Generalization of the Mixture Model Using a Copula Function

Rodrigo Tsai[1] and Luiz K. Hotta[2,3]
[1]Superior Court of Justice, Brazil
[2] University of Campinas, Brazil
[3]Corresponding author: Luiz K. Hotta, hotta@ime.unicamp.br

Abstract

In the mixture distribution model (for continuous and discrete cases) the density function, $\tilde{f}(t)$ is given by a linear combination of k density functions, $f_i(t)$, $i = 1, \cdots, k$, with non-negative weights p_i which must sum to 1.0. We propose a generalization of the mixture model where the weights are not restricted to being constant. The specification of the weight functions is not easy because $\tilde{f}(t)$ must be integrated to one. This is done by defining a random variable which has a density with the desired form. The definition of the random variable includes random variables with the densities $f_i(t)$ and a copula function. The proposed model includes the traditional mixture model, the polyhazard and the fraction of cure models. Real applications are used to illustrate the model.

Key Words: Distribution functions, generalized probability distributions, multimodal hazard functions

1 The model

The generalized mixture model using copula a function is defined as

Definition 1. *Consider U and V uniform random variables in $[0,1]$ such that $(U,V) \sim C$, where C is a copula function (Nelsen, 2006). Define a set A in the square $B = [0,1]^2$ and the distribution functions F and G. The generalized mixture model is defined by the random variable T*

$$T = F^{-1}(r(U,V))\, I_{[(U,V)\in A]} + G^{-1}(s(U,V))\, I_{[(U,V)\in A^c]}, \tag{1}$$

where r and s are functions such that $r(u,v), s(u,v) \in [0,1]$, when $(u,v) \in B$. If $P(r(U,V) = 0) > 0$, there is a_0 such that $F(a_0) = 0$. There is a similar restriction when $P(r(U,V) = 1) > 0$, and for $s(U,V)$. The model is defined by (F,G,C,A,r,s). □

In the following we consider that the distributions F and G are continuous. The distribution of the random variable T is given by:

$$\begin{aligned} P[T \le t] &= P[F^{-1}(r(U,V))I_{[(U,V)\in A]} + G^{-1}(s(U,V))I_{[(U,V)\in A^c]} \le t] \\ &= \int\int_{[(u,v)\in A; r(u,v)\le F(t)]} dC(u,v) + \int\int_{[(u,v)\in A^c; s(u,v)\le G(t)]} dC(u,v). \end{aligned} \tag{2}$$

The continuity of F and G is a necessary condition for the continuity of T. The sufficiency depends on the set A, the functions $r(u,v)$ and $s(u,v)$ and the copula function. When T is a continuous random variable, its density function is given by

$$\tilde{f}(t) = \frac{d}{dx}\left[\iint_{[(u,v)\in A; r(u,v)\leq x]} dC(u,v)\right]_{x=F(t)} f(t) + \frac{d}{dx}\left[\iint_{[(u,v)\in A^c; s(u,v)\leq x]} dC(u,v)\right]_{x=G(t)} g(t), \quad (3)$$

where $f(t)$ and $g(t)$ are respectively the density functions of $F(t)$ and $G(t)$. From (3) we can see that $\tilde{f}(t) = p_1(t)f(t) + p_2(t)g(t)$ with

$$p_1(t) = \frac{d}{dx}\left[\int\int_{[(u,v)\in A; r(u,v)\leq x]} dC(u,v)\right]_{x=F(t)} \text{ and } p_2(t) = \frac{d}{dx}\left[\int\int_{[(u,v)\in A^c; s(u,v)\leq x]} dC(u,v)\right]_{x=G(t)}. \quad (4)$$

The weights are non-negative, but the sum, at each point t is not restricted to 1.0. The definition is given for two distributions but the extension for more variables is mathematically trivial. When $F(x)$ and $G(x)$ are equal to zero for $x < 0$, T is a non-negative random variable and its distribution can be used to model survival data. The next section presents some examples of the generalized mixture model.

2 Examples

Example 1: The traditional mixture function.

In the traditional mixture model of distributions (Mclachlan and Peel, 2000), $F(t)$ and $G(t)$ with weights, respectively, equal to p and $1-p$ and with $p \in (0,1)$, the distribution function is given by:

$$P(T \leq t) = pF(t) + (1-p)G(t). \quad (5)$$

It is straightforward to see that

$$T = F^{-1}\left(\frac{U}{p}\right) I_{[U\leq p]} + G^{-1}\left(\frac{1-U}{1-p}\right) I_{[U>p]}, \quad (6)$$

with $A = [(u,v) \in [0,1]^2; u \leq p]$, $r(u,v) = u/p$, $s(u,v) = (1-u)/(1-p)$, for any copula function, has a distribution equal to (5), because the distributions of U/p conditional to $U \leq p$ and of $(1-U)/(1-p)$ conditional to $U > p$ are uniform $(0,1)$.

