Polyhazard Models with Dependent Causes

Tsai, Rodrigo
Superior Court of Justice and University of Campinas, Department of Statistics
Rua Sérgio Buarque de Holanda, 651
13083-589 Campinas, Brasil
E-mail: rodrigo.tsai@gmail.com

Hotta, Luiz K.
University of Campinas, Department of Statistics
E-mail: hotta@ime.unicamp.br

Introduction

Polyhazard models are a flexible family for fitting lifetime data. Their flexibility comes from the acknowledgment that there are latent causes of failures. There are many examples of application of these models in the literature. See, for instance Mazucheli *et al.* (2001). In all the applications the latent causes of failure are independent. In this paper we extend the independent polyhazard models considering dependence modeled by copula functions. The model is general enough to allow for various form of dependence and also for any marginal distributions of the latent times. The proposed models are able to generate much more flexible risk functions than the independent polyhazard models, including features such as bathtub shape, multimodality and local effects.

There is another approach in the literature in order to construct flexible hazard functions when the distribution is suggested directly. See, for instance Nadarajah *et al.* (2011). The method proposed in this paper is more general. For instance, each of these distributions could be used as a marginal distribution of the latent causes.

The polyhazard model with dependence

Consider the failure time of n independent units of observation with k competing latent causes of failure acting on each unit and denote by X_{ij} the time to failure of the i-th $(i = 1, .., n)$, observed unit due to cause j, $j = 1, .., k$. The distributions of X_{ij}, which depend only on j, denoted by $X_{ij} \sim f_j(\cdot)$ are considered as known except for unknown parameters. Let $\lambda_j(\cdot)$ and $S_j(\cdot)$, respectively the risk and survival functions, related to time of failure due to cause j. For each unit, only the smallest time, denoted by X_i, is observed, i.e., $X_i = min\{X_{ij}, j = 1, .., k\}$. Thus, considering the independence among risks, namely, between the failure times X_{ij}, for any $i = 1, .., n$ and $j = 1, .., k$, the overall survival function of X_i, denoted by $S(t)$, is given by the product of marginal survival functions, i.e.

$$(1) \quad S(t) = P[X_i > t] = P[X_{i1} >, ..., X_{ik} > t] = \prod_{j=1}^{k} S_j(t),$$

and from the density function of X_i, $f(t) = -\partial S(t)/\partial t$, it follows that the hazard function of X_i, $\lambda(t)$, is given by the sum of the marginal hazards, because

$$(2) \quad \lambda(t) = \frac{-\partial \prod_{j=1}^{k} S_j(t)/\partial t}{\prod_{j=1}^{k} S_j(t)} = \sum_{j=1}^{k} \lambda_j(t).$$

From now on we use the notation for $k = 2$ for simplification, but it can be easily generalized. Denoting by $H(.,.)$ and $\bar{H}(.,.)$ the joint distribution and survival functions of the latent variables, respectively, we can write for the survival function of X_i as $S(t) = \bar{H}(t, t)$. In order to model the joint survival function $\bar{H}$ considering dependence between the latent variables we use copula functions. An m-dimensional copula function may be defined as a cumulative distribution function whose marginal distributions are uniform over $[0,1]$ and whose support is the $[0,1]^m$ hypercube. Copula functions have been extensively studied in literature for multivariate modeling, especially when the use of multivariate normal distribution is questionable. An important feature of the copula approach is the possibility of modeling the dependence and the marginal behavior of the related variates separately, which makes copula a very convenient alternative of multivariate modeling.

According to the Sklar's theorem, given an distribution function $H(.,.)$ there is always a copula function C^* such that $H(t_1,t_2) = C^*(F_1(t_1), F_2(t_2))$; C^* is unique if the marginal distributions F_1 and F_2 are continuous. C^* is then called a copula function, because it couples the marginal distributions F_1 e F_2 to their joint distribution H. It is possible to represent the joint survival function directly by $\bar{H}(t_1,t_2) = P[X_1 > t_1, X_2 > t_2] = \tilde{C}(S_1(t_1), S_2(t_2))$, where $\tilde{C}(u,v) = u+v-1+C^*(1-u,1-v)$ is also a copula. On the other hand, for any copula C, $C(S_1(t_1), S_2(t_2))$ is a survival distribution function. Therefore, we can also model the survival function S directly by a copula function C as was done, for instance, by Kashiev *et al.* (2007). This is the approach adopted here because in general it is easier to work analytically with this representation. Then, for the survival function of the polyhazard model with dependence we can write

