Bivariate lifetime geometric distribution in presence of cure fractions

Nasser Davarzani[1*], Jorge Alberto Achcar[2], Evgueni Nikolaevich Smirnov[1], Ralf Peeters[1]

[1] Department of Knowledge Engineering, Maastricht, the Netherlands.
[2]Departamento de Medicina Social, FMRP, Universidade de Sao Paulo,
Ribeirao Preto, SP, Brazil.

Abstract: In this paper, we introduce a Bayesian analysis for bivariate geometric distributions applied to lifetime data in the presence of covariates, censored data and cure fraction using Markov Chain Monte Carlo (MCMC) methods. We show that the use of a discrete bivariate geometric distribution could bring us some computational advantages when compared to standard existing bivariate exponential lifetime distributions introduced in the literature assuming continuous lifetime data as for example, the exponential Block and Basu bivariate distribution. Posterior summaries of interest are obtained using the popular OpenBUGS software. A numerical illustration is introduced considering a medical data set related to the analysis of a diabetic retinopathy data set.

Key words: Bivariate geometric distribution, Censored data, Cure function, Bayesian methods.

1. Introduction

In medical, engineering or other lifetime data applications, we could have more than one lifetime associated to each unit. A special situation is the presence of two lifetimes X_1 and X_2 associated to each unit. This is the case, for example, considering X_1 and X_2 the timing of failure of paired organs like kidney, lungs, eyes, ears, dental implants among many others. In this situation, we could consider some existing bivariate lifetime distributions that has been introduced in the literature for continuous bivariate lifetime data (see for example, Arnold and Strauss, 1988; Marshall and Olkin, 1967; Sarkar, 1987; Block and Basu, 1974).

Usually these bivariate lifetime distributions generalize some popular existing lifetime distributions as exponential, Weibull, gamma or a log-normal distributions (see for example, Lawless, 1982).

Other parametrical lifetime distributions also could be generalized for the bivariate case. One of these models is given by the generalized exponential distribution (see for example, Gupta and Kundu, 2007, Mohsin et al., 2013). A bivariate generalized exponential distribution also was introduced by Kundu and Gupta (2008).

Among all these continuous bivariate lifetime distributions, one model has been very well explored in the literature: the Block and Basu (1974) bivariate exponential distribution, whose marginals are mixtures of exponentials and having an absolutely continuous joint distribution with positive correlation coefficient. This distribution was derived originally by omitting the singular part of the Marshall and Olkin (1967) distribution. The Marshall and Olkin distribution is not absolutely continuous but satisfies the lack of memory property.

It is important to point out that censoring is a key aspect of survival analysis. In this way, many recent papers related to the analysis of bivariate lifetime data are related to the presence of censored data (see for example, Hanagal, 2010; Hanagal and Ahmadi, 2008; Davarzani and Parsian, 2010; Santos and Achcar, 2011; Davarzani et al., 2014).

As an alternative for the use of bivariate parametrical distributions for bivariate continuous lifetime data, we could assume that the lifetimes X_1 and X_2 are discrete random variables assuming any positive integer that is, we round the lifetimes given with decimal part to an integer. It is important to point out, that despite discrete measuring of medical or engineering lifetime data is very common in applications, few papers related to discrete lifetime data are introduced in the literature, (see for example, Grimshaw and et al., 2005; Davarzani and Parsian, 2013).

In this way, Arnold (1975) introduced a general multivariate geometric distribution which showed that it leads in a natural way to the Marshall-Olkin multivariate exponential distribution. Nair and Nair (1988) studied the characterizations of bivariate exponential distributions and geometric distributions.

In this paper, we explore the use of a bivariate geometric distribution to analyze bivariate lifetime data. The bivariate geometric distribution proposed by Arnold (1975) has joint mass probability given by,

$$P(x_1, x_2) = \begin{cases} P_1(x_1, x_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{x_1-1}(1-\theta_2)^{x_2-x_1-1} & x_1 < x_2 \\ 0 & x_1 = x_2 \\ P_2(x_1, x_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{x_2-1}(1-\theta_1)^{x_1-x_2-1} & x_1 > x_2, \end{cases} \tag{1}$$

where the marginal probability mass functions are respectively, given by standard geometric distributions starting at one, that is,

$$p(x_1) = (1-\theta_1)^{x_1-1}\theta_1,\ x_1 = 1,\ 2,\ 3, \ldots$$

and

$$p(x_2) = (1-\theta_2)^{x_2-1}\theta_2,\ x_2 = 1,\ 2,\ 3, \ldots$$

which means, variances, covariance and correlation given respectively, by

$$\mu_1 = E(X_1) = \frac{1}{\theta_1},\ \mu_2 = E(X_2) = \frac{1}{\theta_2},$$

$$\sigma_1^2 = Var(X_1) = \frac{1-\theta_1}{\theta_1^2},\ \sigma_2^2 = Var(X_2) = \frac{1-\theta_2}{\theta_2^2},$$

$$Cov(X_1,\ X_2) = \frac{-1}{1-r},$$

$$\rho_{12} = Corr(X_1,\, X_2) = -\frac{\theta_1\theta_2}{(1-r)[(1-\theta_1)(1-\theta_2)]^{0.5}},$$

where $r = 1 - \theta_1 - \theta_2$; $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$.

