Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data

Vicente G. Cancho, Dipak K. Dey & Francisco Louzada

To cite this article: Vicente G. Cancho, Dipak K. Dey & Francisco Louzada (2015): Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data, Journal of Applied Statistics, DOI: 10.1080/02664763.2015.1071341

To link to this article: http://dx.doi.org/10.1080/02664763.2015.1071341

Published online: 09 Oct 2015.

Journal of Applied Statistics, 2015
http://dx.doi.org/10.1080/02664763.2015.1071341

Taylor & Francis
Taylor & Francis Group

Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data

Vicente G. Cancho[a], Dipak K. Dey[b] and Francisco Louzada[a]*

[a]*ICMC, University of São Paulo, São Carlos, SP, Brazil;* [b]*Department of Statistics, University of Connecticut, Storrs, CT, USA*

(*Received 27 March 2014; accepted 7 July 2015*)

In this paper we propose a new lifetime model for multivariate survival data in presence of surviving fractions and examine some of its properties. Its genesis is based on situations in which there are m types of unobservable competing causes, where each cause is related to a time of occurrence of an event of interest. Our model is a multivariate extension of the univariate survival cure rate model proposed by Rodrigues *et al.* [37]. The inferential approach exploits the maximum likelihood tools. We perform a simulation study in order to verify the asymptotic properties of the maximum likelihood estimators. The simulation study also focus on size and power of the likelihood ratio test. The methodology is illustrated on a real data set on customer churn data.

Keywords: competing risks; cured fraction; maximum likelihood approach; multivariate survival models; unified survival models

1. Introduction

Cure rate models play an important role in survival analysis. The cover situations in that there are sample units insusceptible or cured with respect to the occurrence of the event of interest. The proportion of such units is termed as the cured fraction. In clinical studies, the event of interest can be the death of a patient, hence the terminology. However, cure rate models have been shown to be appropriate for the modeling many other kinds of events, such as criminal recidivism, divorce, child-bearing, unemployment, and costumer churn.

There is a vast literature on cure rate models for survival data (also called survival models with a surviving fraction or long-term survival models), though the majority of these stems are from either one of the *standard mixture cure model* [3,5,26,33], or the *the promotion time cure* model [2,7,17,37,40]. The books by Ibrahim *et al.* [17], Maller and Zhou [29] are references of these two classes, respectively, and covers a wide range of developments.

*Corresponding author. Email: louzada@icmc.usp.br

Although extensions of cure rate models were developed, limited attention has been paid to the research on multivariate cure rate models. In the frequentist framework, Chatterjee and Shih [6] proposed a marginal approach using bivariate copula models. Price and Manatunga [34] imposed frailty to account for correlation and conducted the maximum likelihood estimation under a parametric model assumption. Both methods were based on the mixture cure model. Louzada and Cobre [20] proposed a multiple time scale survival model with a cure fraction. Those methods were based on the mixture cure model. In the Bayesian approach, Chen *et al.* [8] generalized the work of Chen *et al.* [7] to multivariate failure time data by introducing a positive stable frailty and Louzada *et al.* [22] proposed bivariate long-term distribution based on the Farlie–Gumbel–Morgenstern copula model.

In this paper, a new multivariate cure rate survival model is developed under a scenario of latent competing causes (or risks). Our model is a multivariate extension of the univariate survival cure rate model proposed by Rodrigues *et al.* [37], who unified the cure rate survival models proposed by Berkson and Gage [3], Chen *et al.* [7] and Hanin [16]. In the formulation we consider that there are N types of causes of failures in which each cause produces the correspondent event of interest, where these latent variables are modeled by a multivariate Poisson distribution [18].

The main assumption here is that N is a discrete random variable with support at $\{0, 1, \ldots\}$, which represents the unobservable number of causes (or risks). In Section 2, we shall consider some particular discrete distributions such as the binomial, geometric and Poisson. In the case that N is a unknown number of causes (or risks), we have a so called competing risk survival problem. Indeed, assuming a unknown N, we are assuming that there is a unknown number of latent competing causes (or risk). In many situations this information is not available, or it is impossible that the true cause of failure can be specified by an expert. For instance, in reliability, the components can be totally destroyed in the experiment. Further, the true cause of failure can be masked from our view. In modular systems, the need to keep a system running means that a module that contains many components can be replaced without the identification of the exact failing component.

Practical applications of the cure rate models and extensions are well established in biomedical sciences, criminology and engineering as a method for modeling time-to-event data. Also, these models have been used in economical studies. Yamaguchi [44] considered a cure rate model to the analysis of permanent employment in Japan and Tong *et al.* [42] discussed the application of cure rate model to predict time-to-default on a UK personal loan portfolio. Both studies were based on the standard mixture cure model [3].

Here however, we focusing on customers who may abandon a service from an organization (churning), where the main interest is to predict time-to-churn, cure rate models are particularly useful. This is because, in general, a substantial proportion of customers in the organization do not experience churning. In this context we shall denote cure by non-churn, leading to what we call non-churning rate models. We bring into account a sample of costumers taken from a Brazilian retailer customer portfolio, which comprised time up to churn (in years) in two different credit card products, from where we observe a substantial amount of customers who do not experience churning during the company-customer lifetime. The churning can be regarded as a product of attrition between customer and company [4] and is typically driven by different competitive causes, usually latent. Our example as well as more discussion on churning are properly presented in Section 5. For instance, Figure 1 presents the overall Kaplan–Meier survival curve for the Brazilian retailer customer portfolio. A plateau points out to the presence of non-churn fraction on the data.

The paper is organized as follows. In Section 2, we formulate the multivariate cure rate model. Inference methods based on the likelihood approach are developed in occupy the Section 2.2. Simulation study is presented in the Section 3. An application to a real data set is developed in Section 4. Finally, Section 5 concludes with some general remarks.

