Nonparametric predictive inference with parametric copula for survival analysis

N Muhammad*, and *N* Yusoff

Fakulti Sains & Teknologi Industri, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

Abstract. Many real-world problems of statistical inference involve dependent bivariate data including survival analysis. This paper presents new nonparametric methods for predictive inference for survival analysis involving a future bivariate observation. The method combine between bivariate Nonparametric Predictive Inference (NPI) for the marginals with parametric copula to take dependence structure into account. The proposed method is a discretized version of the parametric copula. The NPI fits the marginal and very straight forward computations. Generally, NPI is a frequentist approach which infer a future observation based on past data. The proposed method resulting imprecision is robustness with regard to the assumed parametric copula in the marginal for prediction. This is practical for small data set. The suggestion is to use a basic parametric copula for small data sets. We investigate and discuss the performance of these methods by presenting results from simulation studies. The method is further illustrated via application in survival analysis using data sets from the literature.

1 Introduction

Survival analysis is defined as a set of methods or tools for analysing data where the variable is the time until the occurrence of an event of interest [1, 2, 3, 4]. For example, time to be healed, time to death, length of stay in a hospital, time to marriage, time to divorce, money paid by health insurance or viral load measurement.

In this study, we focus on time-to-event data such as onset of disease in medical and time to failure of mechanical system in engineering. There are many methods use to analysing the data such as proportional hazards and accelerated failure time models have been developed. However, these methods are used when the assumption is independent [5, 6, 7]. For dependent assumption of observed failure time such as clustered survival data or parallel events, those method are unsuitable whereby the dependent of the data are unaccounted. For example, clustered survival data arise when event times belonging to the same cluster are correlated. Consider diagnosis of hip fracture being healed in a dog from [8]. In the study, the time to diagnosis is measured by two different imaging techniques. The first technique is radiography (RX) and the second techniques is an ultrasound (US). This resulting two clustered diagnosis times which should consider the dependence

* Corresponding author: noryanti@ump.edu.my

structure.

Therefore, in this paper, we introduce survival analysis data using Nonparametric Predictive Inference (NPI) for the marginals with parametric copula to take dependence into account. NPI is a frequentist statistical framework for inference on a future observation based on past data observations [9]. The uncertainty in NPI is quantifies through imprecision which only based on a few assumption. Basically, the imprecise probabilities is a classic probability theory which allow for partial probability specification and useful if applicable information is not enough and difficult to obtain. In this paper, we are focusing on predictive inference involving a future bivariate observation for survival analysis data.

2 Methodology

2.1 Bivariate data

Let (X_i, Y_i) be a bivariate random quantity where i = 1, . . . , n. Let X_{n+1} and Y_{n+1} represent the future observations of the random quantities X and Y , respectively and X_{n+1} and Y_{n+1} represent the transformations of X_{n+1} and Y_{n+1} as follows given in [10] and [11]:

$$\left(\tilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \tilde{Y}_{n+1} \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right)\right) \Leftrightarrow \left(X_{n+1} \in \left(x_{(i-1)}, x_{(i)}\right), Y_{n+1} \in \left(y_{(j-1)}, y_{(j)}\right)\right) \quad (1)$$

The transformation is from the real plane R^2 into $[0, 1]^2$, where i, j = 1, 2, ..., n + 1. The method that follows is applied to the transformed data.

2.2 Copula

A Copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables. By the well-known theorem by Sklar's [12], every joint cumulative distribution function F of continuous random quantities (X, Y) can be written as

$$F(x, y) = C(F_x(x), F_y(y)) \quad (2)$$

for all $(x, y) \in R^2$, where $F_x(x)$ and $F_y(y)$ are the continuous marginal distributions and C : $[0, 1] \times [0, 1] \rightarrow [0, 1]$ the unique copula associated to this joint distribution, F (x, y). So, a copula is a joint cumulative distribution function whose arginal are uniformly distributed on [0, 1] [12, 13].

