Brazilian Journal of Probability and Statistics
2013, Vol. 27, No. 3, 357–376
DOI: 10.1214/12-BJPS185

Polyhazard models with dependent causes

Rodrigo Tsai[a,b] **and Luiz Koodi Hotta**[a]

[a]*State University of Campinas*
[b]*Superior Court of Justice*

Abstract. Polyhazard models constitute a flexible family for fitting lifetime data. The main advantages over single hazard models include the ability to represent hazard rate functions with unusual shapes and the ease of including covariates. The primary goal of this paper was to include dependence among the latent causes of failure by modeling dependence using copula functions. The choice of the copula function as well as the latent hazard functions results in a flexible class of survival functions that is able to represent hazard rate functions with unusual shapes, such as bathtub or multimodal curves, while also modeling local effects associated with competing risks. The model is applied to two sets of simulated data as well as to data representing the unemployment duration of a sample of socially insured German workers. Model identification and estimation are also discussed.

1 Introduction

Polyhazard models are a flexible family for fitting lifetime data. Their flexibility stems from the acknowledgment that there are latent causes of failure. There are many applied examples of these models in the literature. Kalbfleisch and Prentice (1980) proposed the poly-log-logistic model for log-logistic competing risks; Berger and Sun (1993) proposed the poly-Weibull model for Weibull competing risks; Louzada-Neto (1999) proposed a generalized polyhazard model which encompasses the poly-Weibull, poly-log-logistic and generalized-poly-gamma models; Kuo and Yang (2000) and Basu et al. (1999) used the poly-Weibull model to model masked-systems, in which the cause of failure may be unknown or partially known; Mazucheli et al. (2001) presented a Bayesian inference procedure for the polyhazard models with covariates; and Louzada-Neto et al. (2004) analyzed the identifiability of the poly-Weibull model. The main advantage of polyhazard models compared to single hazard models is the flexibility to represent hazard rate functions with unusual shapes.

In the applications cited above, the latent causes of failure are independent. In this paper, we extend the independent polyhazard models to encompass dependence modeled by copula functions. The model is general enough to allow for various forms of dependence and also for any marginal distributions for the latent times. The proposed models are able to generate much more flexible risk functions

Key words and phrases. Polyhazard models, copula, competing risks.
Received January 2011; accepted January 2012.

than the independent polyhazard models, including features such as bathtub shape, multimodality and local effects.

The literature also mentions another approach for constructing flexible hazard functions that is not pursued here. In this approach, the authors generalize known distributions. See, for instance, Pham and Lai (2007) and Nadarajah et al. (2011). The method proposed in the present paper, however, is more general. For instance, each of these distributions can be used as a marginal distribution for the latent causes.

The polyhazard model with dependence is proposed in Section 2. In Section 3 identification and estimation of the model through the maximum likelihood method is discussed. Another option would be to use a Bayesian approach; however, this is tangential to the purpose of this paper, as is model estimation, and thus is not discussed in detail. In Section 4 we present applications of simulated data and of data on unemployment duration of German women who are part of the socially secured workforce. General remarks are presented in Section 5.

2 The polyhazard model with dependence

Consider that we observe n units of observations, each one subject to $k \geq 2$ competing latent causes of failure. Let the lifetime related to the jth latent cause of the ith unit of observation, X_{ij}, have a density $f_j(\cdot;\Gamma_j)$, which are considered as known except for the unknown set of parameters Γ_j. Denote the survival and hazard functions by $S_j(\cdot;\Gamma_j)$ and $\lambda_j(\cdot;\Gamma_j)$, respectively. Only $X_i = \min\{X_{ij},\ j = 1,\ldots,k\}$ is observed for each unit of observation. Thus, considering the independence among risks, namely, among the failure times X_{ij}, $j = 1,\ldots,k$, the overall survival function of X_i, denoted by $S(t;\Upsilon)$, where $\Upsilon = (\Gamma_1,\ldots,\Gamma_k)$, is given for any $i = 1,\ldots,n$ by the product of marginal survival functions, that is,

$$\begin{aligned} S(t;\Upsilon) &= P_\Upsilon[X_i > t] \\ &= P_\Upsilon[X_{i1} > t,\ldots,X_{ik} > t] \\ &= \prod_{j=1}^{k} S_j(t;\Gamma_j), \end{aligned} \tag{2.1}$$

and the hazard function of X_i, $\lambda(t;\Upsilon)$, is given by the sum of the marginal hazards, because

$$\begin{aligned} \lambda(t;\Upsilon) &= -\frac{d}{dt}\prod_{j=1}^{k} S_j(t;\Gamma_j) \Big/ \prod_{j=1}^{k} S_j(t;\Gamma_j) \\ &= \sum_{j=1}^{k} \lambda_j(t;\Gamma_j). \end{aligned} \tag{2.2}$$

An example of an application of the independent polyhazard model is given in Mazucheli et al. (2001) where they estimate the poly-Weibull model with covariates using a Bayesian approach. In this paper, we model the failure time X_i with $k = 2$ competing risks, allowing for dependence between the risks. Henceforth, we use the notation for $k = 2$ for simplicity, but the notation for $k > 2$ can be easily generalized. Denoting by $H(\cdot, \cdot; \Upsilon)$ the joint distribution function and by $\bar{H}(\cdot, \cdot; \Upsilon)$ the joint survival function of the latent variables X_{i1} and X_{i2}, we can write the survival function of X_i as

$$\begin{aligned} S(t; \Upsilon) &= P_\Upsilon[X_{i1} > t, X_{i2} > t] \\ &= \bar{H}(t, t; \Upsilon). \end{aligned} \tag{2.3}$$

To model the joint survival function $\bar{H}$, considering dependence between the latent variables, we propose the use of 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 the multivariate modeling literature, especially when the use of the 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, thus making the copula a very convenient alternative in the case of multivariate modeling. Some references for copulas include the textbooks of Nelsen (2006), Joe (1997) and Cherubini et al. (2004) as well as the paper of Trivedi and Zimmer (2005).

