Journal of the Korean Data & Information Science Society
2009, **20**(5), 933–939

한국데이터정보과학회지

Obtaining bootstrap data for the joint distribution of bivariate survival times[†]

Sehyug Kwon[1]

[1]Department of Statistics, Hannam University
Received 13 July 2009, revised 17 September 2009, accepted 21 September 2009

Abstract

The bivariate data in clinical research fields often has two types of failure times, which are mark variable for the first failure time and the final failure time. This paper showed how to generate bootstrap data to get Bayesian estimation for the joint distribution of bivariate survival times. The observed data was generated by Frank's family and the fake date is simulated with the Gamma prior of survival time. The bootstrap data was obtained by combining the mimic data with the observed data and the simulated fake data from the observed data.

Keywords: Bootstrap, Frank's family, mark variable, mixture prior, simulation.

1. Introduction

The bivariate time data would be obtained in AIDS clinical research. The time of HIV screen test being positive is the beginning of the study. The occurrence of AIDS is the mark time (y) and the death is the final failure time (x) in the successive event. In the blindness of eyes research, the time of blindness of one eye is the mark time and the time of that of the other is the failure time. The estimation of bivariate density of (x, y) has been studied by Lin *et al.* (1993) and Huang and Louis (1998). Lin *et al.* (1993) proposed a simple nonparametric estimator for the bivariate distribution function of the gap times between successive events, mark time and survival time as in our notation, when the survival time is right censored. Lin *et al.* (1993) used Gumbel's bivariate distribution function (1960) with unit exponential for (x, y) and uniform distribution on [0,4] for the distribution of censoring. Huang and Louis (1998) proposed the cumulative mark specific hazard function to estimate the joint distribution function. They did numerical studies with Frank's bivariate family with uniform $[0, 1]$ marginal for mark variable and unit exponential marginal for survival time and the exponential distribution with $\lambda = 2$ for the censoring, which had theoretically the one-third of observed survival time censored.

The notation for the real data are defined as follows: x^0 is the survival time or the failure time, y^0 is the mark time and c is the censoring time. On this setting, the survival time is the follow-up time after mark variable, that is $x^0 > y^0$ and only the failure time is subject to the right censoring. The observed data has the following relation in this setting: $x = \min(x^0, c)$, $\Delta = I(x^0 \leq c)$, and $y = y^0\Delta$.

[†] This paper is supported by 2009 Hannam University Research Fund.
[1] Professor, Department of Statistics, Hannam University, Deajon 306-791, Korea.
E-mail: wolfpack@hnu.ac.kr

2. Simulation for the observed and fake data

In the clinical research, we may easily obtain the data that fit the setting discussed in the previous section and use them for estimating the $F(x, y)$. Since the main purpose of this paper is to show the simulation procedure, we use the data simulated from the Frank's bivariate family (Genest, 1987) for the observed data in the real field. The following steps are how to simulate data from the Frank's bivariate family, which fits the previous setting of (x, y).

1. Draw two observation (u, v) from the uniform distribution on $[0, 1]$.
2. Set $t = \alpha u + (1 - \alpha)v$ where $0 < \alpha < 1$.
3. Define $x^0 = -2ln(1 - u)$ and $y^0 = \log_\alpha[t/(t + (1 - \alpha)v)]$.
4. If $x^0 < y^0$, the simulated observation is deleted because it violates the setting mentioned in the previous section. Moreover, when the failure time is same as the previous simulated observation, it is also deleted for simplicity and continuity. That is, there is no tie in the obtained failure times.
5. Do steps 1.-4. until n observations of (x^0, y^0) are obtained.