By using the NPI marginal cumulative distribution functions, we have discretized uniform marginal distributions on [0,1], which therefore fully correspond to copula [10, 11]. Therefore, the transformation shows that the arginal which we use NPI approach can be easily combined with any parametric copulas to reflect the dependence structure.

2.3 Combine NPI with parametric copula

NPI on the arginal can be combined with the estimated parametric copula density, $\hat{\theta}$ as follows [10],

$$h_{ij}\left(\hat{\theta}\right) = P_C\left(\tilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \tilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right) \middle| \hat{\theta}\right) \quad (3)$$

where $i = 1, \ldots, n$ and $PC(\cdot|\hat{\theta})$ represents the copula-based probability with estimated density

$\hat{\theta}$. The values $h_{ij}(\cdot)$ are the main tools of our inferential method and the corresponding cdf,

$$H_{ij}\left(\hat{\theta}\right)=P_C\left(\tilde{X}_{n+1}\le\frac{i}{n+1},\tilde{Y}_{n+1}\le\frac{j}{n+1}\middle|\hat{\theta}\right)=\sum_{k=1}^{i}\sum_{l=1}^{j}h_{kl}\left(\hat{\theta}\right) \tag{4}$$

As given in [10], equation hij ($\hat{\theta}$) can be considered to infer about an event E that involves the next observation (X_{n+1}, Y_{n+1}). Let E(X_{n+1}, Y_{n+1}) denote the event of interest and let $\underline{P}\left(E\left(X_{n+1},Y_{n+1}\right)\right)$ and $\overline{P}\left(E\left(X_{n+1},Y_{n+1}\right)\right)$ be the lower and upper probabilities, based on our parametric method. We further define

$$E\left(x,y\right)=\begin{cases}1 & \text{if } E\left(X_{n+1},Y_{n+1}\right)\\ 0 & \text{else}\end{cases} \tag{5}$$

and $\overline{E}_{ij}=\max_{(x,y)\in B_{ij}}E\left(x,y\right)$, so $\overline{E}_{ij}=1$ if there is at least one $\left(x,y\right)\in B_{ij}$ for which $E\left(x,y\right)=1$, else $\overline{E}_{ij}=0$. Then, we define $\underline{E}_{ij}=\min_{(x,y)\in B_{ij}}E\left(x,y\right)$, so $\underline{E}_{ij}=1$ if $E\left(x,y\right)=1$ for all $\left(x,y\right)\in B_{ij}$, else $\underline{E}_{ij}=0$. As mentioned in [10], the blocks $B_{ij}=(x_{i-1},x_i)\odot(y_{j-1},y_j)$, $i,j=1,\ldots,n+1$ is actually the result of transforming the observed data (x_i, y_i), $i = 1, \ldots, n$, which divide R^2 into $(n+1)^2$. This semi-parametric method presented above leads to the following lower and upper probabilities for the event $E(X_{n+1}, Y_{n+1})$ [10],

$$\underline{P}\left(E\left(X_{n+1},Y_{n+1}\right)\right)=\sum_{ij}h_{ij}\left(\hat{\theta}\right) \tag{6}$$

$$\overline{P}\left(E\left(X_{n+1},Y_{n+1}\right)\right)=\sum_{ij}h_{ij}\left(\hat{\theta}\right) \tag{7}$$

For example, we are interested in the event $D_{n+1} = X_{n+1}/Y_{n+1} > d$ where without loss of generality, $Y > 0$. Then, the lower and upper probability for the event is

$$\underline{S}\left(d\right)=\underline{P}\left(D_{n+1}>d\right)=\sum_{(i,j)\in L}h_{ij}\left(\hat{\theta}\right) \tag{8}$$

where $L = \{(i, j) : x_{(i-1)}/y_{(j-1)} > d\}$, and

$$\overline{S}\left(d\right)=\overline{P}\left(D_{n+1}>d\right)=\sum_{(i,j)\in U}h_{ij}\left(\hat{\theta}\right) \tag{9}$$

where $U = \{(i, j) : x_{(i)}/y_{(j)} > d\}$. Basically, equation (8) and (9) are survival function equations.