Let $F_1(\cdot; \Gamma_1)$ and $F_2(\cdot; \Gamma_2)$ be the distribution functions of X_{i1} and X_{i2}, respectively. It follows from Sklar's theorem that there is always a copula function C^* such that we can write $H(t_1, t_2; \Upsilon) = C^*(F_1(t_1; \Gamma_1), F_2(t_2; \Gamma_2))$ and that 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 and F_2 to their joint distribution H. It is possible to represent the joint survival function directly by $\bar{H}(t_1, t_2; \Upsilon) = P[X_1 > t_1, X_2 > t_2; \Upsilon] = \tilde{C}(S_1(t_1; \Gamma_1), S_2(t_2; \Gamma_2))$, 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, $C(S_1(t_1; \Gamma_1), S_2(t_2; \Gamma_2))$ is a survival distribution function. Therefore, we can also model the survival function S directly by a copula function C as in Kaishev et al. (2007). This is also the approach adopted here because it is generally easier to work analytically with this representation. Then for the survival function of the polyhazard model with dependence given by a copula function C with dependence parameter θ and $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, we can write

$$\begin{aligned} S(t; \Upsilon) &= \bar{H}(t, t; \Upsilon) \\ &= C_\theta(S_1(t; \Gamma_1), S_2(t; \Gamma_2)), \end{aligned} \tag{2.4}$$

where S_1 and S_2 are, in this paper and in almost all practical applications, continuous marginal survival functions. The copula C in (2.4) is called the survival

copula; in this paper, we refer to it as 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 (2.4). From the survival function (2.4), it follows that the probability density and hazard rate functions for the polyhazard model with dependence are obtained in the usual fashion, that is,

$$f(t;\Upsilon) = -\frac{d}{dt}S(t;\Upsilon) \quad \text{and} \quad h(t;\Upsilon) = \frac{f(t;\Upsilon)}{S(t;\Upsilon)}. \tag{2.5}$$

The proposed model is a generalization of the independent polyhazard model in that we allow for dependence while at the same time modeling the marginal behavior of the latent risks. For each combination of copula and marginal survival functions employed, we have another model that allows for the construction of a rich family of competing risks latent models. For instance, in the following sections, we will work with exponential, log-logistic, log-normal, gamma and Weibull distributions for the latent failure causes and Clayton, Gumbel and Frank 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, although it is used as an example in some parts of the paper. These copula functions were selected because they have been widely used in the literature and have different types of dependence. The 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. While it can generate distributions with strong dependence in the center, the dependence in the tails is always small. Thus, in the tails, the hazard function of the competing risks model will be approximately equal to the sum of the 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 $\lambda_U \in [0, 1)$. These features must be taken into consideration when selecting the copula function [see Trivedi and Zimmer (2005) for more properties]. In the above discussion, we always referred to the dependence between the latent variables.

As an example of a specification of the polyhazard model with dependence, consider the Frank copula and Weibull latent failure times such that $X_{ij} \sim \text{Weibull}(\mu_j; \beta_j)$, $j = 1, 2$. This model will be referred to as Frank–Weibull–Weibull, where the first name stands for the copula function and the last two names denote the latent distributions. According to the notation of the proposed model, its parameters can be denoted by $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, where $\Gamma_1 = (\mu_1; \beta_1)$ and $\Gamma_2 = (\mu_2; \beta_2)$. The overall survival function of X_i is given by

$$S(t;\Upsilon) = -\frac{1}{\theta}\log\left(1 - \frac{(1 - e^{-\theta e^{-(t/\mu_1)^{\beta_1}}})(1 - e^{-\theta e^{-(t/\mu_2)^{\beta_2}}})}{(1 - e^{-\theta})}\right), \tag{2.6}$$

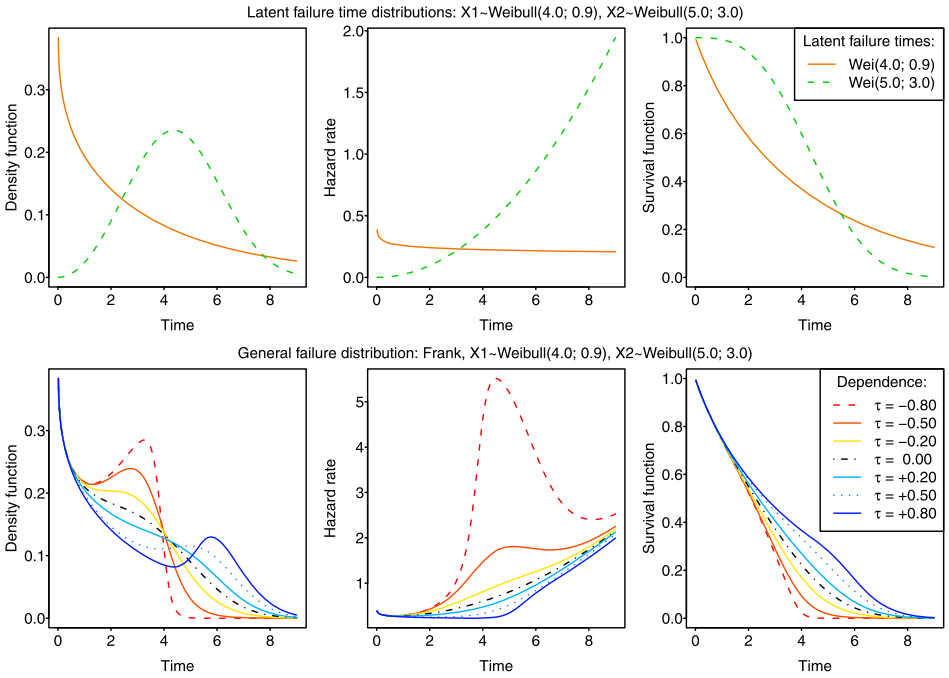

Figure 1 *Examples of density, hazard and survival functions for the single risk Weibull model and polyhazard model with Weibull marginals and dependence through Frank copula and Weibull marginals.*

and the probability density of X_i by

$$f(t;\Upsilon) = \frac{(1-e^{-\theta S_2(t)})e^{-\theta S_1(t)} f_1(t) + (1-e^{-\theta S_1(t)})e^{-\theta S_2(t)} f_2(t)}{(1-e^{-\theta}) - (1-e^{-\theta S_1(t)})(1-e^{-\theta S_2(t)})}, \qquad (2.7)$$

where f_1 and f_2 are the density functions of X_{i1} and X_{i2}, respectively. Figure 1 illustrates some possible shapes for the distribution of X_i for the Frank–Weibull–Weibull 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 from -0.80 to 0.80. The figure shows that different shapes for the hazard rates can result, depending on the shapes of the marginal distributions and the dependence type. Figure 2 shows various hazard rate functions for other specifications of the model in which it is possible to notice local effects and bathtub and multimodal shapes. The two points in the figure denote the 99% and 99.9% quantiles for each specification and the dependence parameter between the latent variables is the Kendall's τ, except for the SJC copula where they denote the lower and upper tail dependence. Henceforth, we use the acronyms Lnor, Llog, Exp, Wei, Gam and Indep for the log-normal, log-logistic, exponential, Weibull and gamma distributions and the independence copula, respectively, when referring to a specification of the poly-