Then, the marginal of X is the exponential distribution with $\lambda = 2$ and that of Y is the uniform on $[0, 1]$. Second, the censoring times of size n, c are obtained from an exponential with $\lambda = 8$, which is independent of the joint distribution of (X^0, Y^0). It makes theoretically 20% right censoring. With (x^0, y^0, c), we can obtain the observed data (x, y, Δ) that would be treated the field observed data by the following steps:

1. Set x be the minimum of x^0 and c.
2. Set Δ be 1 if $x^0 \leq c$. Otherwise, $\Delta = 0$.
3. Set $y = y^0\Delta$.

For the size of $n = 50$, the obtained data of (x, y, Δ) are shown in Table 2.1. It would be considered as the observed data from the real field in this paper, which constructs the empirical distribution function, $F_n(x, y)$.

To compute the posterior densities of θ given data (x, y, Δ), the prior density of θ, $\pi(\theta)$ should be assumed with prior information. To use the conjugate property, we can assume $\pi(\theta)$ and (x, y, Δ) as follows:

1. The prior density of θ is Gamma (α, β),

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta},\ 0 \leq \theta < \infty. \tag{2.1}$$

2. The joint density of (X, Y) given θ is

$$f(x, y|\theta) = \theta^2 e^{-\theta x},\ 0 \leq y < x < \infty. \tag{2.2}$$

   Then, the marginal density of x, $f(x)$ is a Gamma$(2, \theta)$ and that of y, $f(y)$ is a exponential(θ).

Table 2.1 The simulated (x, y, Δ) for the observed data

n	x	y	Δ	n	x	y	Δ	n	x	y	Δ
1	1.618	0.775	1	18	2.930	0.920	1	35	0.917	0.000	0
2	0.444	0.329	1	19	0.380	0.162	1	36	0.607	0.156	1
3	2.894	0.927	1	20	1.384	0.545	1	37	0.156	0.029	1
4	1.982	0.664	1	21	4.533	0.948	1	38	1.685	0.546	1
5	0.323	0.075	1	22	2.417	0.778	1	39	2.313	0.814	1
6	0.773	0.000	0	23	0.321	0.139	1	40	5.551	0.000	0
7	1.642	0.424	1	24	1.526	0.450	1	41	0.273	0.098	1
8	4.096	0.963	1	25	1.502	0.956	1	42	0.875	0.000	0
9	3.405	0.558	1	26	0.482	0.181	1	43	0.434	0.293	1
10	5.326	0.933	1	27	0.894	0.402	1	44	2.603	0.746	1
11	0.133	0.047	1	28	2.325	0.000	0	45	2.781	0.000	0
12	2.251	0.837	1	29	0.769	0.166	1	46	0.907	0.384	1
13	1.115	0.442	1	30	1.674	0.230	1	47	3.341	0.976	1
14	4.677	0.942	1	31	0.141	0.028	1	48	0.613	0.103	1
15	1.104	0.000	0	32	1.848	0.000	0	49	3.347	0.714	1
16	0.409	0.207	1	33	1.912	0.000	0	50	1.233	0.473	1
17	1.608	0.000	0	34	0.741	0.324	1				

3. The censoring time, c is distributed exponential(λ) and independent of (x, y). That is, $g(c) = \lambda e^{-\lambda c},\ 0 \le c < \infty$.

With these priors, the posterior density of θ given (x, y, Δ) is computed by

$$\pi(\theta|x, y, \Delta) = \frac{K(d\theta, dx, dy, d\Delta|\theta)\pi(d\theta)}{\int K(d\theta, dx, dy, d\Delta|\theta)\pi(d\theta)d\theta}. \tag{2.3}$$

Huang and Louis (1998) showed

$$H(x, y, d\Delta) = \prod_{i=1}^{n}[F(x, y|\theta)(1 - G(x))]^{\delta_i}[1 - F(x, -|\theta)G(x)]^{1-\delta_i}. \tag{2.4}$$

Therefore, the $K(dx, dy, d\Delta|\theta)$ in the formula (2.3) can be written

$$\prod_{i=1}^{n}[F(x_i, y_i|\theta)(1 - G(x_i))]^{\delta_i}[1 - F(x_i, -|\theta)G(x_i)]^{1-\delta_i}. \tag{2.5}$$