Example 2: Model with fraction of cure

The survival models with fraction of cure (Berkson and Gage, 1952) can be seen as mixture models. Define the random variable T as

$$T = F^{-1}(V) I_{[U\leq p]} + G^{-1}(V) I_{[U>p]}, \quad (7)$$

where $A = \{(u,v), u \leq p\}$, F is a distribution function such that $F(x) = 0$ when $x < \infty$ and G is a distribution function of a non-negative random variable. In this case

$$\begin{aligned} P[T \leq t] &= P[V \leq F(t), U \leq p] + P[V \leq G(t), U > p] \\ &= C(p, F(t)) + G(t) - C(p, G(t)) = G(t) - C(p, G(t)). \end{aligned} \tag{8}$$

When C is the independent copula, $P[T \leq t] = (1-p)G(t)$, following

$$P[T > t] = 1 - (1-p)G(t) = p + (1-p)\bar{G}(t), \tag{9}$$

where p is the fraction of cure, and $\bar{G}(t) = 1 - G(t)$ is the survival function of the non cured fraction of the population. The use of non-independent copula function allows having a fraction of cure varying in time.

Example 3: Polyhazard model

Define T as:

$$T = F^{-1}(U) I_{[F^{-1}(U) \leq G^{-1}(V)]} + G^{-1}(V) I_{[F^{-1}(U) > G^{-1}(V)]}, \tag{10}$$

where $A = \{(u,v); u \leq F \circ G^{-1}(v)\}$, $r(u,v) = u$, $s(u,v) = v$, and any copula C. We can write the random variable T as $T = \min(T_1, T_2)$, where $T_1 = F^{-1}(U)$ and $T_2 = G^{-1}(V)$ are random variables with distribution F and G, respectively. Thus, we have the traditional polyhazard model (Mazucheli et al., 2012).

When we use the independent copula, we have the traditional polyhazard model, and we have the dependent polyhazard model when we use a non independent copula. In this case, Tsai and Hotta (2011, 2012) model the joint probability directly by the copula function as

$$P(T \leq t) = 1 - P[\min(T_1, T_2) > t] = 1 - P[T_1 > t, T_2 > t] = 1 - C(F(t), G(t)).$$

When $A = \{F^{-1}(U) > G^{-1}(V)\}$, T is equal to the maximum of the latent times, showing the role of the choice of the region A.

Example 4: Examples of distributions

We present some example of distributions generated with latent distributions F and G with Weibull distributions with parameters equal to $(\mu_1; \beta_1) = (2,0; 3,0)$ and $(\mu_2; \beta_2) = (7,0; 6,0)$, respectively; Frank copula with Kendall's τ equal to $-0,8$, 0 (independent copula) and $0,8$; and $A = \{(u,v), u \leq v\}$. We consider three cases:

Model 1: $r(u,v) = u$, and $s(u,v) = v$,
Model 2: $r(u,v) = 1-u$, and $s(u,v) = v$, and

Model 3: $r(u,v)$ and $s(u,v)$ defined as in case 2, but changing the distributions F and G.

Figure 1 presents the density and weight functions of models (1) to (3) for the three values of Kendall's τ, while Figure 2 presents the density, hazard and survival functions of the same models. The figures show that the method is able to generate a rich family of distributions.

3 Final remarks

The main conclusions are:

- The density mixture models is known to be very flexible in constructing densities and (for positive random variables) hazard functions. The generalization of the model increased this flexibility considerably.
- With the generalized model it is possible to construct a rich family of hazard rate functions with bathtub and multimodal shapes with local effects.
- The proposed model fits empirical data sets well (not presented here).
- The maximum likelihood method performed well when applied to simulated and empirical data sets (not presented here).

Acknowledgements: This work was partially supported by CNPq, CAPES and FAPESP. The authors thank Laboratrio EPIFISMA (IMECC-UNICAMP).

References

[1] Berkson, J. and Gage, R. P. (1952) "Survival curves for cancer patients following treatment," *Journal of the American Statistical Association*, 47, 501-515.