(3) $$S(t) = \bar{H}(t,t) = C(S_1(t), S_2(t)),$$

where C is a copula function and S_1 and S_2 are, in this paper and in almost all practical applications, continuous marginal survival functions. The copula C in (3) is called the survival copula, but in this work we just call them the copula function. Notice that the right (left) tail dependence for the latent survival times is equal to the left (right) tail dependence of copula C of (3). By the survival function (3) it follows that the probability density and hazard rate functions for the polyhazard model with dependence are obtained by the usual way, that is

(4) $$f(t) = -dS(t)/dt \quad \text{and} \quad h(t) = f(t)/S(t).$$

The proposed model is a generalization of the independent polyhazard model in the sense that allow for the dependence and at the same time model the marginal behavior of the latent risks. For each combination of copula and marginal survival functions we have a different model allowing us to construct a rich family of competing risk latent models. For instance, in the following we will work with exponential, log-logistic, log-normal, Gamma and Weibull distribution for the latent failure causes and Clayton, Gumbel and copula functions. However, we could work with any distribution and any copula function. The symmetrized Joe Clayton (SJC) copula is not used in the applications, but it is used as example in some parts of the paper. These copula are selected because they have been widely used in the literature and have different type of dependence. Frank copula, with parameter $\theta \in (-\infty, +\infty)$, is a symmetric Archimedean copula with Kendall's $\tau \in (-1,1)$ and Spearman's $\rho \in (-1,1)$, and with lower and upper tail dependence λ_L and λ_U equal to zero. It can generate distributions with strong dependence in the center of the distribution but the dependence in the tails are always small. This means that in the tail the hazard function of the competing risk model will be approximately equal to the sum of marginal hazard functions. For the Clayton copula, the parameter $\theta \in (0, +\infty)$, $\tau = \theta/(\theta+2) \in [0,1)$, $\rho \in [0,1)$, $\lambda_U = 2^{-1/\theta} \in (0,1)$, and $\lambda_L = 0$. For the Gumbel copula, the parameter $\theta \in [1, +\infty)$, $\tau = (\theta-1)/\theta \in [0,1)$, $\rho \in [0,1)$, $\lambda_U = 0$, and $\lambda_L = 2 - 2^{1/\theta} \in [0,1)$. For the SJC copula λ_L and $\lambda_U \in [0,1)$. This features must be taken into consideration when selecting the copula function. λ_L and λ_U are the dependence between the latent variables.

The Figure 1 illustrates some possible shapes for the X_i distribution for the Frank-Weibull-Weibull (Frank copula with marginal Weibull distributions) specification, considering $X_{i1} \sim W(4; 0.9)$ and $X_{i2} \sim W(5; 3)$ and the dependence parameter varying in a range where the Kendall's τ ranges

