D. SURVIVAL COPULA

D.I. Survival models and copulas

Definitions, relationships with multivariate survival distribution functions and relationships between copulas and survival copulas.

D.II. Frailty models

Use of a latent variable to introduce dependence between survival times.

Link with Archimedean copula

D.III. Dependence measures

Particular care should be paid when measuring dependence among survival times.

Properties of Kendall's tau, Spearman's rho and Tail dependences in a survival setting.

D.IV. Competing risk models

Definition and properties

D.V. Estimation

Problems of censoring and truncation.

D.VI. Conclusions

D.I. Survival models and copulas

The term multivariate survival data covers the field where independence between survival times cannot be assumed.

We may parallel the construction of multivariate distribution through the use of copulas in a survival framework.

First we consider the univariate data separately in order to characterize the specific properties of the survival times.

Then we search to describe the joint behavior of the survival times by taking into account the properties exhibited in the first step.

a) Univariate survival notions

Let T denote a survival time with distribution F and density f.

The *survival function* is given by

$$S(t) = P\lfloor T > t \rfloor = 1 - F(t).$$

The *hazard rate* or risk function $\lambda(t)$ is defined as

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P[t \le T \le t + \Delta | T \ge t]}{\Delta}.$$

It can be interpreted as the instantaneous failure rate assuming the system has survived to time t.

It is given by $$\lambda(t) = \frac{f(t)}{S(t)}.$$

The *hazard function* is equal to

$$\Lambda(t) = \int_0^t \lambda(s)ds .$$

It is also known under the name: integrated hazard function or cumulative hazard function.

We get the relationship :

$$S(t) = \exp(-\Lambda(t))$$

In some cases we can incorporate explanatory variables in the modeling of $\lambda(t)$, and we have then

$$\lambda(t) = \exp(X\beta)\lambda_0(t)$$

where $\lambda_0(t)$ is called the *"baseline"* hazard function (Cox proportional hazard rate model).

b) Multivariate survival notions

The previous definitions can be extended to the multivariate case.

The multivariate survival function $S(t)$ is defined by

$$S(t_1,...,t_d) = P\left[T_1 > t_1,...,T_d > t_d\right]$$

where $T_1,...,T_d$ are d survival times with univariate survival functions $S_j(t_j)$.

We have $S_j(t_j) = S(0,...,0,t_j,0,...,0)$.

Note that $S(t_1,...,t_d) \neq 1 - F(t_1,...,t_d)$.

The density is simply

$$f(t_1,...,t_d)$$
$$= \partial_{1,...,d} F(t_1,...,t_d) = (-1)^d S(t_1,...,t_d)$$

Multivariate extensions of the hazard rate and the hazard function are given by

$$\lambda(t_1,\ldots,t_d)$$

$$= \lim_{\max \Delta_j \to 0} \frac{P[t_1 \le T_1 \le t_1 + \Delta_1,\ldots. \mid T_1 \ge t_1,\ldots]}{\Delta_1 \ldots \Delta_d}$$

or equivalently:

$$\lambda(t_1,\ldots,t_d) = \frac{f(t_1,\ldots,t_d)}{S(t_1,\ldots,t_d)}$$

and

$$\Lambda(t_1,\ldots,t_d) = \int_0^{t_1} \ldots \int_0^{t_d} \lambda(s_1,\ldots,s_d) ds_1 .. ds_d$$

Relationship between S and Λ cannot be simply formulated, since conditional hazard rates need to be taken into account.

Copulas are then a natural tools to develop multivariate survival functions from marginal univariate survival functions.

c) Survival copulas

A multivariate survival function S can be represented as follows :

$$S(t_1,...,t_d) = \overline{C}(S_1(t_1),...,S_d(t_d)),$$

where $\overline{C}$ is a copula (Sklar theorem for survival functions).

The survival copula $\overline{C}$ couples the joint survival function to its univariate margins in a manner completely analogous to the way a copula connects the joint distribution function to its margins.

There exists a link between the survival $\overline{C}$ and the copula C.

In the bivariate case it is given by

$$\overline{C}(u_1,u_2) = u_1 + u_2 - 1 + C(1-u_1,1-u_2)$$

Note that we can build a survival function as $S(t_1,...,t_d) = \overline{C}(S_1(t_1),...,S_d(t_d))$ or as $S(t_1,...,t_d) = C(S_1(t_1),...,S_d(t_d))$ for a given copula C.

This will not yield the same survival functions except in some cases.

