E. A. Valdez. (1998). Understanding relationships using copulas. *North American Actuarial Journal 2*(1), 1–25.

Gribkova, S. and O. Lopez. (2015). Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics 42*(4), 925–46.

Hanif, M., S. . Shahzadi, U. Shahzad, and M. Koyuncu (2018). On the adaptive nadaraya-watson kernel estimator for the discontinuity in the presence of jump size. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 511–520.

Laurie, J. A., C. G. Moertel, T. R. Fleming, H. S. Wieand, J. E. Leigh, J. Rubin, G. W. McCormack, J. B. Gerstner, J. E. Krook, and J. Malliard (1989). Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil: The north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology 7*(10), 1447–1456.

Major, P. and L. Rejto. (1988). Strong embedding of the estimator of the distribution function under random censorship. *The Annals of Statistics 16*(3), 1113–32.

Moertel CG, Fleming TR, M. J. H. D. L. J. G. P. U. J. E. W. T. D. G. J. e. a. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine 332*(6), 352–358.

Moertel CG, Fleming TR, M. J. H. D. L. J. T. C. U. J. E. W. T. D. G. J. e. a. (1991). Fluorouracil plus levamisole as an effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of Internal Medicine 122*, 321–326.

Moore, D. F. (2015). *Applied Survival Analysis Using R*. Springer International Publishing.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Routledge.

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

Wand, M. P. and M. C. Jones (1995). *Kernel smoothing*. Boca Raton: Chapman and Hall.

Appendices

Appendix A

Cox PH Model Analysis in Example II

In Appendix I, **Cox PH model** was utilized to analyze which covariates play an important role in the overall survival for Example II in Section 3.2. Further, **Akaike Information Criterion (AIC)** was the criteria to analyze which model fits better. Below is the **R** code using the step function to select the best fit Cox PH model under AIC.

```
> kfit1 <- coxph(Surv(time, status)~ rx + age + sex + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4, mydata,
     id = id)
> result.step <- step(kfit1, scope = list(upper=~rx + age + sex +
    obstruct + perfor + adhere + nodes + differ + extent + surg +
    node4, lower=~1))
Start:  AIC=5420.21
Surv(time, status) ~ rx + age + sex + obstruct + perfor + adhere +
nodes + differ + extent + surg + node4

Df AIC
- perfor    1 5418.2
- sex       1 5418.2
```

```
- adhere    1 5419.8
- differ    1 5420.1
<none>        5420.2
- age       1 5421.5
- obstruct  1 5423.0
- surg      1 5423.2
- nodes     1 5425.4
- rx        2 5426.6
- extent    1 5433.4
- node4     1 5440.3

Step:  AIC=5418.21
Surv(time, status) ~ rx + age + sex + obstruct + adhere + nodes +
differ + extent + surg + node4

Df AIC
- sex       1 5416.2
- adhere    1 5417.9
- differ    1 5418.1
<none>        5418.2
- age       1 5419.5
+ perfor    1 5420.2
- obstruct  1 5421.1
- surg      1 5421.2
- nodes     1 5423.4
- rx        2 5424.6
- extent    1 5431.5
- node4     1 5438.3

Step:  AIC=5416.22
Surv(time, status) ~ rx + age + obstruct + adhere + nodes + differ +
extent + surg + node4
```

```
Df AIC
- adhere    1 5415.9
- differ    1 5416.1
<none>        5416.2
- age       1 5417.5
+ sex       1 5418.2
+ perfor    1 5418.2
- obstruct  1 5419.1
- surg      1 5419.2
- nodes     1 5421.4
- rx        2 5422.7
- extent    1 5429.5
- node4     1 5436.5

Step:  AIC=5415.89
Surv(time, status) ~ rx + age + obstruct + nodes + differ + extent +
surg + node4

Df AIC
<none>        5415.9
- differ    1 5416.2
+ adhere    1 5416.2
- age       1 5417.7
+ perfor    1 5417.8
+ sex       1 5417.9
- obstruct  1 5418.8
- surg      1 5418.9
- nodes     1 5421.1
- rx        2 5422.5
- extent    1 5430.1
- node4     1 5436.2
```