From (1), we observe that this bivariate discrete model assumes that P ($X_1 = X_2$) = 0, but in applications we could have this probability very small, but different of zero, a restrictive property of the model. In some cases, however, like studying the time of blindness of eyes or infection of kidneys, this probability practically is zero. In such cases the proposed model will be more fit to the data rather that the models with positive probability of $X_1 = X_2$.

In the presence of right censored data and covariates, a common situation in applications, we could have some computational difficulties to get standard classical inferences for the parameters of bivariate lifetime distributions, as for example, using maximum likelihood approach. As an alternative for the use of standard classical inference methods, the use of a Bayesian approach is becoming very popular in the analysis of bivariate lifetime data (see for example, Achcar and Leandro, 1998; or Santos and Achcar, 2011).

This paper is organized as follows: in section 2, we introduce the presence of censored data; in section 3, we introduce the presence of cure fraction; in section 4, we introduce an analysis of a diabetic retinopathy data set; finally, in section 5, we present some concluding remarks.

2. Presence of Censored Data

Let (X_{11}, X_{21}), … , (X_{1n}, X_{2n}) be a bivariate random sample of size n from ordinary bivariate geometric (OBG) distribution with joint (pmf) in (1). Suppose that we have the following situations: X_1 and X_2 are complete observations, X_1 and X_2 could be either censored or both X_1 and X_2 are censored at Y_1 or Y_2 , respectively, and that censoring is independent of the lifetimes. Let us subdivide the *n* observations into four classes:

C$_1$: $X_{1i} < Y_{1i}$ and $X_{2i} < Y_{2i}$, thus both x_{1i} or x_{2i} are the real lifetimes;
C$_2$: $X_{1i} < Y_{1i}$ and $Y_{2i} < X_{2i}$, that is, we observe x_{1i} and y_{2i} ;
C$_3$: $Y_{1i} < X_{1i}$ and $X_{2i} < Y_{2i}$, we observe y_{1i} and x_{2i} ;
C$_4$: $Y_{1i} < X_{1i}$ and $Y_{2i} < X_{2i}$, we observe y_{1i} and y_{2i} ;

Now based on the above definitions, the likelihood function of bivariate right censored lifetime in this model is given by

$$L(\theta_1,\,\theta_2) = \prod_{i\in C_1} P(x_{1i},x_{2i}) \prod_{i\in C_2}\left(\sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i})\right) \prod_{i\in C_3}\left(\sum_{x_{1i}=y_{1i}+1}^{\infty} P(x_{1i},x_{2i})\right)$$
$$\prod_{i\in C_4}\left(\sum_{x_{1i}=y_{1i}+1}^{\infty}\ \sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i})\right) \qquad (2)$$

where,

$$\sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i}) = \begin{cases} \theta_1(1-\theta_1-\theta_2)^{x_{1i}-1}(1-\theta_2)^{y_{2i}-x_{1i}} & if \;\; x_{1i} < x_{2i} \\ \theta_1(1-\theta_1-\theta_2)^{y_{2i}}(1-\theta_1)^{x_{1i}-y_{2i}-1} & if \;\; x_{1i} > x_{2i}\,, \end{cases}$$

and

$$\sum_{x_{1i}=y_{1i}+1}^{\infty} P(x_{1i},x_{2i}) = \begin{cases} \theta_2(1-\theta_2)^{x_{2i}-y_{1i}-1}(1-\theta_1-\theta_2)^{y_{1i}} & if \;\; x_{1i} < x_{2i} \\ \theta_2(1-\theta_1)^{y_{1i}-x_{2i}}(1-\theta_1-\theta_2)^{x_{2i}-1} & if \;\; x_{1i} > x_{2i}\,, \end{cases}$$

and

$$\sum_{x_{1i}=y_{1i}+1}^{\infty} \sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i}) = \begin{cases} (1-\theta_2)^{y_{2i}-y_{1i}}(1-\theta_1-\theta_2)^{y_{1i}} & if \;\; x_{1i} < x_{2i} \\ (1-\theta_1)^{y_{1i}-y_{2i}}(1-\theta_1-\theta_2)^{y_{2i}} & if \;\; x_{1i} > x_{2i}. \end{cases}$$

To simulate samples for the joint posterior distribution for θ_1, θ_2 and r we use MCMC methods (see for example, Gelfand and Smith, 1990). In this way we could simulate samples for the joint posterior distribution from the conditional distributions $g(\theta_1 | \theta_2, r, x_1, x_2)$, $g(\theta_2 | \theta_1, r, x_1, x_2)$, and $g(r | \theta_1, \theta_2, x_1, x_2)$, where $x_1 = (x_{11}, x_{12}, ..., x_{1n})$, $x_2 = (x_{21}, x_{22}, ..., x_{2n})$ using the Gibbs sampling algorithm or the Metropolis-Hastings algorithm.