For example it can be shown that for elliptical copulas $\overline{C} = C$ (normal, student). It is also true for the Frank copula.

Then it is equivalent to work with the copula or the survival copula.

D.II. Frailty models

The main idea is to introduce dependence between survival times $T_1,...,T_d$ by using an unobserved random variable W, called the frailty.

It corresponds to a latent (or hidden) variable modeling.

Given the frailty W with distribution G the survival times are assumed to be independent :

$$P\left[T_1 > t_1,...,T_d > t_d \middle| W = w\right]$$
$$= \prod_{j=1}^{d} P\left[T_j > t_j \middle| W = w\right]$$

We take then

$$S(t_1,\ldots,t_n|w) = \prod_{j=1}^{d} S(t_j|W=w) = \prod_{j=1}^{d}\left[\psi_j(t_j)\right]^w,$$

where $\psi_j(t_j)$ is the baseline survival function in a proportional hazard model:

$$\psi_j(t_j) = \exp(-\Lambda_i(t_j)) = \exp(-\int_0^{t_j}\lambda_i(s)ds)$$

The unconditional joint survival function is further defined as

$$S(t_1,\ldots,t_n) = E\left[S(t_1,\ldots,t_n|W)\right] = \int S(t_1,\ldots,t_n|w)dG(w)$$

We only need to integrate w.r.t. the distribution G.

It can be shown that a survival frailty copula is a special case of the construction based on

$$S(t_1,\ldots,t_d) = \overline{C}(S_1(t_1),\ldots,S_d(t_d))$$

where $\overline{C}$ is an Archimedean copula with a generator corresponding to the inverse of the Laplace transform of the distribution of the frailty variable.

Remark that frailty models exhibit a PQD behavior only, which might be an handicap for the modeling of some data.

Recall that an Archimedean copula is such that

$$C(u_1,u_2) = \varphi^{-1}(\varphi(u_1)+\varphi(u_2))$$

where φ is called the generator of the copula.

The name Archimedean comes from one of the mathematical property of this category of copula which is related to the Archimedean axiom: if a,b are positive real numbers, then there exists an integer n such that $na>b$.

Examples are the Frank copula and the Gumbel copula.

They find a wide range of applications since (1) they are easy to construct, (2) there is a large variety of copula families which belong to this class, (3) they have nice mathematical properties.

The high degree of analytical tractability of the class is an advantage, but the number of free parameters is typically low.

This might become an handicap in high dimensions when the dependence structure of the data is complex.

D.III. Dependence measures

a) linear correlation

The traditional way of evaluating dependence in a bivariate distribution is by means of the standard correlation coefficient.

This measure of dependence is natural and unproblematic in the class of elliptical distributions, but it might be misleading in other contexts, typically encountered in survival data.

Here are some usual misinterpretations of the Pearson correlation (counter-examples may be given).

1. T_1 and T_2 are independent if and only if $\text{corr}(T_1, T_2) = 0$.

2. $\text{corr}(T_1, T_2) = 0$ means that there is no perfect dependence between T_1 and T_2.

3. for given margins, the permissible range of $\text{corr}(T_1, T_2)$ is [-1,1].

Survival data are typically positive. Hence the lower bound –1 can never been reached.

It is further difficult to obtain large range of correlation because of the type of distributions generally used in survival modeling. For the Weibull, the interval is often [-1/3,1/2] only.

b) Kendall's tau and Spearman's rho

The Kendall's tau and Spearman's rho of the survival copula and its associated copula are equal.

c) Tail dependence

Tail dependence measures correspond to

Upper tail dependence:

$$\lambda_U = \lim_{u \to 1} P\left[U_2 > u \middle| U_1 > u\right]$$

If $\lambda_U \in (0,1]$, then upper tail dependence.
If $\lambda_U = 0$, then no upper tail dependence.

Lower tail dependence:

$$\lambda_L = \lim_{u \to 0} P\left[U_2 < u \middle| U_1 < u\right]$$

If $\lambda_L \in (0,1]$, then lower tail dependence.
If $\lambda_L = 0$, then no lower tail dependence.

The upper tail dependence of the survival copula will give the lower tail dependence of its associated copula, and vice-versa.

Lower tail dependence in survival copula will characterize “immediate joint death”, while upper tail dependence in survival copula will characterize “long-term joint survival”.

Remark: Normal copula has no upper or lower tail dependence. Student copula may.

D.IV. Competing risk models

Competing risk models correspond to the study of any failure process in which there are different causes of failures.