3 Results and discussions

3.1 Predictive performance

We performed a simulation study to obtain indications of the predictive performance of this approach. Let $\left(x_i^j,y_i^j\right)$ be the jth simulated sample, i = 1, 2, . . . , n and j = 1, 2, . . . , N, and $\left(x_f^j,y_f^j\right)$ be the divisional simulated pair and the corresponding division is denoted by $d_f^j=x_f^j/y_f^j$, j = 1, 2, . . . , N. The first n pairs is used to build the proposed NPI with parametric copula for each simulated sample. Then, the additional pair which consider as a future observation is used to test the prediction performance of the proposed method. The inverse of the lower and upper survival functions of D_{n+1} in $\underline{S}\left(d\right)$ and $\overline{S}\left(d\right)$, can be defined as follows [10] for $q\in(0,1)$,:

$$\underline{d}_q=\underline{S}^{-1}\left(q\right)=\inf_{d\in\mathbb{R}}\left\{\underline{S}\left(d\right)\le q\right\} \tag{10}$$

$$\overline{d}_q = \overline{S}^{-1}(q) = \inf_{d\in\mathbb{R}}\left\{\overline{S}(d) \le q\right\} \tag{11}$$

The proposed method performs well, if the two following inequalities hold,

$$p_1 = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\left(d_f^j \ge \overline{d}_q^j\right) \le q \tag{12}$$

$$p_2 = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\left(d_f^j \ge \underline{d}_q^j\right) \ge q \tag{13}$$

The data were simulated from a copula family and parametric method copula were used in this simulation study;

Table 1. Simulated data from Clayton with estimated density copula, $\hat{\theta}\cdot$

			$n=10$			$n=50$			$n=100$		
τ	θ	q	$\hat{\theta}$	p_1	p_2	$\hat{\theta}$	p_1	p_2	$\hat{\theta}$	p_1	p_2
-0.75	-6.0000	0.25	-9.7782	0.0806	0.5130	-5.8932	0.1859	0.2904	-5.8691	0.2233	0.2741
		0.5		0.2171	0.7735		0.3963	0.5992		0.4428	0.5653
		0.75		0.4770	0.9193		0.7009	0.7992		0.7376	0.7841
-0.5	-2.0000	0.25	-3.1214	0.1581	0.4114	-2.1369	0.2234	0.2732	-2.0693	0.2383	0.2653
		0.5		0.3350	0.6711		0.4526	0.5545		0.4712	0.5286
		0.75		0.5935	0.8427		0.7207	0.7710		0.7377	0.7640
-0.25	-0.6667	0.25	-1.3182	0.1995	0.3742	-0.7863	0.2436	0.2820	-0.7358	0.2405	0.2584
		0.5		0.3919	0.6188		0.4743	0.5312		0.4840	0.5186
		0.75		0.6381	0.8095		0.7235	0.7579		0.7354	0.7528
0.25	0.6667	0.25	1.3232	0.1737	0.2939	0.7934	0.2342	0.2587	0.7349	0.2380	0.2518
		0.5		0.4289	0.5627		0.4784	0.5081		0.4876	0.5018
		0.75		0.7143	0.8119		0.7451	0.7658		0.7457	0.7561
0.5	2.0000	0.25	3.0532	0.1836	0.2953	2.1431	0.2455	0.2711	2.0681	0.2380	0.2516
		0.5		0.4487	0.5522		0.4962	0.5200		0.4916	0.5028
		0.75		0.7091	0.7931		0.7460	0.7651		0.7479	0.7563
0.75	6.0000	0.25	10.1198	0.1970	0.2979	5.8992	0.2342	0.2596	5.8700	0.2458	0.2569
		0.5		0.4587	0.5526		0.4922	0.5098		0.4933	0.5028
		0.75		0.7132	0.8039		0.7337	0.7535		0.7427	0.7529

Based on the performance in table 1, we can see that all the values q ∈ [p1, p2].