In the presence of a covariate vector $z = (z_1, z_2, ..., z_p)$ associated to each bivariate lifetime (X1, X2), we could assume the following regression model:

$$\theta_{1i} = \frac{\exp\{\beta_1' z_i\}}{1+\exp\{\beta_1' z_i\}},$$

$$\theta_{2i} = \frac{\exp\{\beta_2' z_i\}}{1+\exp\{\beta_2' z_i\}}, \qquad (3)$$

where, $\beta_j' = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp}), j = 1, 2,$ is the regression parameter vector and $z_i = (z_{1i}, z_{2i}, \ldots, z_{pi}), i = 1, \ldots n$. Observe that the assumed model (3) guarantees that $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$. We also assume Normal prior distributions for the regression coefficient β_{j1}. Notice that the regression parameters are defined for all real values which justifies the choice of normal prior.

3. Presence of cure fraction

In lifetime data analysis the presence of censored observations is very common in applications. The censored data could be related to individuals lost to follow-up or that will never experience the event of interest. This situation could occur in different areas, such as in cancer studies where the researchers are interested in the proportion of patients cured and where many individuals will die of other causes. In practical work we should carefully interpret the results using a cure fraction model when this situation is not true. As an example of this case, disease recurrence in primary breast cancer; even after many years, the condition can recur. A detailed discussion on the presence of cure fraction is introduced by Lambert (2007). As pointed out by this author, information of cure at the individual level will rarely be available, and so in these models we are concerned with population (or statistical) cure. If reliable information on cause of death is available, then one can perform a cause-specific analysis where deaths not due to the disease of interest can be treated as censored observations. In many situations, the cause of death may either not be recorded or obtained from death certificates, which are often inaccurately recorded (Begg and Schrag, 2002). A possibility in this case: to obtain the expected survival and/or the expected mortality rate from national mortality statistics, and such is usually calculated after matching for age, sex, year of diagnosis, and possibly other covariates (Coleman et al,1999).

In this situation, after a careful preliminary data analysis, we could have a fraction of individuals not expecting to experience the event of interest, that is, these individuals are not at risk (”long term survivors” or ”cured individuals”). Different approaches have been considered to model cure fraction, especially for univariate lifetime data (see for example, Yu et al, 2004; Gamel et al, 1999; Yamaguchi, 1992; Cancho and Bolfarine, 2001; Taylor, 1995; Kannan et al, 1999).

Wienke et al (2006) introduced a model for a cure fraction in bivariate time-to-event data. Let us assume that the population is divided in two groups of individuals: a group of cured individuals with probability $1 - \phi$ and a group of susceptible individuals with a proper survival function $S(x) = P(X > x)$, where X denotes the discrete lifetime of the individual with probability ϕ. In this way, we have a survival function equals to one for all X considering the cured individuals.

Considering univariate lifetimes, a model that incorporates a cure fraction (see for example, Wienke et al., 2006) is given by,

$$S(x) = p + (1-p)\,S_0(x), \tag{4}$$

where $p \in (0, 1)$ is the mixing parameter and $S_0(x)$ denotes a proper survival function for the non-cured group in the population.

A generalization of (3) considering bivariate lifetimes X_1 and X_2 is (see Wienke et al., 2006) given by,

$$\begin{aligned} \mathrm{S}(x_1,\, x_2) &= P(X_1 > x_1,\, X_2 > x_2) \\ &= \phi_{11}\,\mathrm{S}(x_1, x_2) + \phi_{10}\,\mathrm{S}_{10}(x_1) + \phi_{01}\,\mathrm{S}_{20}(x_2) + \phi_{00} \end{aligned} \tag{5}$$

where $S(x_1, x_2) = P(X_1 > x1, X_2 > x_2)$ is the joint survival function; $S(x_1) = P(X_1 > x_1)$ is the marginal survival function for X_1; $S(x_2) = P(X_2 > x_2)$ is the marginal survival function for X_2; $\phi_{11} = P(V_1 = 1, V_2 = 1)$; $\phi_{10} = P(V_1 = 1, V_2 = 0)$; $\phi_{01} = P(V_1 = 0, V_2 = 1)$; $\phi_{00} = P(V_1 = 0, V_2 = 0)$; $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$; V_1 and V_2 are binary variables such that $V_1 = 1$ if the individual is susceptible for lifetime X_1 and that $V_1 = 0$ if the individual is immune; in the same way, $V_2 = 1$ if the individual is susceptible for lifetime X_2 and that $V_2 = 0$ if the individual is immune.

From the above definitions, the likelihood function (2) in presence of cure function and right censoring is given by,

$$L(\theta_1,\,\theta_2) = \prod_{i\in C_1} P_1'(x_{1i},x_{2i}) \prod_{i\in C_2} P_2'(x_{1i},x_{2i}) \prod_{i\in C_3} P_3'(x_{1i},x_{2i}) \prod_{i\in C_4} P_4'(x_{1i},x_{2i}),$$

where

$$P_1'(x_{1i},x_{2i}) = \phi_{11}\, P(x_{1i},x_{2i}),\; i\in C_1,$$

and

$$P_2'(x_{1i},x_{2i}) = \phi_{11} \sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i}) + \phi_{10}\, p_1(x_{1i}),\; i\in C_2,$$

and

$$P_3'(x_{1i},x_{2i}) = \phi_{11} \sum_{x_{1i}=y_{1i}+1}^{\infty} P(x_{1i},x_{2i}) + \phi_{10}\, p_2(x_{2i}),,\; i\in C_3,$$