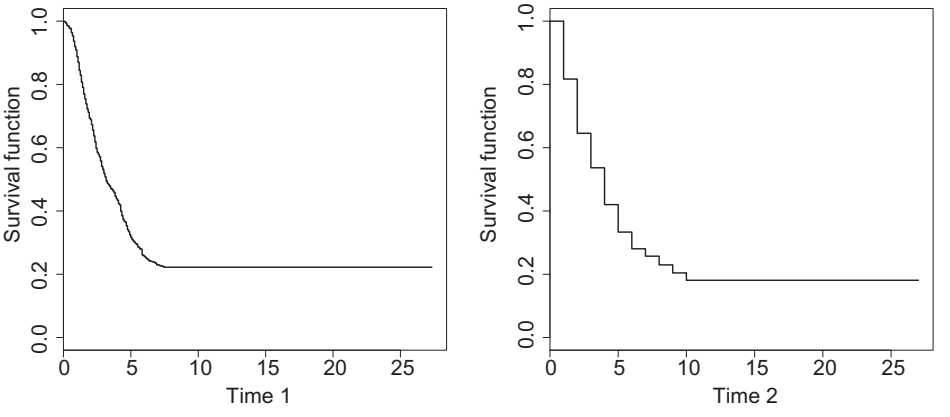

Figure 1. The Brazilian customer churn data. Estimative of Kaplan–Meier of the survival function of Product 1 (right panel) and the Product 2 (left panel).

2. The model

The proposed multivariate cure rate model can be derived as follows. For an individual in the population, let N_k, be the random variable that denote the unobservable number of causes do type k $(k = 1, \ldots, m)$ that can produce the event of interest for this individual. We assume that N_k are discrete random variables taking values in the non-negative integers $\{0, 1, \ldots\}$, with joint probability mass function $P(N_1 = n_1, \ldots, N_m = n_m)$ and probability generating function, $\varphi_{N_1,\ldots,N_m}(w_1, \ldots w_m)$ with $0 \leq w_k \leq 1$. The time for the jth competing cause of type k to produce the event of interest is denoted by $Z_{kj}, k = 1, \ldots, m, j = 1, 2, \ldots$. Given $N_k = n_k$, the $Z_{k1} \ldots, Z_{kn_k}$ are independent and identically distributed random variables with cumulative distribution function $F_k(\cdot) = 1 - S_k(\cdot)$. We also assumed that the latent variables $Z_{1j}, \ldots, Z_{kj}$ are independent. The observable times to event are defined by the random variables $Y_k = \min\{Z_{k1}, \ldots Z_{kN_k}\}$ for $N_k > 0$ and $Y_k = \infty$ if $N_k = 0$ with $P(Y_1 = \infty, \ldots, Y_m = \infty \mid N_1 = 0, \ldots N_m = 0) = 1$. Under this setup we can demonstrate, that the population survival function for $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is given by

$$\begin{aligned}
S_{\text{pop}}(\boldsymbol{y}) &= P[\boldsymbol{N} = \boldsymbol{0}] + \sum_{n_1,\ldots,n_m=1} P[Z_{11} > y_1, \ldots, Z_{1n_1} > y_1, \ldots, Z_{m1} > y_m, \ldots, Z_{mn_m} > y_m] \\
&\quad \times P[N_1 = n_1, \ldots, N_m = n_m] \\
&= P[N_1 = 0, \ldots, N_m = 0] + \sum_{n_1,\ldots,n_m=1} P[N_1 = n_1, \ldots, N_m = n_m] S_1^{n_1}(y_1) \cdots S_m^{n_m}(y_m) \\
&= \sum_{n_1,\ldots,n_m=0} P[N_1 = n_1, \ldots, N_m = n_m] S_1^{n_1}(y_1) \cdots S_m^{n_m}(y_m) \\
&= \varphi_{N_1,\ldots,N_m}(S_1(y_1), \ldots, S_m(y_m)). \qquad (1)
\end{aligned}$$

Note that, the last step of Equation (1) comes from the definition of the probability generating function. The model in Equation (1) is a natural extension of the univariate survival cure rate models.

The survival function $S_{\text{pop}}(\boldsymbol{y})$ in Equation (1) is not a proper survival, that is, $\lim_{y_1,\ldots,y_m \to \infty} S_{\text{pop}}(\boldsymbol{y}) = \varphi_{N_1,\ldots,N_m}(0, \ldots 0) = P[N_1 = 0, \ldots, N_m = 0] > 0$ (the joint cure rate).

From Equation (1) the marginal survival function is obtained as

$$S_{\text{pop}}(y_k) = \varphi_{N_k}(S_k(y_k)), \quad k = 1, \ldots, m. \qquad (2)$$

The marginal survival function (2) is the same as the one proposed by Rodrigues *et al.* [37]. The marginal cure rate proportion is $S_k(\infty) = \varphi_{N_k}(0) = P[N_k = 0] > 0$. In the case that random variables $N_1, \ldots, N_m$ are independent, the survival function $S_{\text{pop}}(\boldsymbol{y})$ in Equation (1) is given by $S_{\text{pop}}(\boldsymbol{y}) = \prod_{k=1}^{m} \varphi_{N_k}(S_k(y_k))$.

2.1 *Some examples*

In this section we apply the results obtained so far for $\boldsymbol{N}$ following the bivariate Bernoulli, bivariate geometric and multivariate Poisson distribution.

An extension of bivariate version of univariate standard mixture cure rate model. In what follows, consider a bivariate Bernoulli random vector $\boldsymbol{N} = (N_1, N_2)$, which takes values from $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ in the cartesian product space $\{0,1\}^2 = \{0,1\} \times \{0,1\}$. Denote $\theta_{ij} = P[N_1 = i, N_2 = j]$, $i,j = 0,1$, such that, $\sum_{i=0}^{1}\sum_{j=0}^{1} \theta_{ij} = 1$. Then the corresponding probability generating function is given by $\varphi_{\boldsymbol{N}}(w_1, w_2) = \theta_{00} + \theta_{10}w_1 + \theta_{01}w_2 + \theta_{11}w_1w_2$. From Equation (1), we obtain the following results related to the standard mixture cure rate model,

$$S_{\text{pop}}(y_1, y_2) = \theta_{00} + \theta_{10}S_1(y_1) + \theta_{01}S_2(y_2) + \theta_{11}S_1(y_1)S_2(y_2).$$

Thus the joint cure fraction is $S_{\text{pop}}(\infty, \infty) = \theta_{00}$. The marginal survival functions are given by $S_{\text{pop}}(y_1) = \theta_{00} + \theta_{01} + (\theta_{10} + \theta_{11})S_1(y_1)$ and $S_{\text{pop}}(y_2) = \theta_{00} + \theta_{10} + (\theta_{01} + \theta_{11})S_2(y_2)$.