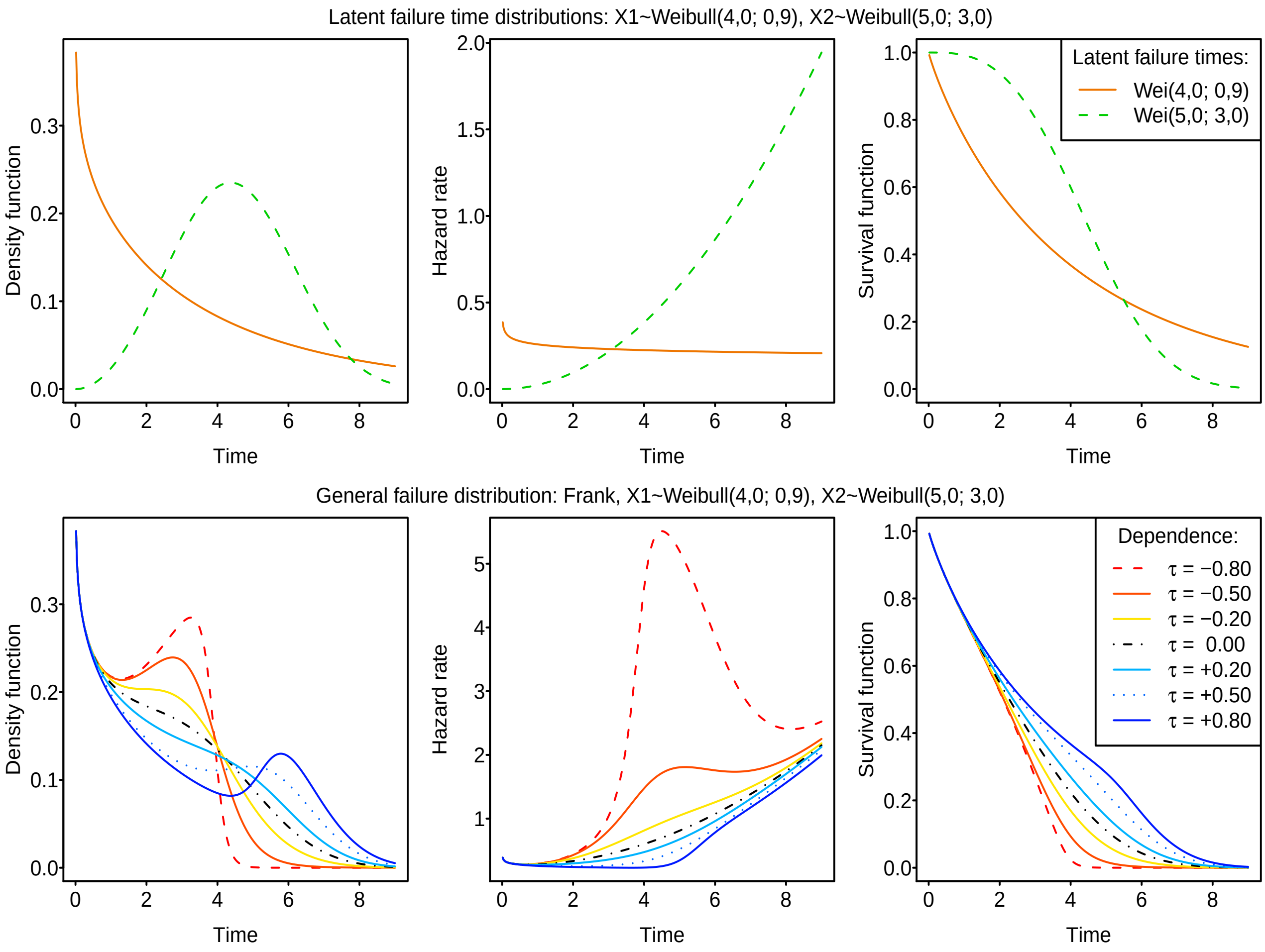

Figure 1: Examples of density, hazard and survival functions for the single risk Weibull model and polyhazard model with dependence with Frank copula and Weibull marginals.

from -0.80 to 0.80. The figure shows that we can have different shapes for the hazard rates, which depends on the shapes of the marginal distributions and also of the dependence type. The Figure 2 shows various hazard rate functions for other specifications of the model in which it is possible to notice local effects, bathtub and multimodal shapes. The two points in the Figure are 99% and 99.9% quartis for each especification and the dependence parameter between the latent variables is the Kendall's τ, except for SJC copula were they are the lower and upper tail dependence.

Model Identification and Estimation

Some models are clearly non identified. Take for instance the model Indep-Exp-Exp (independent copula with both latent variable with exponential distribution). The overall hazard function is constant, say $\lambda > 0$, and the latent hazard function can be any non negative constant, say λ_1 and λ_2, such that $\lambda = \lambda_1 + \lambda_2$. An analysis can be less trivial non identified model, Here we explored the non identification through simulation and numerical analysis. The analysis showed that, except for the specification Indep-Exp-Exp and Gumbell-Exp-Exp, in every especification there is strong evidence of identification. A different point, which is estimability is discussed a little more in the next section. A model being identified does not secure that the parameters can be estimated easily. For instance, when the overall hazard function is dominated by say the first latent cause, it is very difficult to estimate the second latent cause, except where there is a large sample.