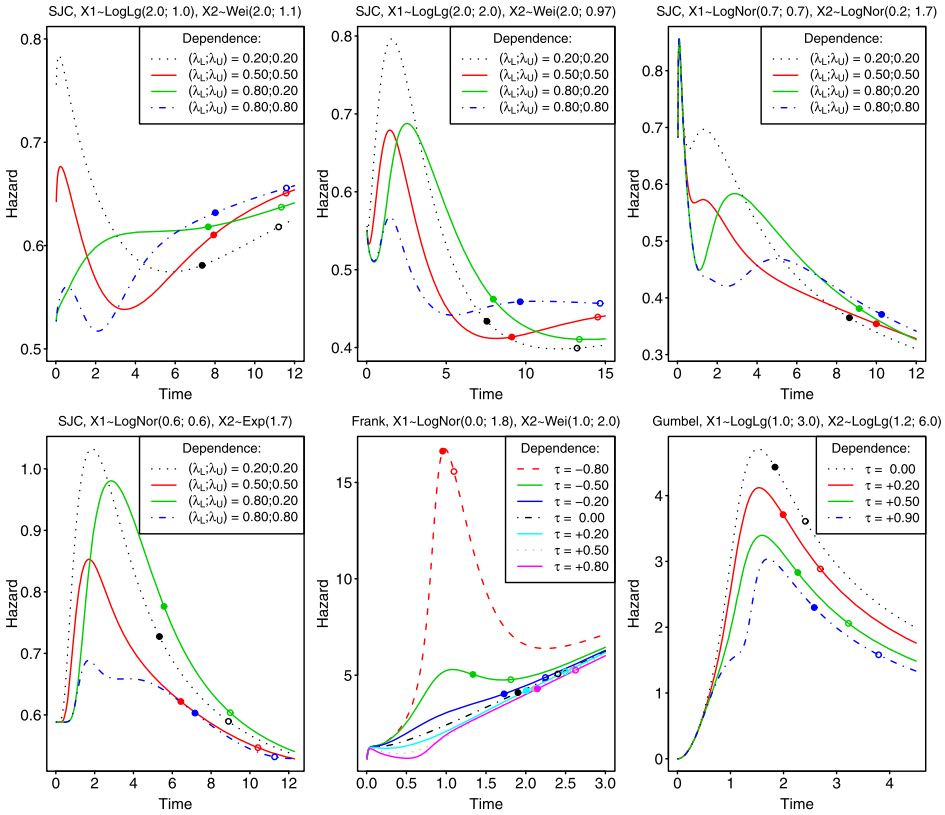

Figure 2 *Examples of hazard rate functions for the polyhazard model with dependence.*

hazard model. For instance, Clayton–Llog–Wei refers to a polyhazard model with the Clayton copula and log-logistic and Weibull latent variables.

3 Model identification and estimation

Some models are clearly nonidentifiable. Consider, for instance, the model Indep–Exp–Exp whose overall hazard function is constant, say, $\lambda > 0$, where the latent hazard function can be any non-negative constant, say, λ_1 and λ_2, such that $\lambda = \lambda_1 + \lambda_2$. A less trivial nonidentifiable model is the dependent polyhazard model Gumbel–Wei–Wei. The Gumbel copula function is given by

$$C(u, v) = \exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{1/\theta}], \qquad u, v \in [0, 1].$$

Therefore, by (2.4), considering Weibull margins with parameters functions (λ_1, β_1) and (λ_2, β_2), the overall survival function is given by

$$\begin{aligned} S(t) &= C(S_1(t), S_2(t)) \\ &= \exp[-\{\lambda_1^\theta t^{\theta\beta_1} + \lambda_2^\theta t^{\theta\beta_2}\}^{1/\theta}], \end{aligned}$$

showing that the model is not identifiable when $\beta_1 = \beta_2 = \beta$ for which any triple $(\lambda_1', \lambda_2', \theta')$ satisfying $(\lambda_1'^{\theta'} + \lambda_2'^{\theta'})^{1/\theta'} = (\lambda_1^{\theta} + \lambda_2^{\theta})^{1/\theta}$ can generate the same model. The same nonidentification problem occurs in the subclass of the Gumbel–Wei–Wei models: Gumbel–Exp–Exp, Indep–Exp–Exp and Indep–Wei–Wei models. Another example of a nonidentifiable model is the Clayton–Llog–Llog model when both marginal distributions have the same shape parameter and the dependence parameter equals 1. In general, we also have nonidentifiability when the distribution of one marginal latent variable is stochastically dominated by the other latent distribution and we use a copula with perfect positive dependence. Usually, it is not that easy to check whether a dependent polyhazard model is identifiable or not by analytical analysis. For this reason, identification of the other models, given by combinations of the Clayton, Gumbel and Frank copulas with the exponential, log-logistic, log-normal, gamma and Weibull latent cause distributions, was conducted by two types of numerical analyses. In the first analysis, the identification of each specification of the model was analyzed by means of an optimization procedure that searched over a region of parametric space for different points representing equal density functions. The analysis covered 1000 points that were sampled uniformly in a hyperspace that was a Cartesian product of individual parameter sets that were considered wide enough to represent the parametric space for real situations. For the dependence parameter, we considered the Kendall's τ in $[-0.99, 0.99]$ for the Frank copula and in $[0.01, 0.99]$ for the Clayton and Gumbel copula functions. For the latent variables parameters: exponential's scale in $[0.01; 4.00]$; gamma's form in $[0.01; 10]$ and scale $[0.01; 8]$; log-logistic's scale in $[0.01; 8]$ and form in $[0.01; 8]$; log-normal's location in $[-3; 3]$ and form in $[0.5; 3]$; and Weibull's scale in $[0.01; 10]$ and shape in $[0.01; 3]$. Then, for each of these 1000 points, its density function was evaluated in a grid of 301 points to serve as reference of a search of another point that could produce the same density function. Denote by $M(\Upsilon)$ the model under investigation, where Υ is its set of parameters in the parametric space E_Υ. For Υ_0, one of the 1000 arbitrarily chosen points in E_Υ, the overall density of $M(\Upsilon_0)$ was evaluated in a grid with 301 points, the $100(0.005 + 0.99i/300)\%$ quantiles, $i = 0, \ldots, 300$. The algorithm looked for a point in the parametric space that minimizes the objective function, $D(\Upsilon, \Upsilon_0)$, the sum of squared errors in which the errors were the differences between the density functions of $M(\Upsilon_0)$ and $M(\Upsilon)$ on the grid. For each Υ_0 the optimization step was repeated by 10 initial values, summing 10,000 cases, so that for the ith initial value denote by $\Upsilon_{0,i}$ the value located by the algorithm. After the optimization analysis the cases where $D(\Upsilon_0, \Upsilon_{0,i}) < 10^{-16}$ and $d(\Upsilon_0, \Upsilon_{0,i}) = \sum_{j=1}^{p} [(\upsilon_{0,i,j} - \upsilon_{0,j})/\upsilon_{0,j}]^2 > 10^{-10}$ were considered as indication of nonidentifiability, where $\Upsilon_0 = (\upsilon_{0,1}, \ldots, \upsilon_{0,p})$ and $\Upsilon_{0,i} = (\upsilon_{0,i,1}, \ldots, \upsilon_{0,i,p})$. Every case satisfying these conditions was analysed individually. The procedure detected the special cases of the Gumbel–Wei–Wei, Indep–Wei–Wei and Clayton–Llog–Llog models mentioned before as nonidentifiable. In the second analysis, in

the applications with the real data set and with the simulated data, we used different initial points, numbering approximately 200, for the optimization of the likelihood function in all cases. Except for a few cases of local maxima, the convergences were at the same values. In this study of convergence, we used more simulated data sets than the two presented in the illustration section. The analysis showed that, except for the cases mentioned previously, there was strong evidence of identification for all other specifications. A different point, estimability, is discussed more in the following paragraphs. An identifiable model does not ensure easy parameter estimation. For instance, when the overall hazard function is dominated by the first latent cause, it is very difficult to estimate the second latent cause, except for large samples.