[2] Mazucheli, J.J., Louzada-Neto, F. and Achcar, J.A. (2012). "The polysurvival model with long-term survivors,"' *Brazilian Journal of Probability and Statistics*, 26, 313-324.

[3] Mclachlan, G. and Peel, D. (2000) *Finite Mixture Models*, John Wiley, New York.

[4] Nelsen, R.B. (2006) *An Introduction to Copulas*, 2nd ed., Springer-Verlag, New York.

[5] Tsai, R. and Hotta, L. K. (2011) "Polyhazard models with dependent causes," *Brazilian Journal of Probability and Statistics*, forthcoming.

[6] Tsai, R. and Hotta, L. K. (2012) "Fitting Distributions with the Polyhazard Model with Dependence," *Communications in Statistics Theory and Methods*, forthcoming.

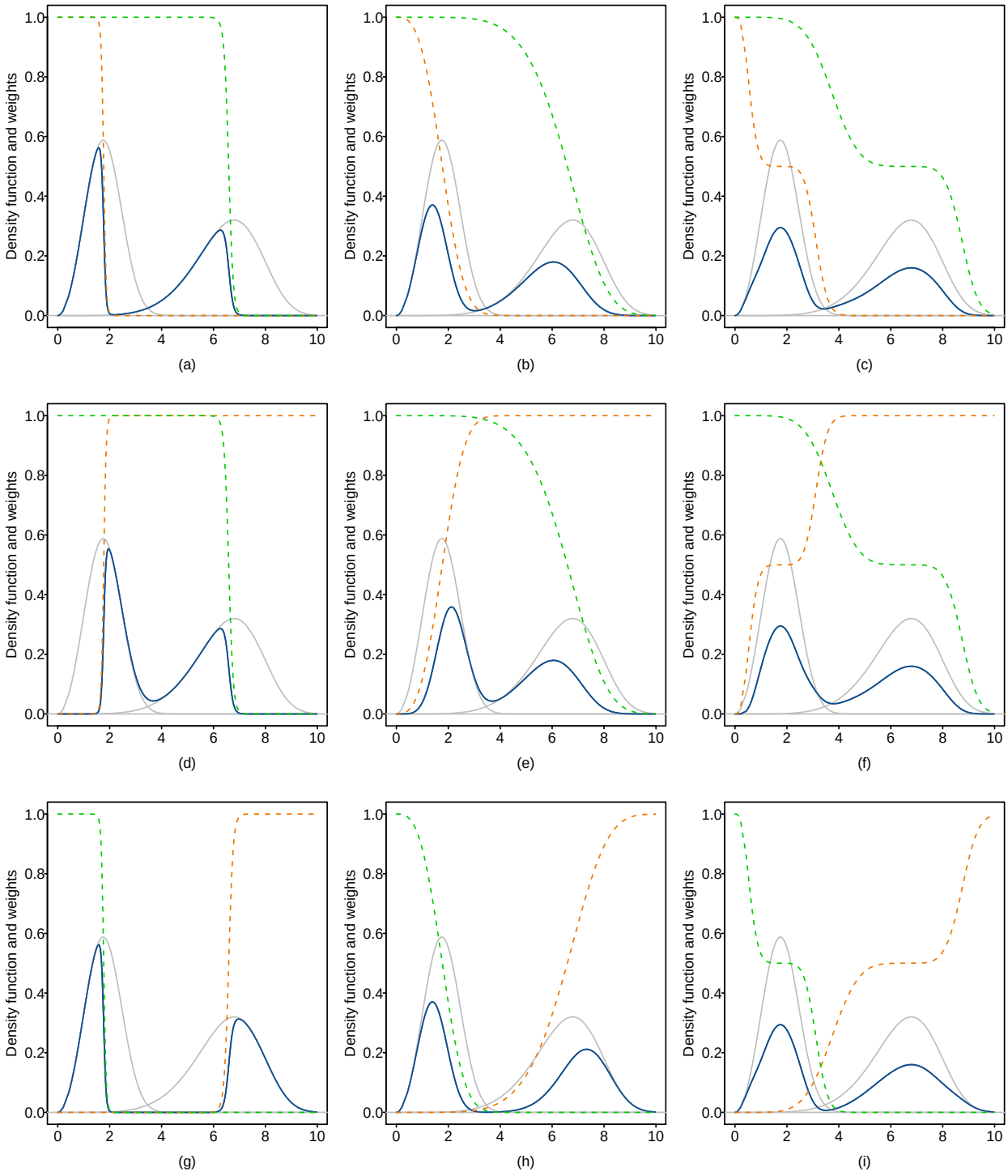

Figure 1: Density and weight functions in Example 4: models 1 ((a) to (c)), 2 ((d) to (f)) and 3 ((g) to (f)) for Kendall's τ equal to -0.8 on the left side, zero (independent copula) in the center and equal to 0.8 on the right side.