With our prior distributional assumptions and the marginal of x of Gamma$(2, \theta)$, we can derive the followings:

$$F(x, y|\theta)(1 - G(x)) = \theta^2 e^{-\theta x}e^{-\lambda x}, \tag{2.6}$$

$$1 - F(x, -|\theta) = \int_x^\infty \theta^2 s e^{-\theta s}ds = (1 + \theta x)e^{-\theta x}, \tag{2.7}$$

$$\text{and } (1 - F(x, -|\theta))G(x) = (1 + \theta x)e^{-\theta x}\lambda e^{-\lambda x}. \tag{2.8}$$

With these results, the $K(dx, dy, d\Delta|\theta)$ can be written

$$\begin{aligned} K(dx, dy, d\Delta|\theta) &= \prod_i^n [\theta^2 e^{-\theta x_i} e^{-\lambda x_i}]^{\delta_i} [(1+\theta x_i) e^{-\theta x_i} \lambda e^{-\lambda x_i}]^{1-\delta_i} \\ &\propto \prod_i^n [\theta^2 e^{-\theta x_i}]^{\delta_i} [(1+\theta x_i) e^{-\theta x_i}]^{1-\delta_i} \\ &= \theta^{2\sum \delta_i} e^{-\theta \sum x_i I(\delta_i=1)} \prod_i^n [(1+\theta x_i) e^{-\theta x_i}]. \end{aligned} \tag{2.9}$$

Let $T(\theta)$ be $\prod_i^n (1+\theta x_i) e^{-\theta x_i}$. Finally, we can get the posterior density of θ given (x, y, Δ) is as follows:

$$\begin{aligned} \pi(\theta|x, y, \Delta) &\propto \theta^{2\sum_i^n \delta_i} e^{-\theta \sum_i^n x_i I_{(\delta_i=1)}} T(\theta) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{2\sum_i^n \delta_i + \alpha - 1} e^{-\theta(\sum_i^n x_i I_{(\delta_i=1)} + \beta)} T(\theta) \\ &\sim \Gamma(2\sum_i^n \delta_i + \alpha, \sum_i^n x_i I_{(\delta_i=1)} + \beta). \end{aligned} \tag{2.10}$$

The posterior density of θ is used to generate a fake data (x^*, y^*, Δ^*) which is considered the prior guess $F_0(x, y)$ on the parameter $F(x, y)$. The fake data (x^*, y^*, Δ^*) is simulated by the following steps:

1. Draw an observation for θ from the prior density of θ.
2. Draw (x, y) from the joint density (x, y) given θ.
3. Draw the censoring time c independently with (x, y).
4. With the setting between (x, y, c) and (x^*, y^*, Δ^*), compute the value of (x^*, y^*, Δ^*).
5. Do steps 2.-4. m times to obtain $(x_i^*, y_i^*, \Delta_i^*)$ for $i = 1, 2, \ldots, m$.

With the our prior distributional assumptions in the previous section, we show here how to generate the fake data (x^*, y^*, Δ^*). When the joint density of (x, y) is assumed to be $f(x, y|\theta) = \theta^2 e^{-\theta x},\ 0 \le y < x < \infty$, we showed that the marginal density of x is Gamma $(2, \theta)$ and that of y is exponential(θ). We can use these facts to guess the parameters (α, β) of the prior density of θ. For the 40 non-censored observations in Table 2.1, the marginal mean of x_i is computed as 1.75, which calculate $\theta = 1.14$ and that of y_i is 0.39 ($\theta = 2.56$). Therefore, we can assume that the prior density of θ is Gamma($\alpha = 5, \beta = 2.7$) which is bell-shaped and has theoretically the mean of 1.85. From the result (2.10), we can simulate the values of θ by Gamma(85, 70.2) to obtain the fake data (x^*, y^*, Δ^*) since $\sum_i^{50} \delta_i$=50 and $\sum_i^{50} x_i I_{(\delta_i=1)}$=67.5 for the 50 observed data in Table 2.1. The fake data would be used for the prior guess $F_0(x, y)$ on the $F(x, y)$.