In the traditional competing risks literature, when the cause of failure is known, there is another type of discussion of identification. See, for instance, Cox (1972) and Tsiatis (1975). In this classical problem, a competing risks model is identifiable if the joint survival function can be calculated or identified by the simple knowledge of the overall survival distribution. 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 on the behavior of the competing risks, this type of identification is not possible. Some papers exhibit results in this direction. Heckman and Honoré (1989) use a function that is similar to a copula based on covariates to overcome, nonparametrically, the identification problem. Carriere (1994) relates the marginal crude probabilities to the net probabilities using copula functions when there is dependence among the risks. Zheng and Klein (1995) show that the identification of the marginal distributions is possible if the copula function is fixed.

The polyhazard model can be seen as a competing risks model with missing values for the cause. Because less information is available, identification of the equivalent competing risks model is necessary but not sufficient for the identification of the polyhazard model. However, even when we have this type of nonidentification in polyhazard models, we can still use these models to model lifetime data and thus benefit from the good characteristics of these models.

The model parameters are estimated by the maximum likelihood method. Considering a random sample X_i, $i = 1, \ldots, n$, with random right censoring in which δ_i is the failure indicator variable and t_i the minimum value of the failure and censoring times, it follows from (2.4) and (2.5) 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 algorithms were written in R and the log-likelihood functions were implemented in C for fast computation. The optimization used the Nelder–Mead algorithm; in all applications, we tested for several initial parameter values to check

for possible problems of local maxima and identification. Except for the issue of local maxima observed in the estimation of the copula specifications, we did not find convergence problems in several applications using both empirical and simulated data.

The analysis of the Hessian matrix shows that, for some specifications, a large number of observations are necessary to have a small variance of the estimator of the copula parameter. This is especially important 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 a large number of overall observations are needed to have a reasonable number of observations in the region of small probability.

4 Illustrations

This section presents illustrations for simulated data, using two models and for the real data on the duration of female unemployment in Germany. For each data set, all models given by the combinations of the exponential, log-logistic, log-normal, gamma and Weibull distributions for the latent failure causes and the Clayton, Gumbel, Frank and Independent copulas were fitted, except for the Indep–Exp–Exp and Gumbel–Exp–Exp models, which are not identifiable. The exponential, log-logistic, log-normal, gamma and Weibull distributions, which are single risk models, were also fitted. Because there are many polyhazard models, we only present the fitting of some of these models. These include all single risk factor models and those polyhazard models selected according to the AIC criterion: the best specification for each copula function and for each data set for models with AIC comparable with that of the best model. In the simulations, we also included for each copula the model with the right marginal specification. We consider data sets with and without censored observations.

4.1 Simulated data

The first data set is a random sample of size $N = 5000$ from a Frank–Lnor–Wei model. The parameter of the Frank copula is given by $\theta = -5.74$, which gives Kendall's τ equal to -0.50. The Frank copula has both tail dependencies equal to zero. For the latent marginal distribution, we used log-normal ($\mu_1 = 0.6$; $\sigma_1 = 1.8$) and Weibull ($\mu_2 = 2.0$; $\beta_2 = 4.0$). A large sample size is necessary for this model to have sufficient observations in the right extreme tail. A random censoring mechanism was applied with uniform distribution $U(0; ax_{(n)})$, where $x_{(n)}$ is the maximum of the simulated latent values and $a = 5.3$. This resulted in 20% of the observations censored, while 43.1% of the observed data came from the first latent cause and 36.9% from the second cause. The upper panel of Figure 3 presents a graph where on the Y-axis, we have the cause of failure (1 for the first cause, 2 for the second cause and 3 if it is censored), and on the X-axis, we have the minimum of

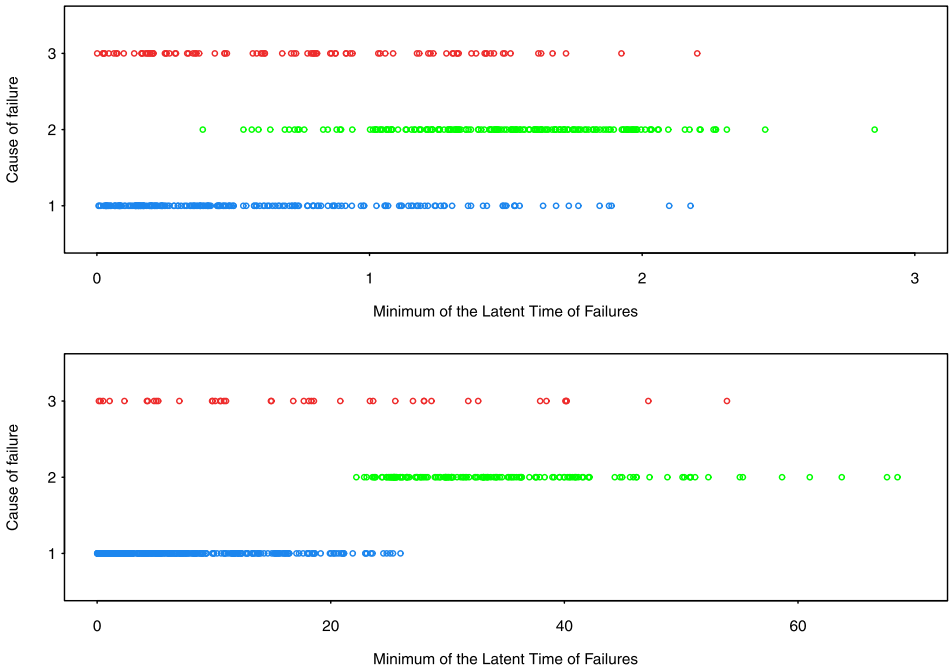

Figure 3 *Dot plot of the minimum between the latent times by cause of failure. Simulation* 1 *in the upper panel and Simulation* 2 *in the lower panel. Cause of failure*: 1: 1*st cause*; 2: 2*nd cause*; *and* 3: *censored value*.