In the traditional competing risk literature, when the cause of failure is known, there is another

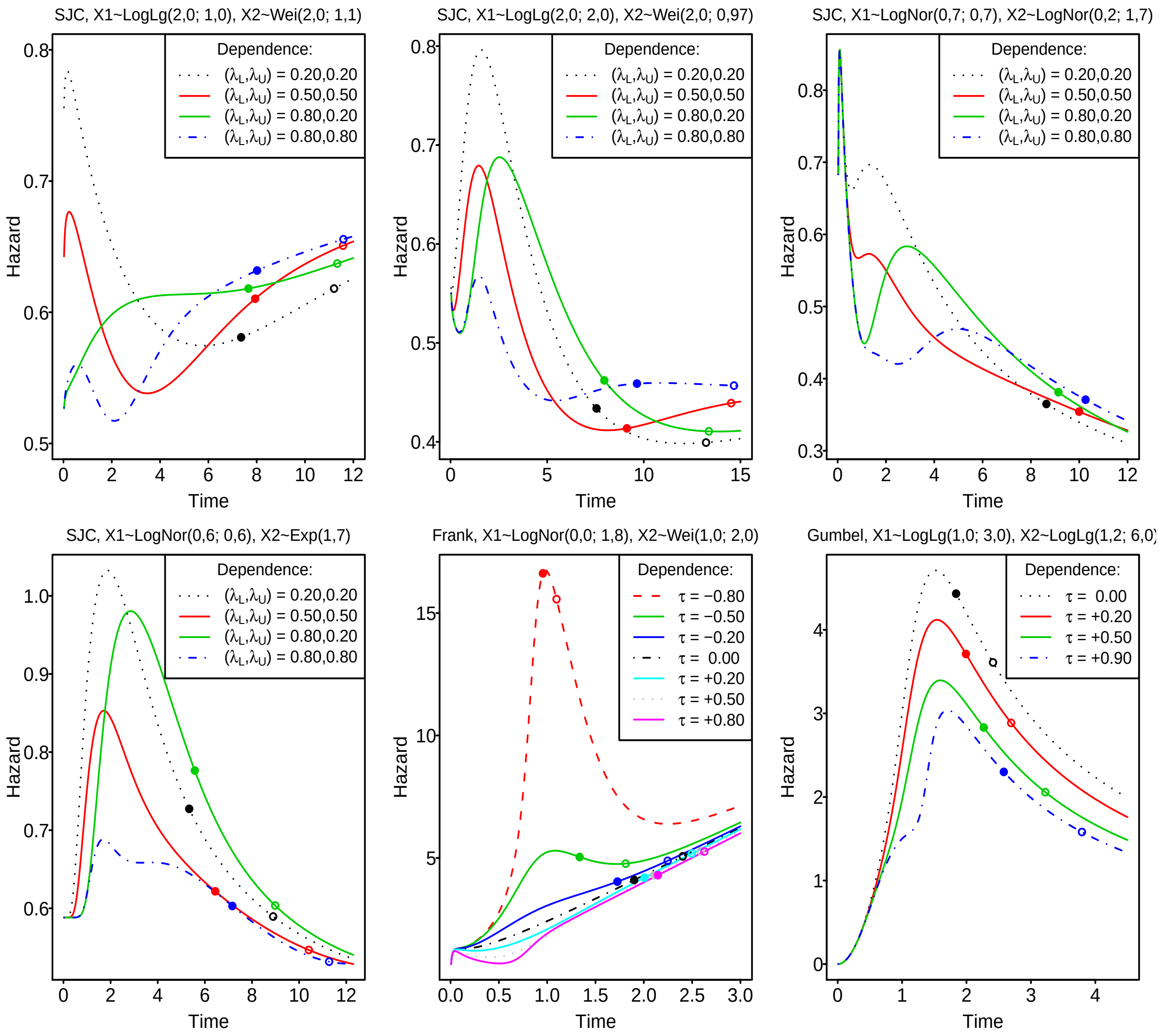

Figure 2: Examples of hazard rate functions for the polyhazard model with dependence.

type of discussion of identification. See, for instance, Tsiatis (1975). In this classical problem a competing risk model is identifiable if the joint survival function can be calculated or identified by the simple knowledge of the marginal survival distributions. Tsiatis (1975) found that, for a model with dependent risks, it is possible to find a set of independent risks that produces the same joint survival distribution. It follows that, unless restrictions are imposed to the behavior of the competing risk, this type of identification is not possible.