Let us consider d survival times $T_1,...,T_d$. In a competing risk model the survival time τ is defined by

$$\tau = \min(T_1,...,T_d).$$

We have then

$$S_\tau(t) = P\left[\min(T_1,\ldots,T_d) \geq t\right]$$
$$= \overline{C}(S_1(t),\ldots,S_d(t))$$

The cdf of the survival time τ is

$$F_\tau(t) = 1 - \overline{C}(S_1(t),\ldots,S_d(t))$$
$$= 1 - \overline{C}(1 - F_1(t),\ldots,1 - F_d(t))$$

and its density is given by

$$f_\tau(t) = \sum_{i=1}^{d} \partial_i \overline{C}(S_1(t),\ldots,S_d(t)) f_i(t)$$

Explicit forms can be found for example for Weibull margins and a Gumbel copulas.

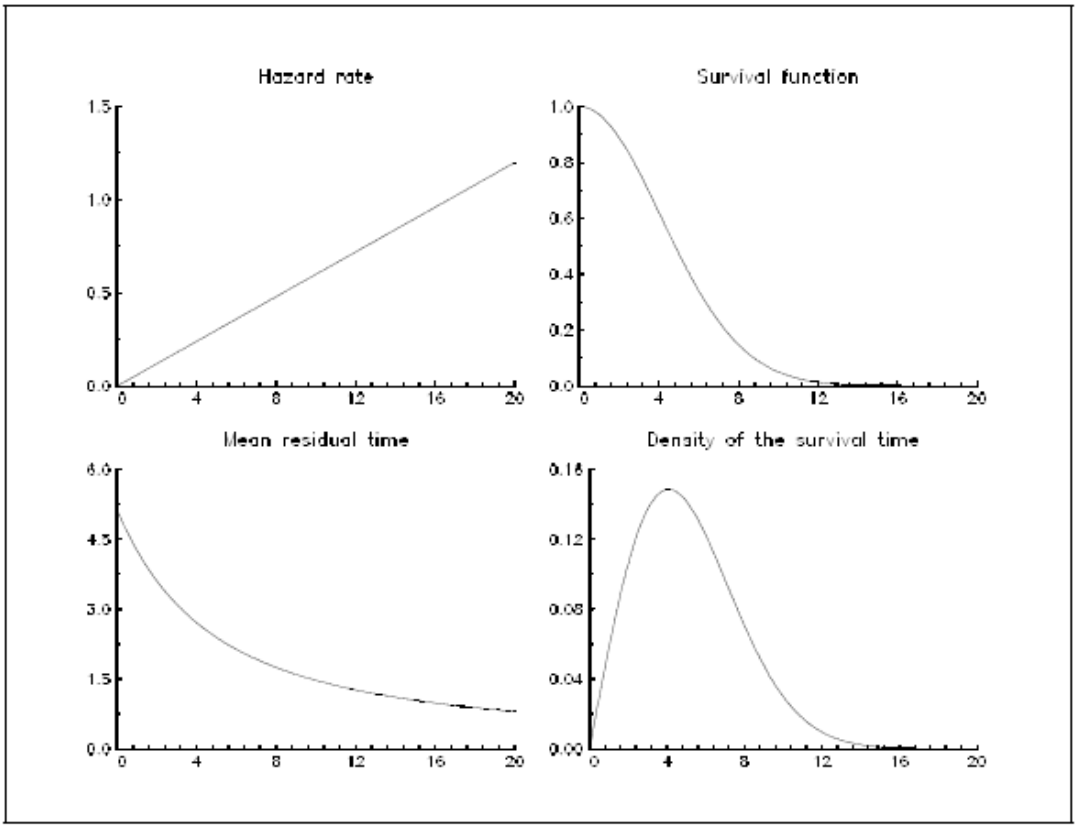

Figure 5: Weibull $(0.03, 2)$ survival time

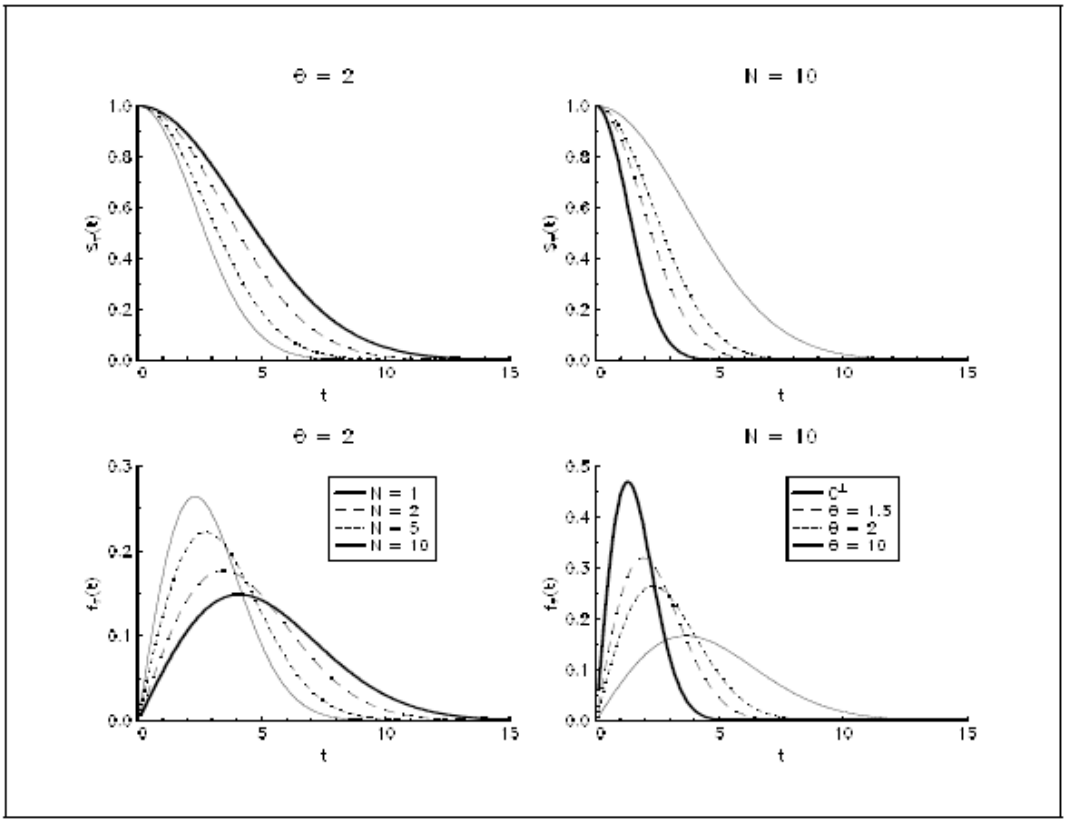

Figure 6: Failure time with Weibull $(0.03, 2)$ survival times

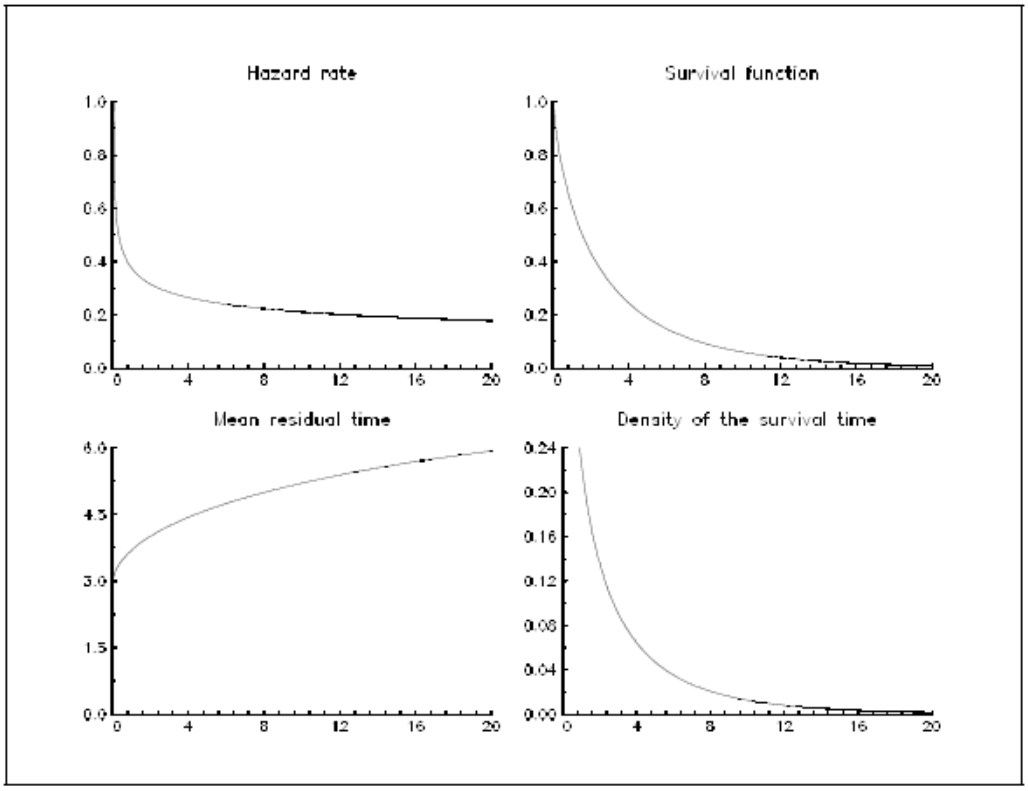