the two latent failure times. We plotted only a sample of 500 observations to be able to visualize the points. We observe that almost all the smallest values came from the first latent cause, while for the large values, we have an inversion, although not as dominant as for small values. Table 1 presents the estimates of some single risk models and for the polyhazard models selected by the Akaike criterion. The Akaike criterion was calculated as $AIC = -2L(\hat{\Upsilon}) + 2k$, where k is the number of parameters and $L(\hat{\Upsilon})$ is the log-likelihood function evaluated at the maximum likelihood estimate. The parameters for the marginal distributions are as follows: exponential(scale); Weibull(form; scale); gamma(form; scale); log-logistic(scale; form) and log-normal(location; scale). The confidence intervals for the estimates are exhibited in parentheses and were calculated numerically from the Fisher information. In this example, the polyhazard models offered a better fit in terms of the AIC and in terms of adjustment to the nonparametric estimation of the density, hazard and survival functions relative to the single risk models. The first 4 models selected by the AIC criterion are Frank copula models (from a total of 63 models tested, 15 are Frank Copula models). In this simulated data set, selection of the right copula was likely facilitated due to the large sample size. Moreover, the Frank copula has no tail dependence and was generated with a negative Kendall' τ coefficient, while both the Gumbell and Clayton copulas have tail dependence and

Table 1 *Simulation* 1. *True model*: *Frank copula* $\theta = -5.74$ (*Kendall's* $\tau = -0.50$) *with* log-normal(0.6; 1.8) *and* Weibull(2.0; 4.0) *marginals and sample size equal to* 5000. *Single risk models and models selected by AIC criterion*: *best polyhazard models*, *best marginals configuration for each copula and the copula model for the right marginal configuration*

Model	AIC	τ	θ	Marginal distribution 1		Marginal distribution 2	
Frank–Lnor–Wei	7417.90	−0.51	−5.90	0.64	1.83	2.03	3.96
		(−0.62; −0.34)	(−8.46; −3.33)	(0.53; 0.76)	(1.74; 1.93)	(1.90; 2.15)	(3.35; 4.57)
Frank–Lnor–Gam	7418.84	−0.66	−9.79	0.61	1.81	7.23	0.29
		(−0.71; −0.60)	(−11.65; −7.93)	(0.50; 0.72)	(1.73; 1.90)	(5.98; 8.47)	(0.23; 0.34)
Frank–Lnor–Llog	7419.91	−0.67	−10.28	0.65	1.84	1.98	4.06
		(−0.72; −0.61)	(−12.29; −8.27)	(0.53; 0.76)	(1.75; 1.93)	(1.93; 2.04)	(3.65; 4.46)
Frank–Lnor–Lnor	7420.83	−0.70	−11.40	0.58	1.79	0.72	0.43
		(−0.74; −0.65)	(−13.46; −9.35)	(0.48; 0.68)	(1.71; 1.88)	(0.69; 0.74)	(0.39; 0.46)
Clayton–Lnor–Gam	7426.42	0.45	1.67	0.54	1.76	15.08	0.09
		(0.36; 0.53)	(1.11; 2.22)	(0.45; 0.63)	(1.69; 1.84)	(12.55; 17.62)	(0.08; 0.11)
Clayton–Lnor–Wei	7427.26	0.58	2.80	0.76	1.90	1.45	3.18
		(0.49; 0.65)	(1.90; 3.70)	(0.64; 0.88)	(1.81; 2.00)	(1.42; 1.48)	(2.96; 3.40)
Gumbel–Lnor–Wei	7427.31	0.39	1.65	0.60	1.80	1.53	3.54
		(0.10; 0.54)	(1.11; 2.19)	(0.49; 0.70)	(1.72; 1.89)	(1.45; 1.61)	(3.14; 3.93)
Indep–Lnor–Wei	7432.81			0.67	1.86	1.69	4.27
				(0.57; 0.78)	(1.77; 1.94)	(1.67; 1.72)	(4.06; 4.48)
Weibull	8605.86			1.22	1.51		
				(1.20; 1.25)	(1.47; 1.54)		
Gamma	8919.70			1.58	0.72		
				(1.52; 1.64)	(0.69; 0.76)		
Exponential	9407.66			1.19			
				(1.16; 1.23)			
Log-logistic	9825.31			0.95	1.77		
				(0.93; 0.98)	(1.73; 1.82)		
Log-normal	10,243.41			−0.18	1.08		
				(−0.21; −0.15)	(1.06; 1.10)		

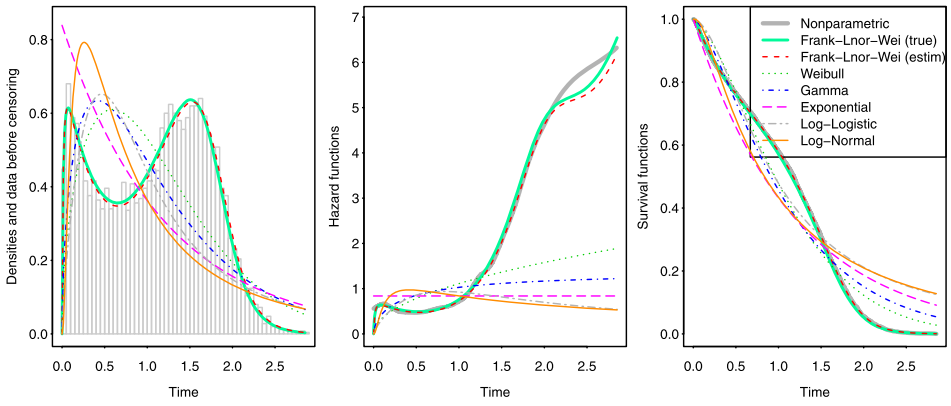

Figure 4 *Simulation* 1. *Comparison of the estimates of the density, hazard and survival functions by single risk models and by the polyhazard models of Table* 1.

positive Kendall' τ coefficients. Table 1 presents the estimation of the fitted models. We also included the first four best models selected by the AIC criteiron, all of which are Frank copula models. Observe that when the lognormal distribution is selected for the model, its estimates are not far from the true marginal distributions, even when the fitted copula is wrong or when the other marginal distribution is specified incorrectly.

Figure 4 presents the theoretical values of the density, hazard and survival functions and their estimates using single risk models, polyhazard models selected by AIC (Frank–Lnor–Wei) and using a nonparametric method. The nonparametric estimate of the survival function is the Kaplan–Meier survival function estimate smoothed by the R-program Loess method. To estimate the hazard function, the derivatives are numerically computed from the smoothed survival function and the Loess filter was again applied to the numerical derivatives. The smoothing parameter was selected empirically for each case. The estimation can depend strongly on the parameter, especially in the extremes. The nonparametric and the polyhazard function methods provide good estimates, while the single risk models are not able to fit the data. This first illustration clearly demonstrates the greater flexibility of the polyhazard models compared to the single risk models. Figure 5 presents the comparison of the fit of some polyhazard models. The estimates of the function density and survival function by all the polyhazard models selected by AIC criterion are close to the true functions. However, only the models with the Frank copula estimate the hazard function well for the entire period. The other specifications fail to fit the theoretical and nonparametric estimates of the hazard function in the right tail.