It can be obtained that under AIC step selection, model with rx, age, obstruct, nodes, differ, extent, surg, and node4 has the smallest value (5415.89). This model is selected as the best fit. Thus, it is worth executing the Cox PH model with these variables for more information.

```
> kfitb <- coxph(Surv(time, status) ~ rx + age  + obstruct + nodes +
    differ + extent + surg + node4, mydata, id = id)
> kfitb
Call:
coxph(formula = Surv(time, status) ~ rx + age + obstruct + nodes +
differ + extent + surg + node4, data = mydata)

coef      exp(coef) se(coef)    z       p
rxLev+5FU -0.325460  0.722195  0.124439 -2.615  0.00891
rxObs      0.038778  1.039540  0.114245  0.339  0.73428
age        0.008067  1.008100  0.004159  1.940  0.05243
obstruct   0.270961  1.311224  0.119362  2.270  0.02320
nodes      0.043774  1.044746  0.015169  2.886  0.00391
differ     0.152127  1.164308  0.100348  1.516  0.12952
extent     0.461501  1.586453  0.118521  3.894 9.87e-05
surg       0.241908  1.273676  0.106109  2.280  0.02262
node4      0.674768  1.963577  0.141757  4.760 1.94e-06

Likelihood ratio test=137.9  on 9 df, p=< 2.2e-16
n= 888, number of events= 430

> cox.zph(kfitb)
chisq   df   p
rx        2.74663  2 0.25327
age       0.93449  1 0.33370
obstruct  6.49167  1 0.01084
nodes     0.30491  1 0.58082
```

```
differ    13.01058  1 0.00031
extent     4.22054  1 0.03994
surg       0.00355  1 0.95250
node4      4.67143  1 0.03067
GLOBAL    35.38326  9 5.1e-05
```

Interpretation of this Cox PH model:

(1) The p-value of covariate **age** is 0.05243, which reveals that there exists significant difference under the condition of 90% confidence level. Also, the Cox PH model passed the regression test with 95% confidence level (0.05).

(2) The exponential coefficient of **age** equals to 1.0081, indicating that the probability of experiencing the cardiac event is 1.0081 times as the value of age increases one unit. The increasing rate of the hazard function is 0.081%.

(3) Since the result from *cox.zph*() shows that all p-values (0.33) are greater than 0.05, the null hypothesis :"PH assumption is not violated" is not rejected. The Cox PH model has a well of fit.

Appendix B

R Code in Example I and II

B.1 Example I

B.1.1 Part I: Loss-ALAE Logarithm Scale Relationship

```
library(ggplot2)
data(loss, package="copula")
loss$censored <- factor(loss$censored)
ggplot(loss, aes(x = log(loss), y = log(alae), shape = censored))+
   geom_point(alpha = 1)+geom_point(size = 2)+scale_shape(name = "
   Censoring", labels = c("Uncensored","Censored"))
```

B.1.2 Part II: Stute Empirical Function and Plot

1. Stute empirical function

```
Stute1 <- function(x.eval, y.eval, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m <- 1
        w <-matrix(NA, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
```

```
                for (j in 1:(i-1)) {
                        a <- mydata$censored[j]
                        b <- ((n-j)/(n-j+1))^a
                        m <- m*b
                }
                c <- (mydata$censored[i])/(n-i+1)
                w[i,1] <- c*m
                m <- 1
        }
        #Stute estimation
        z <- matrix(NA, nrow =n1 , ncol = n1)
        sum1 <- 0
        for (j in 1:n1) {
                for (h in 1:n1) {
                        for (i in 1:n) {
                                if (mydata$loss[i] <= y.eval[j] &&
                                    mydata$alae[i] <= x.eval[h]){
                                        sum1 <- sum1 + w[i,1]
                                }
                        }
                        z[h,j] <- sum1
                        sum1 <- 0
                }
        }
        return(z)
}
```

2. Stute empirical estimator plot

```
library(copula)
data(loss)

source("Example_I_Stute.R")
```

```
n1 <- 15
upx <- max(loss$alae)
upy <- max(loss$loss)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
z = Stute1(x.eval = xstep, y.eval = ystep, mydata = loss)
#plot of Stute estimation
par(mar=c(1.2,2,0,0))
persp(x = xstep, y = ystep, z, theta = 315, phi = 15,
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
   , xlab = "ALAE", ylab = "Loss")
```