and

$$P_4'(x_{1i},x_{2i}) = \phi_{11} \sum_{x_{1i}=y_{1i}+1}^{\infty}\sum_{x_{2i}=y_{2i}+1}^{\infty} P(x_{1i},x_{2i}) + \phi_{10}S_1(x_{1i}) + \phi_{01}S_2(x_{2i}) + \phi_{00},\; i\in C_4,$$

where,

$$p_1(x_{1i}) = \theta_1\,(1-\theta_1)^{x_{1i}-1},\; x_{1i}=1,\,2,\ldots,\; p_2(x_{2i}) = \theta_2\,(1-\theta_2)^{x_{2i}-1},\; x_{2i}=1,\,2,\ldots,$$

and

$$S_1(x_{1i}) = (1-\theta_1)^{x_{1i}},\quad S_2(x_{2i}) = (1-\theta_2)^{x_{2i}}$$

For a Bayesian analysis, we assume a Dirichlet prior for the parameter ϕ_{11}, ϕ_{10} , ϕ_{01} and ϕ_{00}. The Dirichlet prior is used since $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$.

4. Analysis of a diabetic retinopathy data set

In this application, we consider a data set introduced by Huster et al (1989). In this study, we have 197 patients with ”high-risk” diabetic retinopathy as defined by the Diabetic Retinopathy Study (DRS). Each patient had one eye randomized to laser treatment and the other eye received no treatment. For each eye, the event of interest was the time from initiation of treatment to the time when visual acuity dropped below 5/200 two visits in a row (call it ”blindness”). Thus there is a builtin lag time of approximately 6 months (visits were every 3 months). Survival times in this dataset are therefore the actual time to blindness in months, minus the minimum possible time to event (6.5 months). Censoring was caused by death, dropout, or end of the study (Huster et al, 1989). Associated to each individual we have two covariates: Z1 denoting the age at diagnosis of diabetes and Z2 denoting the type of diabetes where 1= juvenile (age at diagnosis < 20) and 0 = for adults.

4.1 A preliminary data analysis

In Figure 1, we have the plots of the non-parametrical Kaplan and Meier (1958) estimators of the survival functions for both times (X1 = time to blindness for the treated eye and X2 = time to blindness for the untreated eye).The non-parametrical estimates for the means (MTTF) of both survival functions obtained from the Kaplan-Meier estimates are given, respectively, by 53.7284 (for X1, with 54 uncensored observations and 143 right censored observations) and 43.5258 (for X2 , with 101 uncensored observations and 96 right censored observations).

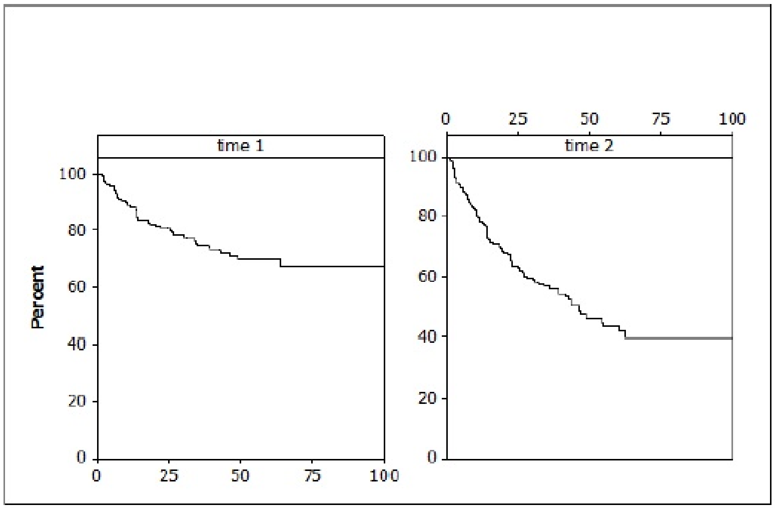


Figure 1: Kaplan-Meier estimates for the survival function of X_1 (time1) and X_2 (time2)-times in months.

From the plots of Figure 1, we observe some indication of cure fraction, or an indication that a great number of individuals that will never become blind.

4.2 A Bayesian analysis

To analyse the diabetic retinopathy data set under a Bayesian approach, let us first assume the bivariate geometric distribution with density (1) since both responses (X_1 = time to blindness for the treated eye and X_2 = time to blindness for the untreated eye) are associated to the same individual. Observe that in this case, we are transforming both continuous lifetimes to discrete lifetimes to assume a bivariate geometric distribution. As a first analysis, we do not consider the presence of the covariates Z_1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1= juvenile (age at diagnosis < 20) and 0 = for adults and not considering the presence of cure fraction. Let us denote this model as ”model 1”.