We used the same data set to fit the eight copula models of Table 1 without censoring and with 10% and 30% of the observations censored. Considering the cases of 10% and 20% of censoring, we have a total of 64 estimates of the risk parameters. Comparing with the estimates found without censoring, the maximum relative

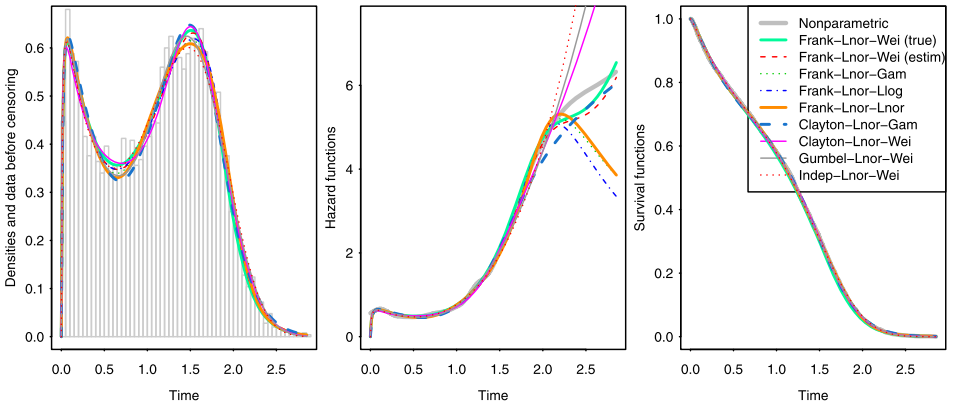

Figure 5 *Simulation* 1. *Comparison among the best copula model fitted. Density, hazard and survival functions for the polyhazard models of Table* 1.

difference was 6% for the point estimates and 23% for their standard deviations. These values were equal to 11% and 23% for the 16 estimates of the Kendall's τ and their standard deviations. The standard deviations were estimated using the delta method. For the 30% censoring the maximum relative difference in the 32 estimates was 98%, and 69% for the point estimates and their standard deviations, respectively. These differences, however, are smaller when we considered that the second last differences were equal to 32%, and 47%.

The same exercise was repeated with sample sizes N equal to 100, 250, 500, 1000, 2000 and 10,000 without censoring and with 20% of censored observations. For every case, the single risk model yielded a bad fit. The AIC selected a dependent copula over the independent copula only when the sample size was larger or equal to $N = 1000$. This is somewhat expected because the main difference between both models occurs in the extreme right tail. The hazard function has a change in the curvature around time 2.15 and another change around time 2.5. To detect this change in the curvatures, it is necessary to have some observations in this region. Thus, it is not surprising that even when we simulated a sample as large as 2000, the estimated hazard function was not accurate at the extreme because, without censoring, the probability of observing a failure larger than 2.5 is 0.0039. That is the main reason why, in this first example, we used a large sample. The estimation of the probability density and survival functions requires fewer observations. For instance, Figure 6 presents the estimation of the same Frank–Lnor–Wei model with sample size equal to 500, 1000, 2000 and 5000, without censoring. All the estimates are close to the theoretical values.

The second example is a random sample of size $N = 1000$ from a model with Clayton copula with parameter $\theta = 18$ (Kendall's $\tau = 0.90$, $\lambda_U = 0.96$ and $\lambda_L = 0$) and log-logistic(13.0; 1.0) and log-normal(3.0; 0.5) latent marginals. More than half (61.6%) of the observed data were obtained from the first latent cause and

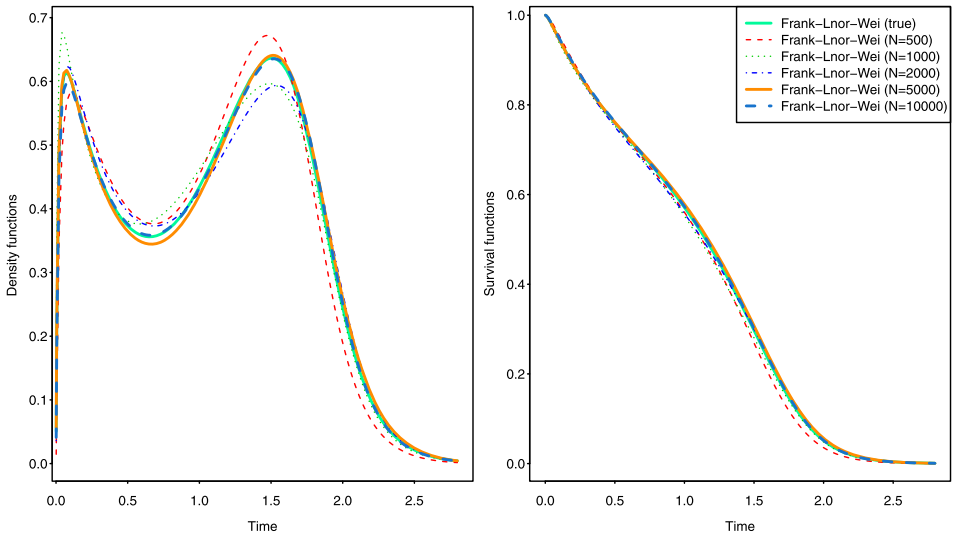

Figure 6 *Simulation* 1. *Frank–Lnor–Wei model fitted to samples of sizes* 500, 1000, 2000, 5000 *and* 10,000.

31.5% were obtained from the second cause. The censoring mechanism was the same as in the previous case with $a = 3$ producing 6.9% of censored observations. The lower panel of Figure 3 presents the same graph as in the first simulated data set, also including only 500 observations. Almost all the smallest values came from the first latent cause and there is less mixture in the middle in comparison with the first example. Table 2 shows the estimates for the polyhazard models with the best fit for each copula according to the Akaike criterion and the models of single risk. Except for the independent copula, which was ranked only 19th in terms of AIC, the other polyhazard models produced a fit close to the nonparametric hazard function estimate. Because the single risk model again produced a bad fit, in Figure 7, we present only the results for the polyhazard models of Table 2. In this example, it is observed that when one or both of the marginals are correctly specified, the parameter estimates of the correctly specified variables are very close to their true value. In this example, we also observed the same facts we observed in respect to the estimates of the marginal distributions and the effect of censoring.