B.1.3 Part III: Empirical Bernstein Estimator and Plot

1. Empirical Bernstein estimator

```
Empirical_Bernstein_estimator1 <- function(m, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m1 <- 1
        w <-matrix(0, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
                for (j in 1:(i-1)) {
                        a <- mydata$censored[j]
                        b <- ((n-j)/(n-j+1))^a
                        m1 <- m1*b
                }
                c <- (mydata$censored[i])/(n-i+1)
                w[i,1] <- c*m1
                m1 <- 1
        }
        #Stute estimation for empirical Bernstein estimator
```

```
upx <- max(mydata$alae)
upy <- max(mydata$loss)
xstep2 <- seq(from = 0, to = upx, length.out = m+1)
ystep2 <- seq(from = 0, to = upy, length.out = m+1)
z0 <- matrix(0, nrow = m+1, ncol = m+1)
sum1 <- 0
for (j in 1:m+1) {
        for (h in 1:m+1) {
                for (i in 1:n) {
                        if (mydata$loss[i] <= ystep2[j] &&
                           mydata$alae[i] <= xstep2[h]){
                                sum1 <- sum1 + w[i,1]
                        }
                }
                z0[j,h] <- sum1
                sum1 <- 0
        }
}
#empirical Bernstein estimator
zb <- matrix(0, nrow = n1, ncol = n1)
k <- seq(0, m, 1)
l <- seq(0, m, 1)
for (j in 0:(n1-1)) {
        for (h in 0:(n1-1)) {
                p1 <- dbinom(k, m, j/(n1-1))
                p1 <- matrix(p1)
                p1 <- t(p1)
                p2 <- dbinom(l, m, h/(n1-1))
                p2 <- matrix(p2)
                zb[h+1,j+1] <- p1%*%z0%*%p2
        }
}
```

```
        #joint cdf matrix
        return(zb)
}
```

2. Empirical Bernstein estimator plot

```
library(copula)
data(loss)

source("Example_I_Binomial.R")
n1 <- 15
zb = Empirical_Bernstein_estimator1(m = 25, mydata = loss)
#plot of Stute estimation
upx <- max(loss$alae)
upy <- max(loss$loss)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
par(mar=c(1.2,2,0,0))
persp(x = xstep, y = ystep, zb, theta = 315, phi = 15,
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
    , xlab = "ALAE", ylab = "Loss", zlab = "z")
```

B.1.4 Part IV: Poisson Estimator and Plot

1. Poisson estimator

```
Poisson_estimator1 <- function(m, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m1 <- 1
        w <-matrix(0, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
                for (j in 1:(i-1)) {
```

```
                a <- mydata$censored[j]
                b <- ((n-j)/(n-j+1))^a
                m1 <- m1*b
        }
        c <- (mydata$censored[i])/(n-i+1)
        w[i,1] <- c*m1
        m1 <- 1
}
#Stute estimation for Poisson estimator
upx <- max(mydata$alae)
mydata$alae <- (mydata$alae)/100000
upy <- max(mydata$loss)
upxnew <- max(mydata$alae)
xstep2 <- seq(from = 0, to = 2*upxnew, length.out = (2*round
   (upxnew)*m+1))
ystep2 <- seq(from = 0, to = (2*upy), length.out = (2*m+1))
z1 <- matrix(0, nrow = (2*m+1), ncol = (2*round(upxnew)*m+1)
   )
sum1 <- 0
for (j in 1:(2*m+1)) {
        for (h in 1:(2*round(upxnew)*m+1)) {
                for (i in 1:n) {
                        if (mydata$loss[i] <= ystep2[j] &&
                           mydata$alae[i] <= xstep2[h]){
                                sum1 <- sum1 + w[i,1]
                        }
                }
                z1[j,h] <- sum1
                sum1 <- 0
        }
}
#Poisson estimator
```

```
        zp <- matrix(0, nrow = n1, ncol = n1)
        k <- seq(0, 2*m, 1)
        l <- seq(0, 2*round(upxnew)*m, 1)
        for (j in 0:(n1-1)) {
                for (h in 0:(n1-1)) {
                        p1 <- dpois(k, m*(j/(n1-1)))
                        p1 <- matrix(p1)
                        p1 <- t(p1)
                        p2 <- dpois(l, m*h*(max(mydata$alae)/(n1-1))
                           )
                        p2 <- matrix(p2)
                        zp[h+1,j+1] <- p1%*%z1%*%p2
                }
        }
        #cdf matrix
        return(zp)
}
```