For a Bayesian analysis of ”model 1”, let us assume a Dirichlet prior distribution for θ_1, θ_2 and r with hyper parameter values $\alpha_1 = \alpha_2 = \alpha_3 = 1$, given by;

$$\pi(\theta_1, \theta_2) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} (1 - \theta_1 - \theta_2)^{\alpha_3 - 1} \quad , \; \theta_1 + \theta_2 < 1, \tag{6}$$

Using the OpenBUGS software (Spiegelhalter et al, 2003) in the simulation of samples of the joint posterior distribution for θ_1, θ_2 and r we discarded the first 10,000 simulated Gibbs samples (burn-in-samples) to eliminate the effect of the initial values for the parameters of the model. Choosing every 50th simulated Gibbs sample, we obtained a final sample of size 1,000 to get the posterior summaries of interest. Convergence of the Gibbs sampling algorithm was monitored using existing methods as time series plots for the simulated samples.

Table 1: Posterior summaries “model 1”.

Parameter	Mean	S.D.	95% Credible interval
μ_1	145.2	18.98	(111.8; 187.6)
μ_2	63.43	6.246	(52.5; 77.41)
r	0.9771	0.001795	(0.9735; 0.9803)
σ_1	144.7	18.98	(111.3; 187.1)
σ_2	62.93	6.246	(52.0; 76.91)
θ_1	0.007003	9.11E-4	(0.005339; 0.008946)
θ_2	0.01592	0.001543	(0.01293; 0.01905)
ρ_{12}	-0.004892	4,71E-01	(-0.005859;-0.004021)

In Table 1, we have the posterior means, the posterior standard-deviations and 95% credible intervals for the parameters for θ_1, θ_2 and r. We also have in Table 1, the posterior summaries for the means, standard-deviations and for the correlation coefficient (see (1)) for the lifetimes X_1 and X_2.

In the presence of the covariates Z1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1=juvenile (age at diagnosis < 20) and 0 = for adults, we now assume the bivariate geometric distribution with density (1) and the following logistic regression model for the parameters θ_1 and θ_2, given by,

$$logit(\theta_{1i}) = \beta_{10} + \beta_{11}(Z_{1i} - \bar{Z}_1) + \beta_{12}\, Z_{2i}$$

$$logit(\theta_{2i}) = \beta_{20} + \beta_{21}(Z_{2i} - \bar{Z}_2) + \beta_{22}\, Z_{2i} \qquad (7)$$

where $i = 1, 2, ..., 197$; $logit(\theta) = log[\theta/(1 - \theta)]$, $\bar{Z}_1$ is the average of the covariate age. Let us denote this model as”model 2”. Observe that $r_i = 1 - \theta_{1i} - \theta_{2i}$. Assuming normal prior distributions N (0, 10) for the regression parameters $\beta_{j0}, j = 1, 2,$ and normal prior distributions $N (0,1)$ for the regression parameters βjl , j = 1, 2; l = 1, 2, and using the OpenBUGS software also considering a ”burn-in-sample” of size 10,000 and final sample of size 1,000 taking every 50th sample, we have in Table 2, the posterior summaries of interest.

It is important to point out that using the approximately normal non-informative $N(0,10)$ for the regression parameters $\beta_{j0}, j = 1, 2,$ and $N (0, 1)$ for the regression parameters β_{jl} , $j = 1, 2; l = 1, 2,$ using the OpenBugs software, we obtained very good convergence of the MCMC simulation algorithm as observed in trace plots of the simulated Gibbs samples. Other very non-informative priors also have been considered as normal $N (0, 10)$ priors for all regression parameters but the obtained posterior summaries were very close to the obtained results given in Table 2, that is, the obtained posterior summaries are not sensible to the choice of other priors, leading to the same inference results.

Table 2: Posterior summaries “model 2”.

Parameter	Mean	S.D.	95% Credible interval
β_{10}	-5.233	0.2265	(-5.69; -4.796)
β_{11}	0.004975	1.0	(-1.986; 1.964)
β_{12}	0.4637	0.2793	(-0.04958; 1.02)
β_{20}	-3.925	0.1394	(-4.209; -3.667)
β_{21}	-0.0203	1.021	(-2.007; 2.007)
β_{22}	-0.3741	0.1999	(-0.7607; 0.01117)

From the results of Table 2, we observe that zero is included in the 95% credible interval for β_{11}, β_{12}, β_{21}, and β_{22}, that is, the lifetime X_1 and X_2 are not affected by the covariates Z_1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1=juvenile (age at diagnosis < 20) and 0 = for adults.

As a third modeling approach, not considering the presence of the covariates Z_1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1= juvenile (age at diagnosis < 20) and 0 = for adults, we assume the bivariate geometric distribution with density (1) considering the presence of cure fraction (see, section 3). Let us denote this model as”model 3”.

Table 3: Poterior summaries “model 3”.

ϕ_{00}	0.04129	0.03845	(0.001019; 0.1384)
ϕ_{01}	0.005489	0.005341	(0.0017; 0.02077)
ϕ_{10}	0.6719	0.05193	(0.5595; 0.7588)
ϕ_{11}	0.2819	0.03815	(0.2135; 0.361)
r	0.9391	0.005585	(0.9275; 0.9497)
θ_1	0.01134	0.001922	(0.008293; 0.01567)
θ_2	0.04958	0.005232	(0.03963; 0.06071)

For a Bayesian analysis of ”model 3”, let us assume a Dirichlet prior distribution (6) for θ_1 , θ_2 and r with hiperparameter values $\alpha_1 = \alpha_2 = \alpha_3 = 1$ and another Dirichlet prior distribution for ϕ_{11}, ϕ_{10} , ϕ_{01} and ϕ_{00} with hyper-parameter values $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$. Note that with these values of hyper-parameters, we are assuming approximately non-informative priors for θ_1, θ_2 and r and for ϕ_{11}, ϕ_{10} , ϕ_{01} and ϕ_{00} .Using the OpenBUGS software also considering a ”burn-in-sample” of size 10,000 and final Gibbs sample of size 1,000 taking every 50th sample, we have in Table 3, the posterior summaries of interest.