3.2 Example: survival analysis data

These data describe the lengths of time required for patients with headaches to achieve relief, each patient receives a standard treatment and a new treatment on separate occasions. The times are recorded to the nearest tenth of a minute [1]. Let (*X, Y*) denote the bivariate variable (the time to relapse of the i^{th} patient on the first treatment, the time to relapse of the i^{th} patient on the second treatment), and suppose that we are interested in the ratio of these two values for the next observation, $D_{n+1} = Y_{n+1}/X_{n+1} > d$ where, without loss of generality, $X > 0$.

For these data, we used bivariate Normal copula, C ($\hat{\theta}$) and the lower and upper probabilities for the event are presented in fig.1.

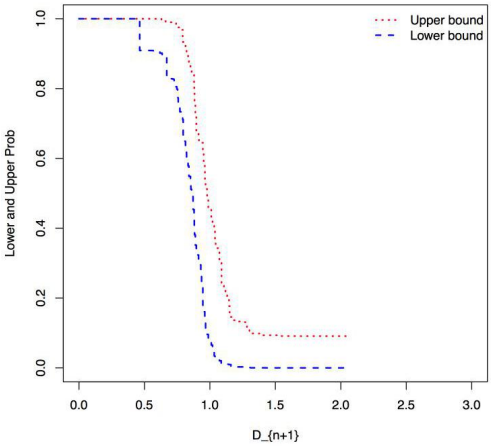

Fig. 1. Upper and Lower Probabilities D_{n+1}.

4 Conclusion

Generally, the main conclusion we draw from the prediction performance of this method is performed well for small values of n, while for larger data sets a nonparametric copula can be used in order to learn more about the dependence structure from the data. The imprecision of the proposed method provides a sufficient robustness which consider to have a good frequentist properties specifically for the predictive inferences. The results is depending on the parametric copulas used, the random quantity and the percentiles studied. The imprecision decreases for increasing sample size.

Acknowledgments

The author gratefully acknowledged the financial support received in the form of university research grant from Universiti Malaysia Pahang (RDU 170359).

References

1. Gross A J and Lam C F 1981 *Biometrics* 505–511
2. Melo L, Schneider R, Manso R, Saucier J P and Fortin M 2017 *Canadian Journal of Forest Research*
3. Wang X, Wang X, Hodgson L, George S L, Sargent D J, Foster N R, Ganti A K, Stinchcombe T E, Crawford J, Kratzke R *et al.* 2017 *The Oncologist* theoncologist–2016
4. Roy D 2017 *Univariate and Multivariate Survival Models with Flexible Hazard Functions* Ph.D. thesis
5. Cook T, Zhang Z and Sun J 2017 *Monte-Carlo Simulation-Based Statistical Modeling* (Springer) pp 319–346 [6] Chen L, Feng Y and Sun J 2017 *Lifetime data analysis* **23** 651–670
6. Vodnala D, Sharma A, Dean E, Kennedy K, Magalski A and Austin B 2017 *Journal of Cardiac Failure* **23** S117
7. Risselada M, van Bree H, Kramer M, Chiers K, Duchateau L, Verleyen P and Saunders J H 2006 *American Journal of Alzheimer's Disease and Other Dementiasournal of Veterinary Research* **67** 1354–1361

8. Coolen F P A 2011 *International Encyclopedia of Statistical Science* ed Lovric M (Springer) pp 968–970
9. Muhammad N 2016 *Predictive inference with copulas for bivariate data* Ph.D. thesis Durham University
10. Coolen-Maturi T, Coolen F P and Muhammad N 2016 *Journal of Statistical Theory and Practice* **10** 515–538
11. Sklar A W 1959 *Publications de lSInstitut de Statistique de lSUniversité de* Paris **8** 229–231
12. Joe H 1997 *Multivariate models and multivariate dependence concepts* vol **73** (CRC Press)
13. Nelsen R B 2007 *An Introduction to Copulas* (Springer Series in Statistics)