2. Poisson estimator plot

```
library(copula)
data(loss)

source("Example_I_Poisson.R")
n1 <- 15
zp = Poisson_estimator1(m = 25, mydata = loss)
#plot of Stute estimation
upx <- max(loss$alae)
upy <- max(loss$loss)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
par(mar=c(1.2,2,0,0))
persp(x = xstep, y = ystep, zp, theta = 315, phi = 15,
```

```
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
   , xlab = "ALAE", ylab = "Loss", zlab = "z")
```

B.2 Example II

B.2.1 Part I: Age-time Relationship Plot

```
library(ggplot2)
mydata <- read.csv("Example2_data.csv")
mydata$status <- factor(mydata$status)
ggplot(mydata, aes(x = age, y = time, shape = status))+geom_point(
   alpha = 1)+geom_point(size = 2.5)+scale_shape(name = "Censoring",
    labels = c("Censored","Dead"))
```

B.2.2 Part II: Stute Empirical Function and Plot

1. Stute empirical function

```
Stute2 <- function(x.eval, y.eval, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m <- 1
        w <-matrix(NA, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
                for (j in 1:(i-1)) {
                        a <- mydata$status[j]
                        b <- ((n-j)/(n-j+1))^a
                        m <- m*b
                }
                c <- (mydata$status[i])/(n-i+1)
                w[i,1] <- c*m
                m <- 1
```

```
        }
        #Stute estimation
        z <- matrix(NA, nrow = n1, ncol = n1)
        sum1 <- 0
        for (j in 1:n1) {
                for (h in 1:n1) {
                        for (i in 1:n) {
                                if (mydata$time[i] <= ystep[j] &&
                                    mydata$age[i] <= xstep[h]){
                                        sum1 <- sum1 + w[i,1]
                                }
                        }
                        z[h,j] <- sum1
                        sum1 <- 0
                }
        }
        return(z)
}
```

2. Stute empirical estimator plot

```
library(survival)
data_1 <- read.csv("Example2_data.csv")

source("Example_II_Stute.R")
n1 <- 15
upx <- max(data_1$age)
upy <- max(data_1$time)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
z = Stute2(x.eval = xstep, y.eval = ystep, mydata = data_1)
#plot of Stute estimation
par(mar=c(1.2,2,0,0))
```

```
persp(x = xstep, y = ystep, z, theta = 315, phi = 15,
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
   , xlab = "Age", ylab = "Time")
```

B.2.3 Part III: Empirical Bernstein Estimator and Plot

1. Empirical Bernstein estimator

```
Empirical_Bernstein_estimator2 <- function(m, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m1 <- 1
        w <-matrix(0, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
                for (j in 1:(i-1)) {
                        a <- mydata$status[j]
                        b <- ((n-j)/(n-j+1))^a
                        m1 <- m1*b
                }
                c <- (mydata$status[i])/(n-i+1)
                w[i,1] <- c*m1
                m1 <- 1
        }
        #Stute for empirical Bernstein estimator
        upx <- max(mydata$age)
        upy <- max(mydata$time)
        xstep1 <- seq(from = 0, to = upx, length.out = m+1)
        ystep1 <- seq(from = 0, to = upy, length.out = m+1)
        z0 <- matrix(0, nrow = m+1, ncol = m+1)
        sum1 <- 0
        for (j in 1:m+1) {
                for (h in 1:m+1) {
```

```
                        for (i in 1:n) {
                                if (mydata$time[i] <= ystep1[j] &&
                                    mydata$age[i] <= xstep1[h]){
                                        sum1 <- sum1 + w[i,1]
                                }
                        }
                        z0[j,h] <- sum1
                        sum1 <- 0
                }
        }
        #empirical Bernstein estimator
        zb <- matrix(0, nrow = n1, ncol = n1)
        k <- seq(0, m, 1)
        l <- seq(0, m, 1)
        for (j in 0:(n1-1)) {
                for (h in 0:(n1-1)) {
                        p1 <- dbinom(k, m, j/(n1-1))
                        p1 <- matrix(p1)
                        p1 <- t(p1)
                        p2 <- dbinom(l, m, h/(n1-1))
                        p2 <- matrix(p2)
                        zb[h+1,j+1] <- p1%*%z0%*%p2
                }
        }
        #joint cdf matrix
        return(zb)
}
```