Now, if we consider a random vector $\boldsymbol{N} = (N_1, N_2)$, following a bivariate Geometric distribution with probability mass function

$$P[N_1 = n_1, N_2 = n_2] = \binom{n_1 + n_2}{n_1} (1 - \theta_1 - \theta_2)\theta_1^{n_1}\theta_2^{n_2},$$

where $n_j = 0, 1, \ldots, 0 < \theta_j < 1, j = 1, 2$ and $0 < 1 - \theta_1 - \theta_2 < 1$. From Equation (1), we obtain the new bivariate cure rate model as

$$S_{\text{pop}}(y_1, y_2) = \frac{1 - \theta_1 - \theta_2}{1 - \theta_1 S_1(y_1) - \theta_2 S_2(y_2)},$$

with the joint cure fraction $S_{\text{pop}}(\infty, \infty) = 1 - \theta_1 - \theta_2$. The marginal survival functions are given by $S_{\text{pop}}(y_1) = (1 - \theta_1 - \theta_2)/(1 - \theta_2 - \theta_1 S_1(y_j))$ and $S_{\text{pop}}(y_2) = (1 - \theta_1 - \theta_2)/(1 - \theta_1 - \theta_2 S_2(y_j))$, respectively. This marginal model is similar to the one proposed by Gu *et al.* [15].

Finally, if we assume that $\boldsymbol{N} = (N_1, \ldots, N_m)$ follows a multivariate Poisson distribution with probability mass function

$$P[N_1 = n_1, \ldots, N_m = n_m] = \mathrm{e}^{-\{\sum_{i=1}^{m} \theta_i\}} \prod_{i=1}^{m} \frac{\theta_i^{n_i}}{n_i!} \sum_{i=0}^{s} \prod_{j=1}^{m} \binom{n_j}{i!} i! \left(\frac{\theta_0}{\prod_{i=1}^{m} \theta_i}\right)^i, \tag{3}$$

where $n_j = 0, 1, \ldots, \theta_j > 0, j = 0, 1, \ldots, m$ and $s = \min\{n_1, \ldots, n_m\}$. The above multivariate distribution allows for positive dependence between the two random variables. Marginally each random variable follows a Poisson distribution with $E(N_j) = \theta_j + \theta_0$ and, $\text{Cov}(N_i, N_j) = \theta_0$, $i \neq j = 1, \ldots, m$ and hence θ_0 is a measure of dependence between the two random variables. If $\theta_0 = 0$ then the variables are independent and the multivariate Poisson distribution reduces to the product of m independent Poisson distributions. For a comprehensive treatment of the multivariate Poisson distribution the reader can refer to Karlis [18]. The probability generating function,

of $\boldsymbol{N}$ is given by

$$\varphi_{\boldsymbol{N}}(w_1,\ldots,w_m)=\exp\left\{-\sum_{i=1}^{m}\theta_i(1-w_i)-\theta_0\left(1-\prod_{i=1}^{m}w_i\right)\right\}.$$

From Equation (1), we obtain the new multivariate cure rate model as

$$S_{\text{pop}}(\boldsymbol{y})=\exp\left\{-\sum_{i=1}^{m}\theta_i(1-S_i(y_i))-\theta_0\left(1-\prod_{i=1}^{m}S_i(y_i)\right)\right\}. \tag{4}$$

The survival function $S_{\text{pop}}(\boldsymbol{y})$ in Equation (4) is not a proper survival, that is, $\lim_{y_1,\ldots,y_m\to\infty}S_{\text{pop}}(\boldsymbol{y})=\exp\{-\sum_{i=0}^{m}\theta_i\}>0$ (the joint cure fraction). Note that when $\theta_0=0$ in Equation (4), the joint survival function reduces to the product of m independent survival functions. The marginal distribution of each component in Equation (4) has a proportional hazards structure if the covariates enter the model only through $(\theta_0,\ldots,\theta_m)$. This is a desirable feature of the proposed model that leads to attractive theoretical properties. From Equation (4) the marginal survival functions are

$$S_{\text{pop}}(y_k)=\exp\{-(\theta_k+\theta_0)F_k(y_k)\},\quad k=1,\ldots,m. \tag{5}$$

Equation (5) indicates that the marginal survival function has a cure rate structure with probability of cure $p_{0k}=\mathrm{e}^{-\theta_k-\theta_0}$ for Y_k, $k=1,\ldots,m$. It is important to note in Equation (5) that each marginal survival function has the structure of the promotion time cure model [7,43]. In Equation (5) that each marginal survival function has a proportional hazards structure as long as the covariates, only enter through θ_k and θ_0. The marginal hazard function is given by, $(\theta_k+\theta_0)f_k(y_k)$ which satisfies the conditions for the proportional hazards model [12].

Without loss of generality , considering the bivariate distribution of (Y_1,Y_2), then joint survival function in Equation (4) is given by

$$S_{\text{pop}}(y_1,y_2)=\exp\{-\theta_1(1-S_1(y_1))-\theta_2(1-S_2(y_2))-\theta_0(1-S_1(y_1)S_2(y_2))\}. \tag{6}$$

The parameter θ_0 is a measure of association between (Y_1,Y_2). As $\theta_0\to 0$, this implies less association between (Y_1,Y_2) which can be seen from Equation (6). Following Clayton [9] and Oakes [32] , we can compute a local measure of association, denoted by $\vartheta^*(Y_1,Y_2)$, as a function of θ_0. This measure of association is defined as

$$\vartheta^*(Y_1,Y_2)=\frac{S_{\text{pop}}(y_1,y_2)(\partial^2/\partial y_1\partial y_2)S_{\text{pop}}(y_1,y_2)}{(\partial S_{\text{pop}}(y_1,y_2)/\partial y_1)(\partial S_{\text{pop}}(y_1,y_2)/\partial y_2)}. \tag{7}$$

The measure in Equation (7), has the interpretation as the ratio of the hazard rate of the conditional distribution of Y_1 (Y_2), given $Y_2=y_2$ $(Y_1=y_1)$, to that of Y_1 (Y_2) given $Y_2>y_2$ $(Y_1>y_1)$. For more discussion of Equation (7), see [9]. For the bivariate cure rate model in Equation (6), $\vartheta^*(y_1,y_2)$ is well defined and is given by

$$\vartheta^*(y_1,y_2)=1+\theta_0\{[\theta_1+\theta_0S_2(y_2)][\theta_2+\theta_0S_1(y_1)]\}^{-1}. \tag{8}$$

In Figure 2, we see that $\vartheta^*(y_1,y_2)$ in Equation (8) increases in (y_1,y_2). That is, the association between (y_1,y_2) is less when (y_1,y_2) are small and the association increases over time.