Similarly to the first example, we fitted models with different sample sizes, with and without censoring. The copula parameter was often estimated in the border of the parametric space for sample sizes up to 500. The result was worst with censoring, when in many cases the independent copula was selected by the AIC criterion. When the sample size was increased to 1000, the AIC criterion seldom selected the independent copula, and the nonparametric and the parametric estimation (by the correct Clayton–Llog–Lnor model) were close to the theoretical hazard function. Even when the wrong copula was fitted, the fit was good, except in the right tail.

Table 2 *Simulation* 2. *True model*: *Clayton copula* $\theta = 18$ (*Kendall's* $\tau = 0.90$) *and* log-logistic(13.0; 1.0) *and* log-normal(3.0; 0.5) *marginals*, *and sample size equal to* 1000. *Single risk models and models selected by AIC criterion*: *best polyhazard models*, *best marginals configuration for each copula and the copula model for the right marginal configuration*

Model	AIC	τ	θ	Marginal distribution 1		Marginal distribution 2	
Frank–Llog–Llog	7035.63	0.77	15.60	12.95	0.97	22.89	4.65
		(0.33; 0.87)	(3.24; 27.96)	(11.41; 14.49)	(0.90; 1.04)	(21.00; 24.78)	(3.93; 5.37)
Clayton–Llog–Llog	7035.72	0.76	6.28	13.03	0.97	22.38	4.55
		(0.54; 0.84)	(2.38; 10.17)	(11.46; 14.60)	(0.90; 1.04)	(21.03; 23.73)	(3.88; 5.22)
Gumbel–Llog–Llog	7035.73	0.83	5.87	12.94	0.97	22.88	4.70
		(−0.02; 0.91)	(0.98; 10.75)	(11.41; 14.46)	(0.90; 1.04)	(20.84; 24.92)	(3.85; 5.55)
Clayton–Llog–Lnor	7037.27	0.83	9.86	12.92	0.97	3.04	0.44
		(0.62; 0.89)	(3.21; 16.50)	(11.38; 14.45)	(0.90; 1.04)	(2.98; 3.10)	(0.39; 0.49)
Gumbel–Llog–Lnor	7037.69	0.91	10.98	12.85	0.98	3.04	0.44
		(0.62; 0.95)	(2.60; 19.35)	(11.35; 14.36)	(0.91; 1.04)	(2.98; 3.11)	(0.39; 0.49)
Frank–Llog–Lnor	7037.75	0.87	29.61	12.87	0.98	3.05	0.44
		(0.52; 0.93)	(6.03; 53.19)	(11.36; 14.37)	(0.91; 1.04)	(2.98; 3.11)	(0.39; 0.49)
Indep–Llog–Lnor	7049.22			32.79	5.85	2.64	1.88
				(31.39; 34.19)	(5.07; 6.62)	(2.50; 2.78)	(1.76; 2.01)
Gamma	7181.87			0.93	18.80		
				(0.86; 1.00)	(16.82; 20.77)		
Weibull	7184.91			17.30	0.98		
				(16.14; 18.47)	(0.93; 1.03)		
Exponential	7183.37			17.41			
				(16.29; 18.52)			
Log-logistic	7378.90			10.87	1.27		
				(9.93; 11.81)	(1.20; 1.33)		
Log-normal	7400.32			2.25	1.41		
				(2.16; 2.34)	(1.35; 1.48)		

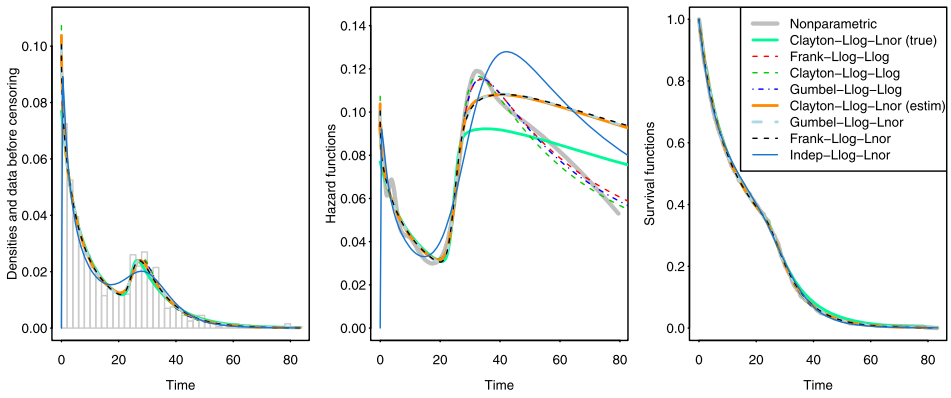

Figure 7 *Simulation* 2: *Comparison among the best copula model fitted. Density, hazard and survival functions for the polyhazard models of Table* 2.

The reasons for this are the same as in example 1: few observations in the extreme and incorrect tail dependency.

The simulation was also conducted with different copula parameter values. The copula parameter was chosen to have Kendall's τ equal to 0.7, 0.5 and 0.3. When τ is equal to 0.3 or 0.5, the likelihood of the models with independent copula was very close to that of models with dependent copulas, and, in general, the AIC criterion selected the independent copula. For $\tau = 0.7$, the AIC criterion almost always selected a dependent copula.

4.2 Unemployment duration data

The unemployment duration data set was previously studied by Wichert and Wilke (2008), who described it as follows: "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." At the time the data were collected, certain rules governed the administration of the two basic benefits related to unemployment: the unemployment benefit and unemployment assistance. The unemployment benefit was granted at the beginning of the individual's unemployment and could last from six to 32 months. The benefit had mechanisms to incentivize the insured individual's return to the job market, for instance, by suspending the benefit of a person who refused a job offer that would pay a salary comparable with that of his or her last job. The unemployment assistance could be granted immediately after the end of the unemployment benefit; it had additional criteria for eligibility, its value was lower than that of the unemployment benefit and it could last indefinitely in time.

The available data consist of the duration of the withdrawals of an individual from one or both of the benefits. Therefore, the date when an individual began and finished his or her withdrawals from the unemployment insurance is the only available measure. The end of the benefit may occur due to several causes, such as emigration, finding another job or starting a 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. Only the 8109 observations of women in the data set were used. We considered as censored observation cases when the woman was still unemployed by the end of the observation period (the year of 2001) or when she was unemployed when the benefit reached its maximum duration. There are 15.8% censored observations.

Table 3 shows the estimates for each copula for the best AIC polyhazard models fitted to the unemployment data; estimates for the single risk models are also provided. Estimates of the density, hazard and survival functions are presented in Figure 8. The polyhazard models exhibit a good fit to the data, and are clearly superior to the single risk models. The estimated hazard function has a peak at the beginning and a maximum at approximately 1.4 months, followed by a subsequent decline. A minimum value is reached at approximately one year and four months, after which the function increases again. Except for the model with the Frank copula, the estimates show dependence between the latent variables. Independently of the model, the estimates of the density, hazard and survival functions are very close, showing again that the estimation of these functions is robust to the model misspecification.