The polyhazard models can be seen as a competing risk model with missing values for the cause. This means that we have less information, and therefore the identification of the equivalent competing risk model is a necessary but not sufficient condition for the identification of the polyhazard model. However, even when we have this type of non identification in polyhazard models we can still use these models in order to model lifetime data and take advantage of the good characteristic of these models.

The model parameters are estimated by the method of maximum likelihood. Considering a random sample X_i, $i = 1, \cdots, n$, with random right censoring in which δ_i is the failure indicator variable and t_i the minimum between the failure and censoring, it follows from (3) and (4) that the likelihood is given by $L(\Upsilon) = \prod_{i=1}^{n} f(t_i; \Upsilon)^{\delta_i} S(t_i; \Upsilon)^{1-\delta_i}$, where Υ denotes the parameters for the copula function and the marginal distributions. The algorithm written in R uses Nelder-Mead optimization that is performed by several starting points in order to check for possible problem of local maximum and identification. We did not find any convergence problem in several applications using both empirical and simulated data.

The analysis of the Hessian matrix shows that for some specifications it is necessary a large number of observations to have a small variance of the estimator of the copula parameter. This is particularly true when the difference between the polyhazard model with dependence and the independent polyhazard model lies in a region with small probability. This is expected because we need a large number of observations in order to have a reasonable number of observations in the region.

Illustration: Unemployment duration data

We ran some simulations which produced good results but present only the resultd for the unemployment duration data set, which was previously studied by Wichert and Wilke (2008), where it was described as: "it is a sample of German administrative individual unemployment duration data. it is extracted from the IAB-Employment Sample 1975-2001 (IABS-R01) which contains employment trajectories of about 1.1 million individuals from West-Germany and about 200K individuals from East-Germany. It is a 2% random sample of the socially insured workforce." There are two basic benefits related to unemployment, the unemployment benefit and the unemployment assistance. The unemployment benefit is granted at the begining of the state of individual's unemployment, and may lasts, by the time of the data, from six to 32 months. The benefit has mechanisms to incentivate the insured individual to return to the job market, for instance, by suspending the benefit for a person who refuse a job offer that pays a sallary that is compatible with the last job. The unemployment assistance may be granted right after the end of the unemployment benefit, it has additional criteria for eligibility, and its value is lower than unemployment benefit and lasts indefinitely in time.

The available information is the duration of the withdrawals of an individual by one of the benefits or both. Therefore, it is only known the date when an individual began and finished his or her withdrawals by the unemployment insurance. The end of the benefit may occur due to several causes as emigration, finding another job or even starting business, but this information is not available. Thus, we believe that there are risks competing for the end of the unemployment duration of an individual. We considered as censored when the woman was still unemployed at the end of the observation period, the year of 2001, or she was unemployed at the end of the benefit duration's period. Only the 8,109 observations of women in the data set were used. There are 15.8% censored observations.

Table 1 shows the estimates for the best AIC polyhazard models specifications of each copula fitted to the unemployment data as well as the single risk models and the Figure 3 presents the estimates of the density, hazard and survival functions. The polyhazard models exhibit a good fit to the data, which is clearly better than the single risk models fit. The estimated hazard function has a sharp values at the beginning with a maximum around 1,4 months and decline with a minimum around one year and four months and increases again. Except for the model with Frank copula the estimates show a dependence between the latent variables. Independently of the model the estimates of the density, hazard and survival functions are very close, showing again (observed previously with simulated data) that the estimation of these function are robust to the model misspecification.

Final remarks

We showed that the dependent polyhazard models is a flexible way of constructing hazard functions. The use of copulas to model the dependence of the latent factors increased considerably this flexibility. With this generalized polyhazard models it is possible to construct a rich family of hazard rate functions with bathtub and multimodal shapes with local effects. The proposed model