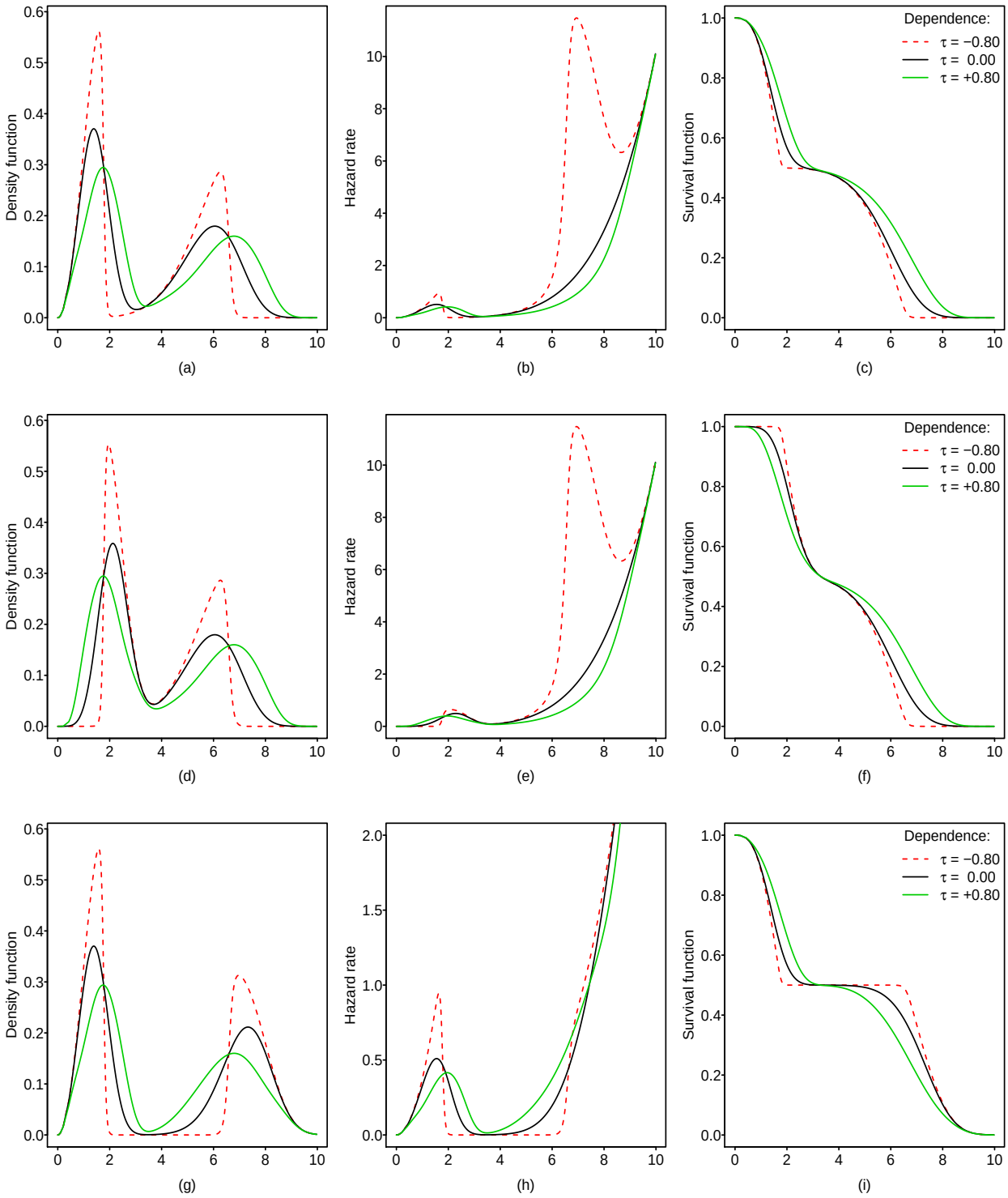

Figure 2: Density, hazard and survival functions in Example 4: models 1 ((a) to (c)), 2 ((d) to (f)) and 3 ((g) to (f)) for Kendall's τ equal to -0.8, zero (independent copula) and 0.8.