The procedures to generate the fake data (x^*, y^*, Δ^*) with the observed data (simulated and shown in Table 2.1 for example) can be summarized as follows:

1. Draw an observation θ from Gamma(85, 70.2).
2. Draw y and z independently from the exponential(θ). Set $x = y + z$ because the conditional density of X given y is $f(x|y) = \theta e^{-\theta(x-y)}$, $y < x < \infty$. We assume that there is no tie in the simulated value of x like the observed (obtained by simulation) data in Table 2.1.
3. When the marginal density of X is Gamma$(2, \theta)$ and the probability density of the censoring time is assumed exponential(λ) as we assumed, the probability of not being censored, $P(x \le c)$ is $(\theta/(\theta+\lambda))^2$. For our observed data in Table 2.1, the probability of being censored is 0.2. Therefore, for our observed data, we can draw the censoring time c from the exponential($0.118\theta^*$) that satisfies $(\theta/(\theta+\lambda))^2 = 0.8$.
4. Set $\Delta^* = 1$ if $x \le c$. Otherwise, set $\Delta^* = 0$. Set x^* be the minimum value of x and c and $y^* = y\Delta^*$.
5. Do steps 2.-4. m times to obtain (x^*, y^*, Δ^*) for $i = 1, 2, \ldots, m$.

The simulated fake data of size 50 with the previous steps is shown in Table 2.2. The size of fake data, m depends on the size of bootstrap data which will be discussed in the next section.

Table 2.2 The simulated fake data of (x^*, y^*, Δ^*)

n	x	y	Δ	n	x	y	Δ	n	x	y	Δ
1	2.000	1.421	1	18	2.406	0.252	1	35	0.865	0.155	1
2	0.155	0.068	1	19	0.796	0.525	1	36	2.859	2.692	1
3	1.370	0.000	0	20	1.547	1.487	1	37	1.799	0.789	1
4	1.938	0.624	1	21	1.221	0.000	0	38	1.291	0.284	1
5	0.301	0.000	0	22	1.991	1.367	1	39	0.178	0.007	1
6	1.724	1.062	1	23	1.185	0.053	1	40	0.872	0.611	1
7	0.572	0.029	1	24	0.108	0.000	0	41	0.432	0.000	0
8	1.265	0.918	1	25	1.670	0.179	1	42	0.330	0.000	0
9	0.810	0.310	1	26	3.691	0.000	0	43	1.274	0.708	1
10	0.687	0.000	0	27	1.674	0.617	1	44	3.811	0.132	1
11	2.848	1.105	1	28	2.617	0.849	1	45	1.623	0.796	1
12	1.151	1.045	1	29	1.243	0.939	1	46	1.129	1.074	1
13	1.396	1.223	1	30	1.300	0.251	1	47	1.782	1.350	1
14	0.993	0.197	1	31	0.108	0.065	1	48	2.011	0.547	1
15	0.497	0.043	1	32	2.401	1.076	1	49	1.081	0.002	1
16	1.430	0.273	1	33	1.309	1.157	1	50	1.925	1.587	1
17	1.022	0.001	1	34	0.158	0.016	1				

3. Bootstrap data and discussion

Now, we can generate $(x, y, \delta)^b$ given θ by using the bootstrap technique shown in Muliere and Secchi (1996) to estimate the $F(x, y)$. The observations $(x, y, \delta)^b$ are generated from $(n+\alpha)^{-1}(\alpha F_0 + nF_n)$ where F_n is the empirical distribution of the observed data and F_0 is the proper distribution from the prior guess on F. The value of α would be the degree of faith in the guess on F and the fake data (x_0, y_0, Δ_0) discussed in the previous section would be used for F_0.