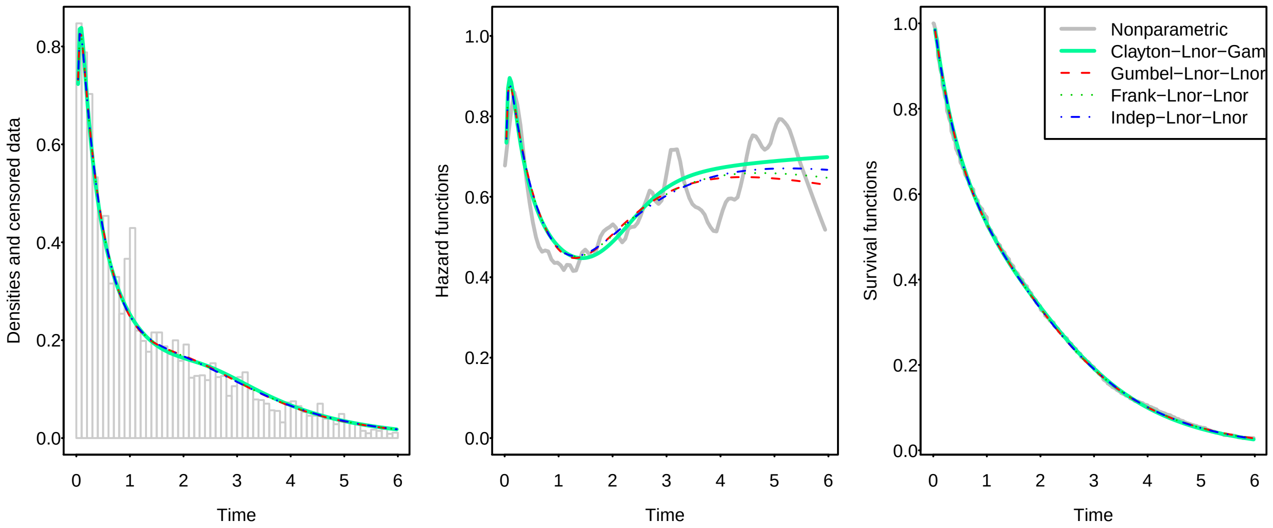

Figure 3: Density, hazard and survival functions of the models fitted to the women unemployment data. Polyhazard models of Table3.

Table 1: Summary of the models fitted to the Unemployment Data. Single risk models and selected polyhazard models. For each copula it is only presented the specification selected by the AIC criterion.

Model	AIC	$\tau = 0.90$	θ	Par-Marg1		Par-Marg2	
Clayton-lnor-gam	-20429.48	0.75	5.90 (0.79)	0.24 (0.043)	1.62 (0.030)	1.45 (0.060)	1.31 (0.048)
indep-lnor-lnor	-20434.62			0.13 (0.024)	1.65 (0.022)	1.33 (0.021)	0.48 (0.018)
Gumbel-lnor-lnor	-20436.03	0.53	2.14 (1.68)	0.13 (0.024)	1.65 (0.022)	0.85 (0.37)	0.55 (0.10)
Frank-lnor-lnor	-20436.46	-0.05	-0.43 (1.15)	0.13 (0.024)	1.65 (0.021)	1.38 (0.14)	0.49 (0.037)
-Wei-	-20822.76			1.66 (0.022)	0.92 (0.009)		
-gam-	-20832.89			0.88 (0.013)	1.95 (0.04)		
-exp-	-20906.22			1.70 (0.021)			
-lnor-	-21170.94			-0.08 (0.016)	1.40 (0.012)		
-llog-	-21333.81			0.99 (0.016)	1.23 (0.012)		

was applied to simulated data and to unemployment duration resulting in the presence of competing risks. Even when it is not possible to infer for the latent times due to the identification issue resulting of the lack of information of the cause of failure, the proposed model has a structure that allows to impose restrictions in the type of dependence, i.e., negative, positive or tail dependence, and also allows to associate covariates directly to the behavior of the latent times.

REFERENCES

Mazucheli, J. , Louzada-Neto, F. Achcar, J. A. (2001). Bayesian inference for polyhazard models in the presence of covariates. *Computational Statistics & Data Analysis* **38**, 1-14.

Nadarajah, S.S., Cordeiro, G. M. and Ortega, E. M. (2011). General results for the beta modified Weibull distribution. *Journal of Statistical Computation and Simulation.* DOI: 10.1080/00949651003796343

Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA* **7**2, 20-22.

Wichert, L. and Wilke, R. (2008). Simple nonparametric estimators for unemployment duration analysis. *Journal of the Royal Statistical Society - Series C* **57** 117-126.