As pointed out by Chen *et al.* [8] in multivariate survival models with cure rate can not determine global dependence measures (such as correlation coefficient, which depends on time) due to the fact that the random variables can take infinite with positive probability. The same authors recommend using measures of local dependency as presented here.

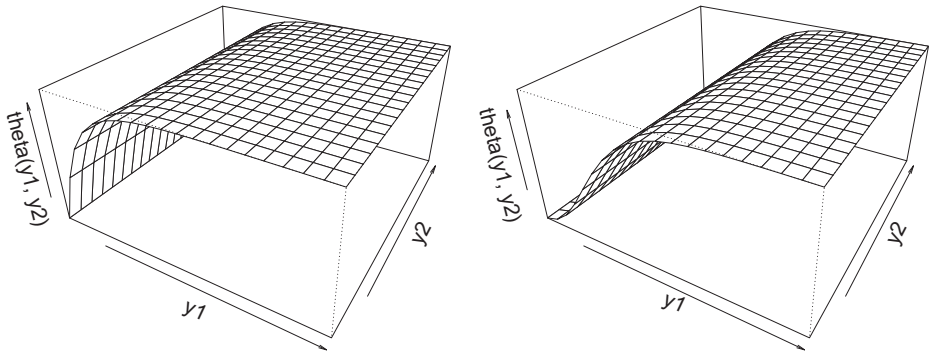

Figure 2. Local measure of association for model with $\theta_1 = 0.2$, $\theta_2 = 0.3$, $\theta_0 = 0.2$ (left panel) and $\theta_0 = 2$ (right panel).

In the next section, we develop the inferential procedures based on likelihood theory for the bivariate survival cure rate model given in Equation (6). Hereafter we assume a Weibull distribution for the unobserved time Z with $F_k(z \mid \boldsymbol{\gamma}_k) = 1 - S_k(z \mid \boldsymbol{\gamma}_k) = 1 - \exp(-z^{\gamma_{1k}} \mathrm{e}^{\gamma_{k2}})$ and $f_k(z \mid \boldsymbol{\gamma}_k) = \gamma_{k1} z^{\gamma_{k2}-1} \exp(\gamma_{k2} - z^{\gamma_{k1}} \mathrm{e}^{\gamma_{k2}})$, for $z > 0$, $\gamma_{k1} > 0$, is the shape parameter and $\gamma_{k2} \in \mathbb{R}$ is the scale parameter and $\boldsymbol{\gamma_k} = (\gamma_{1k}, \gamma_{2k})^\top$, $k = 1, 2$.

2.2 *Inference for bivariate survival cure rate model*

Let us consider the situation when the failure times (Y_1, Y_2) in Section 2 are not completely observed and are subject to right censoring. Let C_{ki} denote the censoring time of k component, $k = 1, 2$. Suppose that $(Y_{1i}; Y_{2i})$ and $(C_{1i}; C_{2i})$ are independent. For each individual i, observed quantities are represented by the random variables $t_{ki} = \min(Y_{ki}, C_{ki})$ and $\delta_{ki} = I(y_{ki} = Y_{ki})$, which denotes a censorship indicator, $k = 1, 2$, $i = 1, \ldots, n$. Assuming that C_{1i} and C_{2i} are independents, and both are non-informative censoring, then we can express the likelihood of $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \theta_0)$ as,

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} S_{\text{pop}}(t_{1i}, t_{2i}) \prod_{k=1}^{2} [f_k(t_{ki} \mid \boldsymbol{\gamma}_k)]^{\delta_{ki}} [\theta_0 + (\theta_2 + \theta_0 S_1(t_{1i} \mid \boldsymbol{\gamma}_1))(\theta_1 + \theta_0 S_2(t_{2i} \mid \boldsymbol{\gamma}_2))]^{\delta_{1i}\delta_{2i}}$$
$$\times (\theta_1 + \theta_0 S_2(t_{2i} \mid \boldsymbol{\gamma}_2))^{\delta_{1i}(1-\delta_{2i})} (\theta_2 + \theta_0 S_1(t_{1i} \mid \boldsymbol{\gamma}_1))^{\delta_{2i}(1-\delta_{1i})}, \tag{9}$$

where $S_{\text{pop}}(t_1, t_2)$ is survival function given in Equation (6) whereas $f_k(t_{ki} \mid \boldsymbol{\gamma}_k)$ and $S_k(t_{ki} \mid \boldsymbol{\gamma}_k)$, $k = 1, 2$ are, respectively, the density and survival functions of the Weibull distribution. The maximum likelihood estimates (MLEs) of $\boldsymbol{\vartheta}$ can be obtained by maximizing the log–likelihood function $\ell(\boldsymbol{\vartheta}) = \log L(\boldsymbol{\vartheta})$, which is equivalent to solve the following nonlinear equation system $\partial \ell(\boldsymbol{\vartheta}) / \partial \vartheta_k = 0$, for $k = 1, \ldots, \dim(\boldsymbol{\vartheta})$. Under suitable regularity conditions, following Cox and Hinkley [11], it can be shown that $\mathbf{R}(\boldsymbol{\vartheta})(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) \xrightarrow{D} N(\mathbf{0}, \mathbf{I})$, as $n \to \infty$, where $\mathbf{I}$ denotes the identity matrix and $\mathbf{R}(\boldsymbol{\vartheta})$ denotes Choleski decomposition of the observed information matrix $\mathcal{I}_{\text{obs}}(\boldsymbol{\vartheta})$, that is, $\mathbf{R}(\boldsymbol{\vartheta})^\top \mathbf{R}(\boldsymbol{\vartheta}) = \mathcal{I}_{\text{obs}}(\vartheta)$. Thus, the approximate distribution of $\hat{\boldsymbol{\vartheta}}$ (MLE) in large samples is a multivariate normal distribution with mean vector $\boldsymbol{\vartheta}$ and covariance matrix $\mathcal{I}^{-1}_{\text{obs}}(\boldsymbol{\vartheta})$, which can be estimated by $\{-\partial \ell \boldsymbol{\vartheta}) / \partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top\}^{-1}$ evaluated at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$. Second derivative computations can be obtained numerically.