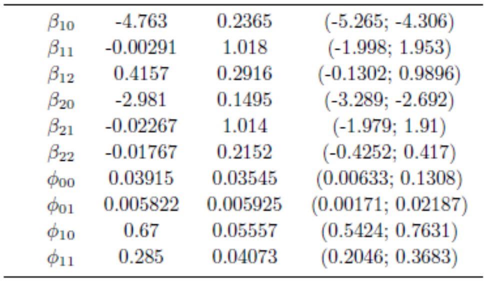
From the results of Table 3, we observe that the Monte Carlo estimate for the posterior mean for ϕ_{10} (probability of observations pairs with $V_1 = 1$ for susceptible and $V_2 = 0$ for immune) based on the 1,000 simulated Gibbs samples is approximately equal to 0.67 and for ϕ11 (probability of $V_1 = 1$ and $V_2 = 1$) is approximately equal to 0.28 with large 95% credible intervals. It is important to point out that more accurate Bayesian inferences could be obtained using more informative priors for the parameters ϕ_0 , ϕ_{01}, ϕ_{10} and ϕ_{11} based on expert opinions or using empirical Bayesian methods.

Finally, considering the presence of the covariates Z_1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1 = juvenile (age at diagnosis < 20) and 0 = for adults, we assume the bivariate geometric distribution with density (1) and regression model (3) considering the presence of cure fraction (see, section 3). Let us denote this model as ”model 4”.

For a Bayesian analysis of ”model 4”, let us assume a Dirichlet prior distribution for ϕ_{11}, ϕ_{10}, ϕ_{01} and ϕ_{00} with hyper parameter values $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$ and the regression model (3) with the following priors for the regression parameters: $\beta_{10} \sim N(-5, 1)$, $\beta_{11} \sim N(0.005, 1)$, $\beta_{12} \sim N(0.5, 1)$, $\beta_{20} \sim N(-4, 1)$, $\beta_{21} \sim N(-0.02, 1)$ and $\beta_{22} \sim N(-0.4, 1)$. Observe that we are using informative priors for the regression parameters based on the information obtained from ″model 2″ (Use of empirical Bayesian methods, see for example, Carlin and Louis, 2000.)

Using the OpenBUGS software also considering a ”burn-in-sample” of size 10,000 and a final Gibbs sample of size 1,000 taking every 50th sample, we haveing Table 4, the posterior summaries of interest.

Table 4: Posterior summaries “model 4”



β_{10}	-4.763	0.2365	(-5.265; -4.306)
β_{11}	-0.00291	1.018	(-1.998; 1.953)
β_{12}	0.4157	0.2916	(-0.1302; 0.9896)
β_{20}	-2.981	0.1495	(-3.289; -2.692)
β_{21}	-0.02267	1.014	(-1.979; 1.91)
β_{22}	-0.01767	0.2152	(-0.4252; 0.417)
ϕ_{00}	0.03915	0.03545	(0.00633; 0.1308)
ϕ_{01}	0.005822	0.005925	(0.00171; 0.02187)
ϕ_{10}	0.67	0.05557	(0.5424; 0.7631)
ϕ_{11}	0.285	0.04073	(0.2046; 0.3683)

From the results of Table 4, we also observe that zero is included in the 95% credible interval for β_{11} , β_{12} , β_{21}, and β_{22} , that is, the lifetime X_1 and X_2 are not affected by the covariates Z_1 denoting the age at diagnosis of diabetes and Z_2 denoting the type of diabetes where 1= juvenile (age at diagnosis < 20) and 0 = for adults. Similar Bayesian inferences were obtained for the parameters ϕ_{00}, ϕ_{01}, ϕ_{10} and ϕ_{11} (see Table 3).

As a comparison for the four proposed models, we could use a Bayesian discrimination criterion, as for example, the DIC (Deviance Information Criterion) introduced by Spiegelhalter et al. (2002), (see the appendix) and given automatically by the OpenBUGS software. Assuming ”model 1” (bivariate geometric distribution without the presence of covariates and cure fraction), we have a Monte Carlo estimate for DIC given by the value 1684.0; assuming ”model 2” (bivariate geometric distribution in presence of covariates but not considering presence of cure fraction) the DIC value is given by 1679.0; assuming ”model 3” (bivariate geometric distribution not considering the presence of covariates but considering the presence of cure fraction) the DIC value is given by 1553.0 and assuming ”model 4” (bivariate geometric distribution considering the presence of covariates and the presence of cure fraction) the DIC value is given by 1556.0. That is, ”model 3” and ”model 4” (presence of cure fraction) give better fit for the diabetic retinopathy data set (smaller value for DIC). Since both covariates do not show significant effects on the lifetimes X_1 and X_2, and the DIC values for ”model 3” and ”model 4” are very similar, we could conclude that ”model 3” is the best model to be fitted by the data set.