Figure 7: Weibull $(0.5, 0.75)$ survival time

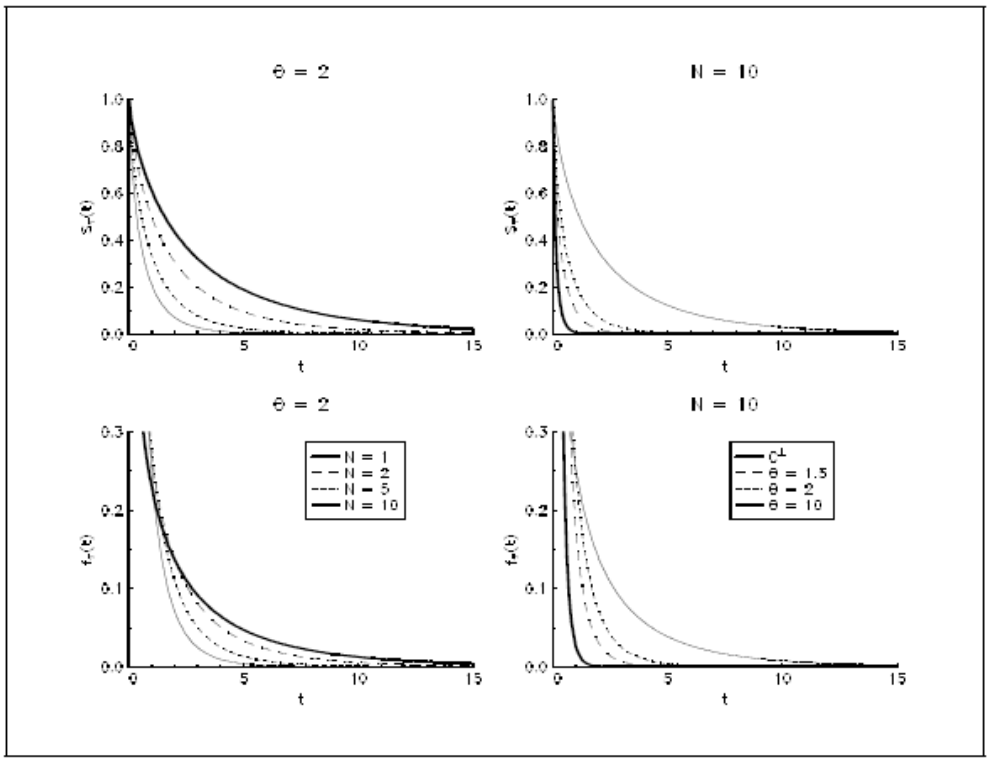

Figure 8: Failure time with Weibull $(0.5, 0.75)$ survival times

Under an iid scheme we get

$$F_\tau(t) = 1 - (1 - F_1(t))^d$$

and $$f_\tau(t) = d(1 - F_1(t))^{d-1} f_1(t)$$

D.V. Estimation

The estimation by maximum likelihood are exactly the same as before when observations are complete.

Indeed ML estimation relies on the joint density of the survival times.

However dealing with survival times is not as simple, because records on survival traits are often incomplete: survival data are often censored or truncated.

Under *left truncation* we only observe data above a fixed threshold. We have no information about the behavior below the limit (only reported losses above a given level).

Under *censoring* we have usually a mixture between complete and incomplete data.

For example under right censoring we observe T if it is below a threshold C or the threshold C itself if it is above. The threshold C may be fixed or random.

Estimation under these schemes are much more difficult, especially when dealing with nonparametric estimation.

For example under left truncation it is impossible to identify nonparametrically the part of the distribution below the threshold (we have no information!).

D.VI. Conclusions

The joint behavior of survival times can be easily modeled through copulas.

It is a powerful tool to analyze the dependence structure among these data, especially because symmetric distributions are not natural candidates for these data.

Estimation procedures are also available in such a setting but are more difficult to implement when censoring or truncation mechanisms are present.