Besides estimation, hypothesis testing is another key issue. Let $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ be proper disjoint subsets of $\boldsymbol{\vartheta}$. Suppose that we have interest in testing $H_0 : \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_{10}$ versus $H_1 : \boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_{10}$, with $\boldsymbol{\vartheta}_2$ being a nuisance parameter. Let $\hat{\boldsymbol{\vartheta}}_0$ be the MLE under H_0 and define the log-likelihood ratio

statistic given by, $\Lambda_n = 2\{\ell(\hat{\boldsymbol{\vartheta}}) - \ell(\hat{\boldsymbol{\vartheta}}_0)\}$. Under H_0 and some regularity conditions, Λ_n converges in distribution to a chi-square distribution with $\dim(\boldsymbol{\vartheta}_1)$ degrees freedom [31]. Another hypothesis test of interest, which is studied further here, is to test the independence model ($H_0 : \theta_0 = 0$) against the dependent one ($H_1 : \theta_0 > 0$). The null distribution of the likelihood ratio test under H_0 is however non-standard [39] and it has been found that the distribution can be approximated by a 50–50 mixture of the chi-square distribution with 1 degree of freedom and a degenerated distribution at zero. That is, the statistic Λ_n converge to $0.5 + 0.5P[\chi_1^2 \leq x]$ distribution, where χ_1^2 is chi-square distribution with *one* degree freedom. We investigate the sample properties of the distributions of the tests via a simulation study.

3. Simulation study

To evaluate the performance of the maximum likelihood estimative of the parameter of the bivariate survival cure rate model, we carry out a simulation study. In this study we consider the bivariate cure rate model with a Weibull distribution for the event time (Z_{kj}, $k = 1, 2$, $j = 1, 2, \ldots$.), with parameter, $\gamma_{k1} = 1.4$ and $\gamma_{k2} = 2.0$. For each individual i, $i = 1, \ldots, n$, the number of causes of the event of interest (N_1, N_2) is generated from a bivariate Poisson distribution with parameter $\theta_0 = 0.5$, $\theta_1 = 0.2$ and $\theta_2 = 0.3$, so that the joint cured fraction and marginals are $p_{00} = 0.37$, $p_{01} = 0.50$ and $p_{02} = 0.45$, respectively. The censoring times C_{ki} are sampled from the uniform distribution on the interval $(0, \tau_k)$, where τ_k is a set in order to control the proportion of censored observations. In this study the proportion of censored observations is on an average approximately equal to 55% and 45%, respectively.

We choose five sample sizes, $n = 50, 100, 200, 300$ and 600. For each configuration, we conduct 1,000 simulations and then calculate the average of the MLEs of the parameters of the model and the cured fraction, the standard deviation (SD) of the MLEs, the square RMSE of the MLEs and the coverage probability (CP) of the 95% confidence intervals.

The simulation results are shown in Table 1. We observe that the averages of the MLEs of the parameters of the model and the cured fraction are close to the true values. We also observe, the SDs and RMSEs decrease as the sample size increases. Also, from Table 1, we observe that PCs are closer to the nominal value as the sample size increases.

Table 1. Averages of maximum likelihood estimates (AMLEs), SD and square root of mean square error (RMSE) of the parameters of bivariate cure rate model and cured fractions p_{01}, p_{02} and p_{00}.

		Parameter							Cure fraction		
n		θ_1	θ_2	θ_0	γ_{11}	γ_{12}	γ_{21}	γ_{22}	p_{01}	p_{02}	p_{00}
50	AML	0.222	0.328	0.495	2.056	1.436	2.044	1.435	0.494	0.445	0.358
	SD	0.107	0.129	0.136	0.299	0.304	0.302	0.303	0.072	0.072	0.068
	REQM	0.109	0.132	0.136	0.304	0.308	0.305	0.306	0.072	0.089	0.068
	CP	0.913	0.921	0.919	0.922	0.930	0.913	0.909	0.912	0.909	0.923
100	AMLE	0.211	0.302	0.499	2.047	1.416	2.047	1.424	0.495	0.452	0.367
	SD	0.077	0.091	0.096	0.227	0.219	0.222	0.220	0.053	0.054	0.052
	REQM	0.077	0.091	0.096	0.232	0.221	0.226	0.223	0.054	0.070	0.052
	CP	0.938	0.942	0.939	0.942	0.941	0.940	0.937	0.935	0.935	0.948
200	AMLE	0.203	0.303	0.500	2.018	1.409	2.030	1.410	0.497	0.450	0.368
	SD	0.057	0.066	0.068	0.155	0.161	0.154	0.153	0.037	0.038	0.0362
	REQM	0.057	0.066	0.068	0.156	0.163	0.157	0.155	0.037	0.060	0.036
	CP	0.942	0.943	0.949	0.942	0.953	0.950	0.948	0.949	0.945	0.950
300	AMLE	0.201	0.298	0.502	2.008	1.391	2.021	1.403	0.496	0.450	0.369
	SD	0.047	0.051	0.054	0.133	0.130	0.131	0.129	0.031	0.031	0.029
	REQM	0.047	0.051	0.054	0.133	0.130	0.132	0.130	0.031	0.055	0.029
	CP	0.951	0.950	0.954	0.949	0.952	0.951	0.953	0.952	0.950	0.952

Table 2. Empirical rejection rates of the null hypothesis $H_0 : \theta_0 = 0$ at a nominal significance level of 5%.

	n			
θ_0	50	100	200	300
0.0	0.028	0.043	0.047	0.052
0.1	0.175	0.326	0.576	0.745
0.2	0.443	0.706	0.936	0.985
0.5	0.810	0.987	0.992	0.998
1.0	0.935	0.995	0.999	0.999

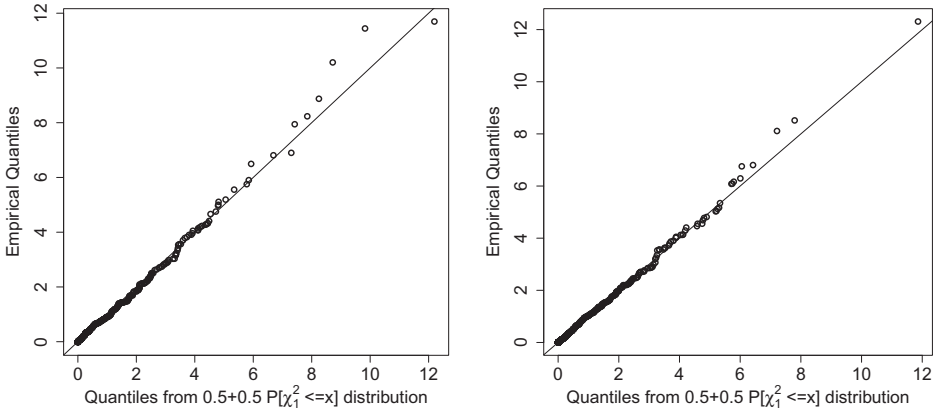

Figure 3. Q–Q plot of empirical Λ_n against the mixture of the Chi-square distribution with 1 degree of freedom and the degenerated distribution at 0.