5. Concluding remarks

The search for new lifetime models is of great interest for researchers in statistics and in different areas of applications as medical or engineering studies. In some situations, we have bivariate lifetime data, that is, two lifetimes are measured for each unit where it is important, to have a model that captures the possible dependence between both responses. Assuming continuous lifetimes, the literature presents many bivariate parametrical distributions, as for example, the popular Block-Basu (1974) and the Marshall Olkin (1967) distributions among many others. A Bayesian analysis of the Block and Basu distribution in presence of cure fraction is introduced by Achcar et al., 2013. As an alternative, we propose the use of a bivariate geometric distribution which could be a new and suitable alternative for the analysis of bivariate lifetime data, especially under a Bayesian approach and using MCMC methods. The use of discrete probability distributions still is not well explored in the literature for the analysis of lifetime data. This family of distributions could bring us some improvements in the computational work to obtain the inferences of interest. In many applications, especially in medical studies, we could have the presence of censored data, the presence of one or more covariates and the presence of cure fraction, a situation that could bring us great computational difficulties to get the usual classical inferences for the parameters of the proposed model like the convergence of the iterative numerical algorithm used to get the maximum likelihood estimators for the parameters of the model. It is also important to point out that these classical inference procedures are based on the asymptotical normality of maximum likelihood estimators (see for example, Lawless,1982).This could be a problem when we have small data sizes, a situation very common in medical studies. In this way, we use Bayesian methods to analyze bivariate lifetime data assuming a bivariate geometric distribution. Also it is important to point out that the use of available Bayesian software like the OpenBUGS software, only requires the specification of the distribution for the data and prior distributions for the parameters of the model, giving us a great simplification to obtain posterior summaries of interest, as it was observed in the application considering a medical data set introduced in section 4. These results could be of great interest for applications in medical and engineering studies.

References

[1] Achcar,J.A., Coelho-Barros,E.A. and Mazucheli, J. (2013). Block and Basu bivariate lifetime distribution in the presence of cure fraction, *Journal of Applied Statistics*, **40**(**9**), 1864-1874.

[2] Achcar, J.A. and Leandro, R.A. (1998). Use of Markov Chain Monte Carlo methods in a Bayesian analysis of the Block and Basu bivariate exponential distribution, *Annals of the Institute of Statistical Mathematics* **50**, pp. 403-416.

[3] Arnold, B. (1975). A characterization of the exponential distribution by multivariate geometric compounding. *Sankhya* **37**, p.164-173.

[4] Arnold, B.C. and Strauss, D. (1988). Bivariate distributions with exponential conditionals, *J. Amer. Statist. Assoc.*, **83**,pp. 522-527.

[5] Begg, C.B. and Schrag, D. (2002). Attribution of deaths following cancer treatment. *J. Natl. Cancer Inst.* **94**, pp. 10441045.

[6] Block, H.W. and Basu,A.P. (1974). A continuous bivariate exponential extension, *J. Amer. Statist. Assoc.*, pp. 1031-1037.

[7] Cancho, V.G., Bolfarine, H. (2001). Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, **28**(**6**), 659-671.

[8] Carlin, B.P. and Louis, T.A. (2000). Bayes and Emperical Bayes methods for data analysis, 2nd edition, Chapman and Hall/CRC, London.

[9] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, Amer. Statist., pp. 327-335.

[10] Coleman, M., Babb,P., Damiecki,P., Grosclaude,P., Honjo,S., Jones,J., Knerer,G., Pitard,A., Quinn,M.J., Sloggett,A. and De Stavola, B.: Cancer survival trends in England and Wales, 1971-1995, deprivation and NHS Region.London: Office for National Statistics, (1999).

[11] Davarzani, N. and Parsian, A. (2010). Bayesian inference in dependent right censoring, Communications in Statistics Theory and Methods **39**, pp.1270-1288.

[12] Davarzani, N. and Parsian, A. (2013). Inference under right censoring in a discrete setup, Communications in Statistics Theory and Methods **42**, pp.2362-2375.

[13] Davarzani, N., Parsian, A. and Peeters, R. (2014). Dependent Right Censorship in the Marshall-Olkin Bivariate Weibull Distribution. Communications in Statistics Theory and Methods, DOI:10.1080/03610926.2013.766342.

[14] Gamel, J.W., Mclean, I.W., Rosenberg, S.H. (1999). Proportion cured and mean log-survival time as functions of tumor size. *Statistics in Medicine* **9**, 999-1006.

[15] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, pp. 398-409.

[16] Grimshaw, S.D., McDonald, J., McQueen, G.R., and Thorley, S. (2005). Estimating hazard functions for discrete lifetimes. *Communications in Statistics, Part B-Simulation and Computation* **34**, pp. 451-463.

[17] Gupta, R.D. and Kundu, D. (2007). Generalized exponential distribution: existing methods and recent developments, *Journal of the Statistical Planning and Inference*, **vol. 137**, **no. 11**, 3537-3547.