2. Empirical Bernstein estimator plot

```
library(survival)
data_1 <- read.csv("Example2_data.csv")
```

```
source("Example_II_Binomial.R")
n1 <- 15
zb = Empirical_Bernstein_estimator2(m = 25, mydata = data_1)
#plot of Stute estimation
upx <- max(data_1$age)
upy <- max(data_1$time)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
par(mar=c(1.2,2,0,0))
persp(x = xstep, y = ystep, zb, theta = 315, phi = 15,
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
   , xlab = "Age", ylab = "Time", zlab = "z")
```

B.2.4 Part IV: Poisson Estimator and Plot

1. Poisson estimator

```
Poisson_estimator2 <- function(m, mydata){
        #KM estimator
        n <- length(mydata[[1]])
        m1 <- 1
        w <-matrix(0, nrow = n, ncol = 1)
        w[1,1] <- 0
        for (i in 2:n) {
                for (j in 1:(i-1)) {
                        a <- mydata$status[j]
                        b <- ((n-j)/(n-j+1))^a
                        m1 <- m1*b
                }
                c <- (mydata$status[i])/(n-i+1)
                w[i,1] <- c*m1
                m1 <- 1
        }
```

```
#Stute estimation for Poisson estimator
upx <- max(mydata$age)
mydata$age <- (mydata$age)/10
upxnew <- max(mydata$age)
upy <- max(mydata$time)
xstep2 <- seq(from = 0, to = 2*upxnew, length.out = (2*round
   (upxnew)*m+1))
ystep2 <- seq(from = 0, to = (2*upy), length.out = (2*m+1))
z1 <- matrix(0, nrow = (2*m+1), ncol = (2*round(upxnew)*m+1)
   )
sum1 <- 0
for (j in 1:(2*m+1)) {
        for (h in 1:(2*round(upxnew)*m+1)){
                for (i in 1:n) {
                        if (mydata$time[i] <= ystep2[j] &&
                           mydata$age[i] <= xstep2[h]){
                                sum1 <- sum1 + w[i,1]
                        }
                }
                z1[j,h] <- sum1
                sum1 <- 0
        }
}
#Poisson estimator
zp <- matrix(0, nrow = n1, ncol = n1)
k <- seq(0, 2*m, 1)
l <- seq(0, 2*round(upxnew)*m, 1)
for (j in 0:(n1-1)) {
        for (h in 0:(n1-1)) {
                p1 <- dpois(k, m*(j/(n1-1)))
                p1 <- matrix(p1)
                p1 <- t(p1)
```

```
                        p2 <- dpois(l, m*h*(max(mydata$age)/(n1-1)))
                        p2 <- matrix(p2)
                        zp[h+1,j+1] <- p1%*%z1%*%p2
                }
        }
        #plot of Poisson estimator
        xstep <- seq(from = 0, to = upx, length.out = n1)
        ystep <- seq(from = 0, to = upy, length.out = n1)
        persp(x = xstep, y = ystep, zp, theta = 315, phi = 15,
           expand = 0.5, col = "lightblue", shade = 0.75, ticktype =
            "detailed", xlab = "Age", ylab = "Time", zlab = "z")
        #cdf matrix
        return(zp)
}
```

2. Poisson estimator plot

```
library(survival)
data_1 <- read.csv("Example2_data.csv")

source("Example_II_Poisson.R")
n1 <- 15
zp = Poisson_estimator2(m = 50, mydata = data_1)
#plot of Stute estimation
upx <- max(data_1$age)
upy <- max(data_1$time)
xstep <- seq(from = 0, to = upx, length.out = n1)
ystep <- seq(from = 0, to = upy, length.out = n1)
par(mar=c(1.2,2,0,0))
persp(x = xstep, y = ystep, zp, theta = 315, phi = 15,
expand = 0.5, col = "lightblue", shade = 0.75, ticktype = "detailed"
   , xlab = "Age", ylab = "Time", zlab = "z")
```

Vita Auctoris

NAME: Luchen Liu

PLACE OF BIRTH: Xi'an, Shaanxi, P.R.C.

YEAR OF BIRTH: 1998

EDUCATION: Beijing Institute of Technology, B.Sc., Beijing, P.R.C., 2016-2020

University of Windsor, M.Sc., Windsor, ON, 2020-2022