Additionally, we conduct a simulation study to search the null distribution of the likelihood ratio test, Λ_n, to test the hypotheses $H_0 : \theta_0 = 0$ versus $H_1 : \theta_0 > 0$. Table 2 summarizes the results of the simulation study considering different sample sizes. The rejection rates are close to 5% for moderate sample sized. Besides, the power of the test increases as sample size increases. To further examine the null distribution of the tests, we plot the simulated null distributions of Λ_n for sample size 100 (left panel) and 600 (right panel). They are displayed in Figure 3. The plots show that the mixture of chi-square distribution with 1 degree of freedom with a degenerated distribution at zero provides reasonable approximation to the null distribution of Λ_n.

4. The Brazilian customer churn data

In order to illustrate our proposed modeling discussed so far, in this section we consider a sample of customers taken from a Brazilian retailer customer portfolio, as already stated in Section 1, which comprised of time up to churn (in years) in two different credit card products.

Usually known as customer churn, but also, customer turnover or customer defection, the loss of customers is a major concern of companies or service providers today. The central focus is on voluntary churn, which occurs due to a personal decision of the client, who decides to migrate to another company or service provider.

For the Brazilian customer churn data we observe two times up to churn, Y_1 (in years) and Y_1 (in years), in two credit card products, hereafter Product 1 and Product 2. The study was carried out with 945 credit card holders of a large Brazilian retailer in order to measure the customer churn. Censoring is approximately 23% and 19% for each one of the products, respectively,

and they can be regarded as the loyal customer fractions, which remain in the product even after several years. Here, though the cause of churn is latent, we can conjecture some possible competitive causes, which may prevent a customer from being loyal to a particular product, such as: more attractive management fees or exemption from fees, greater facilities in using the credit card, more solicitous managers, faster problem solving, easier to pay bills, offering major advantages such as high score to earn award miles (on airlines) and best after-sales help is provided. The main objectives here are to understand the relationship between the times up to

Table 3. MLEs for the bivariate cure rate model for the Brazilian customer churn data.

Parameter	Estimate	Standard error	Conf. interval (95%)	
			L	*U*
θ_0	1.390	0.066	1.261	1.519
θ_1	0.103	0.024	0.056	0.150
θ_2	0.293	0.041	0.213	0.374
γ_{11}	1.921	0.059	1.806	2.036
γ_{12}	− 2.736	0.085	− 2.903	− 2.569
γ_{21}	1.839	0.0495	1.743	1.936
γ_{22}	− 3.060	0.089	− 3.235	− 2.886

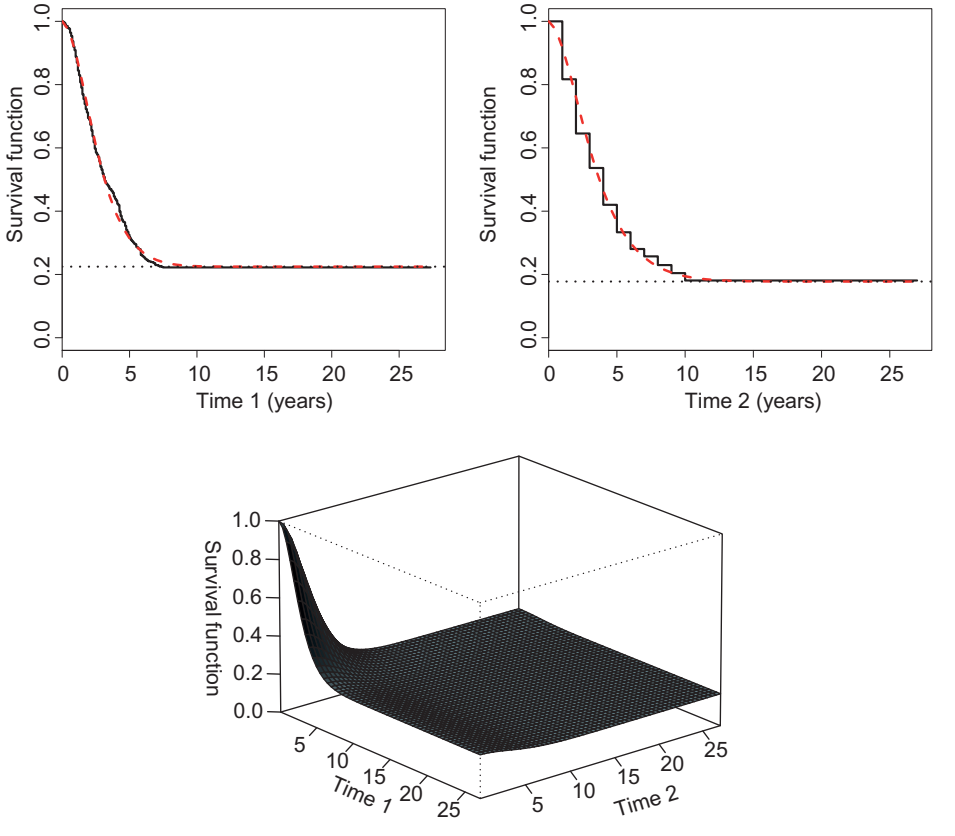

Figure 4. The Brazilian customer churn data. Kaplan–Meier survival curve together with MLEs of the survival of Product 1 (top right panel) and the Product 2 (top left panel). The bivariate survival surface (bottom panel).

churn in the two credit card products and estimate the loyal customer fractions. The idea is to provide the retailer with tools for enhancing customer retention.