[18] Hanagal, D.D. (2010). Modeling heterogeneity for bivariate survival data by the compound poisson distribution. *Model Assisted Statistics and Applications*, **5**, 1-9.

[19] Hanagal, D.D. and Ahmadi, K.A. (2008). Estimation of parameters by EM algorithm in bivariate exponential distribution based on censored samples. *Economic Quality Control*, **20**(**2**), 257-66.

[20] Huster, W.J., Brookmeyer, R. and Self, S.G. (1989). Modelling paired survival data with covariates, *Biometrics* **45**, pp. 145-156.

[21] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**: 457-481.

[22] Kannan, N., Kundu, D., Nair, P., Tripathi, R.C. (2010). The generalized exponential cure rate model with covariates, *Journal of Applied Statistics*,**37**(**9-10**), 1625-1636.

[23] Kundu, D. and Gupta, R.D. (2009). Bivariate generalized exponential distribution, *Journal of Multivariate Analysis*, **vol. 100**, **no. 4**, 581-593.

[24] Lambert, P.C. (2007). Modeling of the cure fraction in survival studies, *TheStata Journal*, 7, pp. 351-375.

[25] Lawless, J.F. (1982). Statistical Models and Methods for Lifetime Data,Wiley, NewYork. Mohsin, M., Kazianka, H., Pilz, J. and Gebhardt, A. (2013). A new bivariate exponential distribution for modeling moderately negative dependence, Stat. Meth. appl., DIO 10.1007/s10260-013-0246-3.

[26] Marshall, A.W. and Olkin,I. (1967). A generalized bivariate exponential distribution, *J. Appl. Probab.*, pp. 291-302.

[27] Nair, K.R.M. and Nair, U. (1988). On characterizing the bivariate exponential and geometric distributions. *Annals of the Institute of Statistical Mathematics*, **40**(**2**), pp.267-271.

[28] Santos, C.A. and Achcar, J.A. (2011). A Bayesian analysis for the Block and Basu bivariate exponential distribution in the presence of covariates and censored data, *Journal of Applied Statistics*, **38**, pp. 2213-2223.

[29] Sarkar, S.K. (1987). A continuous bivariate exponential distribution. *Journal of the American Statistical Association*, (**82**), 667-675.

[30] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002).
Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, **B**(**64**), 583-639.

[31] Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R. (2003). WinBUGS User Manual (version 1.4). MRC Biostatistics Unit, Cambridge, UK.

[32] Taylor, J.M.G. (1995). Semiparametric estimation in failure time mixture models. *Biometrics* **51**, 899-907.

[33] Wienke, A., Locatelli, I., Yashin, A.I. (2006). The modelling of a cure fraction in bivariate time-to-event data. *Austrian Journal of Statistics* **35**(**1**), 67-76.

[34] Yamaguchi, K. (1992). Accelerated failure-time regression model with a regression model for the survining fraction: an application to the analysis of permanent employment in japan. *Journal of the American Statistical Association* **87**, 284-292.

[35] Yu, B., Tiwari, R.C., Cronin, K.Z. (2004). Cure fraction estimation from the mixture cure models for grouped survival times. *Statistics in Medicine*, **23**, 1733-1747.

Received November 10, 2015; accepted March 10, 2015.

Nasser Davarzani
Department of Knowledge Engineering, Maastricht University
Maastricht, Netherlands.
Maastricht University, Maastricht, Netherlands.
Tel: +31 638190220
n.davarzani@maastrichtuniversity.nl

Jorge Alberto Achcar
Departamento de Medicina Social, FMRP, Universidade de Sao Paulo
Ribeirao Preto, SP, Brazil.
Universidade de Sao Paulo,Ribeirao Preto, SP, Brazil.
achcar@fmrp.usp.br.jchan@maths.usyd.edu.au

Evgueni Nikolaevich Smirnov
Department of Knowledge Engineering, Maastricht University
Maastricht, Netherlands.
Maastricht University, Maastricht, Netherlands.
maastrichtuniversity.nl jchan@maths.usyd.edu.au

Ralf Peeters
Department of Knowledge Engineering, Maastricht University
Maastricht, Netherlands.
Maastricht University, Maastricht, Netherlands.
ralf.peeters@maastrichtuniversity.nl

Appendix. Deviance Information Criterion (DIC)

The Deviance Information Criterion (DIC) is a criterion specially useful for selection models under the Bayesian approach where samples of the posterior distribution for the parameters of the model are obtained using MCMC methods.

The deviance is defined by

$$D(\theta) = -2logL(\theta) + c,$$

Where θ is a vector of unknown parameters of the model, L(θ) is the likelihood function of the model and c is a constant that does not need to be known when the comparison between models is made.

The DIC criterion defined by (20) is given by

$$DIC = D(\hat{\theta}) + 2n_D$$

Where $D(\hat{\theta})$ is the deviance evaluated at the posterior mean $\hat{\theta} = E(\theta|data)$ and n_D is the effective number of parameters of the model given by $n_D = \overline{D} - D(\hat{\theta})$, where $\overline{D} = E(D(\theta)|data)$ is the posterior deviance measuring the quality of the data fit for the model. Smaller values of DIC indicates better models. Note that these values could be negative.