5 Final remarks

Independent polyhazard models are known to be a flexible tool for the construction of hazard functions. The use of copulas to model the dependence of the latent factors considerably increases this flexibility. With generalized polyhazard models, it is possible to construct a rich family of hazard rate functions with bathtub and multimodal shapes as well as local effects. The proposed model yields a strong fit to simulated data and unemployment duration data representing effects resulting from the presence of competing risks. Although it was not possible to infer the latent times due to the identification issue resulting from the lack of information about the cause of failure, the proposed model conveniently allows for restrictions on dependence (negative, positive or tail dependence), and also allows for the direct examination of the association between covariates and the behavior of the latent times.

Acknowledgments

The authors would like to thank two anonymous referee for carefully reading the paper and for their comments which greatly improved the paper. We also thank

Table 3 *Summary of the models fitted to the unemployment data. Single risk models and selected polyhazard models. For each copula, only the specification selected by the AIC criterion is presented*

Model	AIC	τ	θ	Marginal distribution 1		Marginal distribution 2	
Clayton–Lnor–Gam	20,429.48	0.75	5.90	0.24	1.62	1.45	1.31
		(0.68; 0.79)	(4.34; 7.45)	(0.15; 0.32)	(1.56; 1.68)	(1.33; 1.57)	(1.22; 1.41)
Gumbel–Lnor–Lnor	20,436.03	0.53	2.14	0.85	0.55	0.13	1.65
		(0.00; 0.82)	(1.00; 5.44)	(0.12; 1.58)	(0.35; 0.75)	(0.08; 0.17)	(1.61; 1.69)
Frank–Lnor–Lnor	20,436.46	−0.05	−0.44	1.38	0.50	0.13	1.65
		(−0.28; 0.19)	(−2.69; 1.81)	(1.11; 1.65)	(0.42; 0.57)	(0.08; 0.18)	(1.61; 1.69)
Indep–Lnor–Lnor	20,434.62			0.13	1.65	1.33	0.48
				(0.08; 0.18)	(1.61; 1.70)	(1.29; 1.37)	(0.45; 0.52)
Weibull	20,822.76			1.66	0.92		
				(1.62; 1.71)	(0.90; 0.93)		
Gamma	20,832.89			0.88	1.95		
				(0.86; 0.91)	(1.87; 2.03)		
Exponential	20,906.22			1.70			
				(1.66; 1.74)			
Log-normal	21,170.94			−0.08	1.40		
				(−0.11; −0.05)	(1.38; 1.42)		
Log-logistic	21,333.81			0.99	1.23		
				(0.96; 1.02)	(1.20; 1.25)		

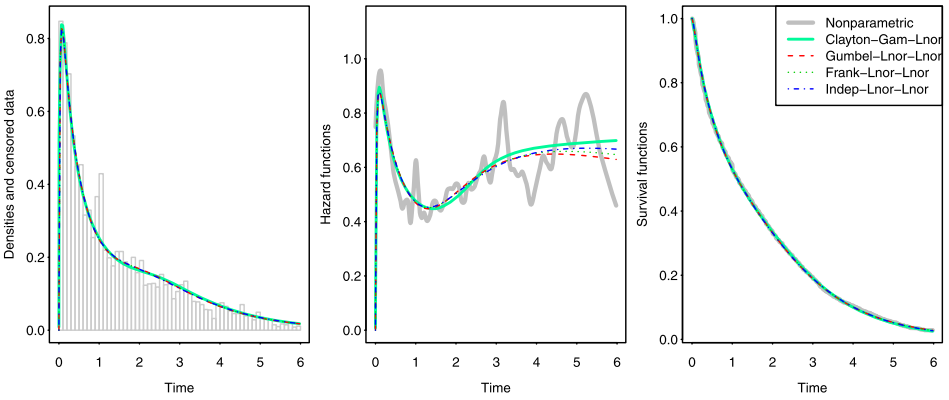

Figure 8 *Density, hazard and survival functions of the models fitted to the women unemployment data. Polyhazard models of Table 3.*

Epifisma Laboratory (UNICAMP). This work was partially supported by grants from CNPq, CAPES and FAPESP.

References

Basu, S., Basu, A. P. and Mukhopadhyay, C. (1999). Bayesian analysis for masked system failure data using non-identical Weibull models. *Journal of Statistical Planning and Inference* **78**, 255–275. MR1705552

Berger, J. M. and Sun, D. O. (1993). Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association* **88**, 1412–1418. MR1245378

Carriere, J. (1994). Dependent decrement theory. *Transactions of the Society of Actuaries* **46**, 45–74.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula Methods in Finance*. Chichester: Wiley. MR2250804

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Ser. B* **34**, 187–220. MR0341758

Heckman, J. J. and Honoré, B. E. (1989). The identifiability of the competing risks model. *Biometrika* **76**, 325–330. MR1016023

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall/CRC. MR1462613

Kaishev, V. K., Dimitrova, D. S. and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insurance Mathematics and Economics* **41**, 339–361. MR2364559

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley. MR0570114

Kuo, L. and Yang, T. M. (2000). Bayesian reliability modeling for masked system lifetime. *Statistics and Probability Letters* **47**, 229–241. MR1747483

Louzada-Neto, F. (1999). Polyhazard models for lifetime data. *Biometrics* **55**, 1281–1285.

Louzada-Neto, F., Andrade, C. S. and Almeida, F. R. Z. (2004). On the non-identifiability problem arising on the poly-Weibull model. *Communications in Statistics—Simulation and Computation* **33**(3), 541–552. MR2090953

Mazucheli, J., Louzada-Neto, F. and Achcar, J. A. (2001). Bayesian inference for polyhazard models in the presence of covariates. *Computational Statistics & Data Analysis* **38**, 1–14. MR1869477

Nadarajah, S., Cordeiro, G. M. and Ortega, E. M. M. (2011). General results for the beta modified Weibull distribution. *Journal of Statistical Computation and Simulation* **81**, 1211–1232.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed. New York: Springer. MR2197664

Pham, H. and Lai, C. D. (2007). On recent generalizations of the Weibull distribution. *IEEE Transactions on Reliability* **56**, 454–458.

Trivedi, P. K. and Zimmer, D. M. (2005). Copula modelling: An introduction for practitioners. *Foundations and Trends in Econometrics* **1**, 1–111.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA* **72**, 20–22. MR0356425

Wichert, L. and Wilke, R. A. (2008). Simple non-parametric estimators for unemployment duration analysis. *Journal of the Royal Statistical Society, Ser. C* **1**, 117–126. MR2412670

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82**(1), 127–138. MR1332844

IMECC-UNICAMP
State University of Campinas
Rua Sérgio Buarque de Holanda, 65 1
Campinas, São Paulo, 13083-970
Brazil
E-mail: hotta@ime.unicamp.br