Then, the bivariate cure rate (BCR) model proposed in Equation (6) is fitted to the data. Table 3 presents the maximumm likelihood estimates (MLEs) and the corresponding 95% confidence intervals of the BCR model parameters. We observe that the 95% confidence interval of θ_0 not including zero which indicates a significative dependence between Y_1 and Y_2. Alternatively, we can compute the likelihood ratio statistics (LRS) to verify dependence between the random variables Y_1 and Y_2, that is, $H_0 : \theta_0 = 0$ versus $H_1 : \theta_0 > 0$, where under the null hypothesis H_0, the LRS, Λ_n, is assumed to be asymptotically distributed as a symmetric mixture of a chi-squared distribution with one degree of freedom and a point-mass at zero. Thus, Λ_n is equal to 678.43, with a $p-$value $= 0.00001$, which is a strong evidence in favor of the H_1.

Figure 4 exhibits the Kaplan–Meier estimates of the survival function of customers churn together with the MLEs of the marginal survival function (top panel) based on the multivariate cure rate model and the joint survival function (bottom panel). Finally, the MLEs of the non-churning fractions for the plots in Figure 4 are, respectively: $\hat{p}_{01} = 0.225$ [0.195, 0.254], $\hat{p}_{02} = 0.186$ [0.160, 0.212] and $p_{00} = 0.168$ [0.143, 0.92]. The quantities in brackets are the 95% asymptotic confidence intervals, after applying the delta method. We observe that $\hat{p}_{01}$ and $\hat{p}_{02}$ are different from zero, indicating the presence of customer non-churning fractions for both credit products after a period of almost 7 years for Product 1 and 10 years for Product 2. While, the $\hat{p}_{00} = 0.168$ indicates the presence of joint customer non-churning fraction. Thus, actions for customer retention are desirable to reduce churn in periods shorter than 7 and 10 years, respectively, since churning of long-term customers is worth to the retailer than newly recruited ones. Moreover, the retailer's shares are much facilitated by the fact that the mean number of competing latent causes for Product 1 and Product 2 are estimated to be equal to 1.493 and 1.683, respectively, indicating a small amount of competitive latent causes to be investigated.

5. Final comments

In this paper we proposed a new multivariate survival model with cure rate, having examined some of its properties. Some of its particular cases are focused, that is, the bivariate Bernoulli, bivariate geometric and multivariate Poisson distributions. The model is a multivariate extension of the unified survival cure rate model proposed by Rodrigues *et al.* [37]. The model is useful for jointly modeling lifetime data with a cure rate fraction. The model can be easily extended to incorporate covariate information into the model parameters. It is of practical use in settings were the competing causes (or risks), related to the occurrence of the phenomenon, are latent, in the sense there is no information about which one was responsible for occurrence of the phenomenon. Its applicability is discussed in a Brazilian customer churn data set, from where we discovered the BCR model delivers the best fit. Moreover, from our modeling we discover there is presence of customer non-churning fractions of about 20% for both credit products and presence of a small joint customer non-churning fraction, directing to actions for customer retention in order to reduce churn. The modeling considered in this article can be fitted using standard available software [35], which makes the approach quite powerful and accessible to practitioners in the field.

There are evident gain by the multivariate model over the independent model. In general, the differences between populations (groups) do not depends on only one variable but a set of them. The use of only one variable can produce erroneous results. There are some cases, for example in the univariate study indicates a group (or population or treatments) to be the best or most appropriate. However, when considering other variables, jointly, other treatments may shown to be the most appropriate. There are still situations in which, when the variables are analyzed separately, significant differences between populations (or treatments or groups) are not detected

for the study variables. However, when the analysis is made globally, by a multivariate approach, differences are highlighted and are detected appropriately by statistical tests.

There are a large number of further possible developments of the current work. Future developments of our work may involve other parametric or semi-parametric extensions of the multivariate lifetime distribution under the proposed setup. Macera *et al.* [28] proposed an simple exponential-Poisson model for multivariate data. Besides, we envisage the relaxation of the assumption that $Y_k = \min\{Z_{k1}, \ldots Z_{kN_k}\}$. Indeed, we may consider the modeling of its counterpart, $Y_k = \max\{Z_{k1}, \ldots Z_{kN_k}\}$, corresponding to a complementary risk scenarios as discussed by Louzada-Neto [25] and further developed by Barriga *et al.* [2], Roman *et al.* [38], Flores *et al.* [14], Louzada *et al.* [21] and Tojeiro *et al.* [41] among others. Moreover, following Cooner *et al.* [10] and Louzada *et al.* [19], we envisage a generalization of our framework by assuming Y_k as random, such as we may scan all possible Y_k values from the first to last order statistics.

Accelerated life tests is common in practice. Our approach should be investigate in this context. A possible approach is to consider the accelerated life test schemes adopted by Achcar and Louzada-Neto [1], Rodrigues *et al.* [36], or in a regression fashion as Louzada-Neto [24], Mazucheli *et al.* [30] and Louzada-Neto *et al.* [27]. Finally, influence diagnostics is important in the context of regression modeling. Influence diagnostics should be developed for our multivariate survival model with cure rate. A possible approach is to consider the influential diagnostic scheme developed by Fachini *et al.* [13] and the one developed by Louzada *et al.* [23] for a bivariate promotion lifetime model.

Acknowledgements

The authors thank the reviewers for their comments and suggestions, which led to a substantial improvement of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research was sponsored by the Brazilian organizations CNPq and FAPESP through their research grant programs.

References

[1] J.A. Achcar and F. Louzada-Neto, *A Bayesian approach for accelerated life tests considering the Weibull distribution*, Comput. Statist. Q. 7 (1992), pp. 355–368.
[2] G.D.C. Barriga, F. Louzada, and V.G. Cancho, *The complementary exponential power lifetime model*, Comput. Statist. Data Anal. 55 (2011), pp. 1250–1259.
[3] J. Berkson and R.P. Gage, *Survival curve for cancer patients following treatment*, J. Amer. Statist. Assoc. 47 (1952), pp. 501–515.
[4] W. Buckinx and D. Van den Poel, *Customer base analysis:partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting*, European J. Oper. Res. 164(1) (2005), pp. 252–268.
[5] V.G. Cancho and H. Bolfarine, *Modeling the presence of immunes by using the exponentiated-Weibull model*, J. Appl. Stat. 28(6) (2001), pp. 659–671.
[6] N. Chatterjee and J. Shih, *A bivariate cure-mixture approach for modeling familial association in diseases*, Biometrics 57(3) (2001), pp. 779–786.
[7] M.-H. Chen, J.G. Ibrahim, and D. Sinha, *A new Bayesian model for survival data with a surviving fraction*, J. Amer. Statist. Assoc. 94 (1999), pp. 909–919.
[8] M.-H. Chen, J.G. Ibrahim, and D. Sinha, *Bayesian inference for multivariate survival data with a cure fraction*, J. Multivariate Anal. 80(1) (2002), pp. 101–126.