Fist of all, we should have mimic data for the observed data (x, y, Δ) if we need to select more than n observations from the observed data to have bootstrap data. Since the fake data (x_0, y_0, Δ_0) is simulated with some prior distributional assumptions, we can have enough the fake observations that the bootstrap data is obtained without selecting the same observation from the fake date more that once. But, the size of the observed data may not be enough big to generate the bootstrap data without replacement, and there would be tied values in the failure time x of the bootstrap data. Therefore, we need to mimic more observed data by using the observed data. Suppose we generate bootstrap data of size b with the prior degree α. Then, we select observations from the observed data theoretically with the probability of $(n+\alpha)^{-1}\alpha$ to obtain the bootstrap data. That is, $b(n+\alpha)^{-1}\alpha$ observations of the mimic data are obtained from the observed data. Suppose that $b(n+\alpha)^{-1}\alpha = 3n$ where n is the size of the observed data (x, y, Δ). The first s observations of the mimic data is the observed data (x, y, Δ). The second n observations of the mimic data is obtained from the observed data by $(x+u, y, \Delta)$ where u is randomly assigned by u_1 or u_2 with the probability of 0.5. The values of u_1 and u_2 depends on the scale of the observed x, which would be are respectively -0.0005 and 0.0005 which would be set the half of the last digit of data. The final n observations for the mimic data would be $(x + u, y, \Delta)$ where the value of u is selected in the second stage. If we need more than $3n$, the value of u is randomly selected from $(u_1, u_2, ..., u_p)$ with the same probability of $1/p$ in the first stage. In the next stage, we can use the value of u which is randomly selected from u_i except the selected value in the previous stage with the same probability of $1/(p-1)$. We do the similar step until the mimic data of desired size is obtained. With the mimic data of the observed data and the fake data, we generate a bootstrap data of size b by the following steps:

1. Choose an value of α, the degree of faith on the prior information. If we assign $\alpha = 0$ (no prior), the bootstrap results would be the same as the result of the Frequentists.

2. Generate an observation w from Bernoulli$(\alpha/(n + \alpha))$.

3. Choose an observation either from the fake data (F_0) in Table 2.1 if $w = 1$ or from the observed data (F_n) in Table 2.2 if $w = 0$.

4. Do steps 2.-3. b times to have $(x, y, \delta)^b$ for $i = 1, 2, \ldots, b$.

We show how to generate the bootstrap data of size 200 when the observed data of size n in Table 2.1 is 50 the prior guess α is 10. The bootstrap data would be obtained from either the fake data with the probability of 1/6 or the mimic data of the observed data with the probability of 5/6. That is, we need the mimic data of at least size 167 and the fake data of al least size 34. Hence, we generate the fake data of size 50 and the mimic data of size 200. For making the mimic data of size 200 from the 50 observed data, the following steps is used:

1. The first 50 observation of the mimic data is exactly same as the observed data in Table 2.1.

2. Select an observation u randomly from (u_1=0.00075, u_2=0.0005, u_3=0.00025, u_4=0.000 25, u_5=0.0005, u_6=0.00075) with the probability of 1/6 and make the second 50 observation of the mimic data by $(x + u, y, \Delta)$.

3. Select an observation u randomly from the values of u except the selected value in the Step 2 with the probability of 1/5 and make the second 50 observation of the mimic data by $(x+u, y, \Delta)$.

4. The final 50 observations of the mimic data is obtained by $(x+u, y, \Delta)$ using the value of u selected independently from the values of u except the values of pre-selected u_i in the Step 2 and 3 with the probability of 1/4.

For example, a bootstrap data $(x, y, \delta)^b$ of size $b = 200$ with $\alpha = 10$ when a value of θ^* is selected from $Gamma(85, 70.2)$. That is, 200 observations of $(x, y, \delta)^b$ from $(1/6)F_0+(5/6)F_n$ are generated where F_n is the mimic data of the observed data in Table 2.1 and F_0 is the fake data in Table 2.2. The bootstrap data $(x, y, \delta)^b$ would be used to estimate $F(x, y)$, the joint distribution of two survival times.

References

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, **74**, 549-555.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, **55**, 698-707.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*, **76**, 751-761.

Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, **86**, 573-581.

Muliere, P. and Secchi, P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics*, **48**, 663-673.