[9] D.G. Clayton, *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence*, Biometrika 65(1) (1978), pp. 141–151.
[10] F. Cooner, S. Banerjee, B.P. Carlin, and D. Sinha, *Flexible cure rate modeling under latent activation schemes*, J. Amer. Statist. Assoc. 102 (2007), pp. 560–572.
[11] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*, CRC Press, London, 1979.
[12] D. Cox and D. Oakes, *Analysis of Survival Data*, Chapman & Hall, London, 1984.
[13] J.B. Fachini, E.M.M. Ortega, and F. Louzada-Neto, *Comparing several accelerated life models*, Stat. Methods Appl. 17 (2008), pp. 413–433.
[14] J. Flores, P. Borges, V.G. Cancho, and F. Louzada, *The complementary exponential power series distribution*, Braz. J. Probab. Stat. 27 (2013), pp. 565–584.
[15] Y. Gu, D. Sinha, and S. Banerjee, *Analysis of cure rate survival data under proportional odds model*, Lifetime Data Anal. 17(1) (2011), pp. 123–134.
[16] L.G. Hanin, *Iterated birth and death process as a model of radiation cell survival*, Math. Biosci. 169(1) (2002), pp. 89–107.
[17] J.G. Ibrahim, M.-H. Chen, and D. Sinha, *Bayesian Survival Analysis*, Springer, New York, 2001.
[18] D. Karlis, *An EM algorithm for multivariate Poisson distribution and related models*, J. Appl. Stat. 30(1) (2003), pp. 63–77.
[19] F. Louzada, E.M.P. Bereta, and M.A.P. Franco, *On the distribution of the minimum or maximum of a random number of i.i.d. lifetime random variables*, Appl. Math. 3 (2012), pp. 350–353.
[20] F. Louzada and J. Cobre, *A multiple time scale survival model with a cure fraction*, Test 21 (2012), pp. 355–368.
[21] F. Louzada, V. Marchi, and J. Carpenter, *The complementary exponentiated exponential geometric lifetime distribution*, J. Probab. Statist. (2013). Available at http://dx.doi.org/10.1155/2013/502159
[22] F. Louzada, A.K. Suzuki, and V.G. Cancho, *The FGM long-term bivariate survival copula model: Modeling, Bayesian estimation, and case influence diagnostics*, Comm. Statist. Theory Methods 42(4) (2013), pp. 673–691.
[23] F. Louzada, A.K. Suzuki, and V.G. Cancho, *On estimation and influence diagnostics for a bivariate promotion lifetime model based on the fgm copula: A fully Bayesian computation*, TEMA Tend. Mat. Apl. Comput. 14 (2013), pp. 441–461.
[24] F. Louzada-Neto, *Extended hazard regression model for reliability and survival analysis*, Lifetime Data Anal. 3 (1997), pp. 367–381.
[25] F. Louzada-Neto, *Poly-hazard regression models for lifetime data*, Biometrics 55 (1999), pp. 1121–1125.
[26] F. Louzada-Neto, J. Mazucheli, and J.A. Achcar, *Mixture hazard models for lifetime data*, Biom. J. 44 (2002), pp. 355–368.
[27] F. Louzada-Neto, J. Mazucheli, and J.A. Achcar, *Mixture hazard models for lifetime data*, Biom. J. 44 (2002), pp. 3–14.
[28] M.A.C. Macera, F. Louzada, V.G. Cancho, and C.J.F. Fontes, *The exponential-Poisson model for recurrent event data: An application to a set of data on malaria in Brazil*, Biom. J. 57 (2015), pp. 201–214.
[29] R.A. Maller and X. Zhou, *Survival Analysis with Long-Term Survivors*, Wiley, New York, 1996.
[30] J. Mazucheli, F. Louzada-Neto, and J.A. Achcar, *Bayesian inference for polyhazard models in the presence of covariates*, Comput. Statist. Data Anal. 28 (2001), pp. 1–14.
[31] H.S. Migon, D. Gamerman, and F. Louzada, *Statistical Inference: An Integrated Approach*, CRC Press, London, 2014.
[32] D. Oakes, *A model for association in bivariate survival data*, J. R. Stat. Soc. Ser. B Methodol. (1982), pp. 414–422.
[33] G.S.C. Perodona and F. Louzada, *A general hazard model for lifetime data in the presence of cure rate*, J. Appl. Stat. 38 (2011), pp. 1395–1405.
[34] D.L. Price and A.K. Manatunga, *Modelling survival data with a cured fraction using frailty models*, Stat. Med. 20(9–10) (2001), pp. 1515–1527.
[35] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.
[36] J. Rodrigues, H. Bolfarine, and F. Louzada-Neto, *Comparing several accelerated life models*, Comm. Statist. Theory Methods 22 (1994), pp. 2297–2308.
[37] J. Rodrigues, V.G. Cancho, M. de Castro, and F. Louzada-Neto, *On the unification of long-term survival models*, Statist. Probab. Lett. 79 (2009), pp. 753–759.
[38] M. Roman, F. Louzada, V.G. Cancho, and J.G. Leite, *A new long-term survival distribution for cancer data*, J. Data Sci. 10 (2012), pp. 241–258.
[39] S.G. Self and K.-Y. Liang, *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions*, J. Amer. Statist. Assoc. 82(398) (1987), pp. 605–610.
[40] C. Tojeiro, F. Louzada, M. Roman, and P. Borges, *The complementary Weibull geometric distribution*, J. Stat. Comput. Simul. 84 (2012), pp. 1345–1362.

[41] C. Tojeiro, F. Louzada, M. Roman, and P. Borges, *The complementary Weibull geometric distribution*, J. Stat. Comput. Simul.on 84 (2014), pp. 1345–1362.
[42] E.N.C. Tong, C. Mues, and L.C. Thomas, *Mixture cure models in credit scoring: If and when borrowers default*, European J. Oper. Res. 218(1) (2012), pp. 132–139.
[43] A.Y. Yakovlev and A.D. Tsodikov, *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore, 1996.
[44] K. Yamaguchi, *Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment' in Japan*, J. Amer. Statist. Assoc. 87(418) (1992), pp. 284–292.