"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2016, 606-623 | Science Signpost Publishing

On Statistical Analysis of Competing Risks with Application to the Time of First Goal

Petr Volf

Inst. of Information Theory and Automation, Czech Ac. Sci., Prague 8, Czech Republic.

Abstract

The contribution deals with the analysis and application of the competing risks model. First, the problem of identifiability of marginal and joint distributions of competing random variables is discussed. Then, the notion of copula is recalled and used to express the dependence in the competing risks scheme. Copula model is then utilized to the analysis of times of scoring the first goal in a football (soccer) match. It is assumed that competing latent times are exponentially distributed with parameters depending on attack and defence strengths of teams, while their mutual dependence is described with the aid of a conveniently chosen copula ensuring the model identifiability. As a real example the data from the season 2014-2015 of the Czech premier league (called Synot League) are analyzed. It is shown that the correlation in the framework of proposed model is, as a rule, negative, and its value is related to the first goal importance. The outcomes of analysis are discussed both from theoretical and practical point of view.

Keywords: survival analysis, competing risks, copula, sports statistics.

1. Introduction

The interest in the problem of competing risks dates back to 70-ties of the last century. It arises often in the field of survival analysis, namely in reliability, biostatistics and medical studies, simultaneously it is also studied in demography, labor statistics, insurance, and in econometrics generally. From the beginning it was revealed that in the competing risks setting the background model may not be identifiable. A proof and an example of this phenomenon is given in Tsiatis (1975), some instances of identifiable (or not)

Corresponding author: Petr Volf, Inst. of Information Theory and Automation, Czech Ac. Sci., Prague 8, Czech Republic. E-mail: volf@utia.cas.cz.

models are presented in Basu and Ghosh (1978). In these classical studies the notion of copula has not been used yet. Just later it was recognized that the use of copula for multi-dimensional continuous distribution can lead to a 'nice' closed form of the model.

In the present contribution we shall apply the approach to the analysis of distribution of time to the first scored goal in a football (soccer) match. Consequently, we shall deal with continuous-type distributions of random times and with the scheme of competing risks of two random variables, two latent times to scoring the first goal of two teams. The starting point will be a basic probabilistic model for final score of a football match proposed for instance in Maher (1982). It consists of two conditionally independent Poisson random variables. It means that they are dependent just through shared parameters or covariates. More flexible models are obtained by generalizations, for instance the distribution of number of scored goals can be inflated to cover better certain more frequent results. Another generalization can consist in considering changes or/and a time development of model parameters as well as covariates during the match (see for instance Dixon and Robinson, 1998, or Volf, 2009), in such a way a model based on random birth process or on time-continuous counting process scheme is obtained. The present contribution deals with another aspect of basic model improvement, namely with models considering an explicit form of dependence of both teams scoring distributions. Let us mention here at least two relevant papers. Thus, in Karlis and Ntzoufras (2003) a special case of bivariate Poisson distribution was employed. In the same context, McHale and Scarf (2011) have described the dependence with the aid of a copula model. Interesting is the comparison of conclusions of both approaches. While the correlation in the former model is non-negative (by definition), the latter paper concludes that the correlation is negative and is absolutely larger in more competitive matches. It has to be also said that the use of copula in discrete distribution models is not easy technically (and then computationally), because marginal distribution functions are as a rule expressed by sums of point probabilities, not having a reasonably closed form. In this aspect the present paper differs, it analyzes continuous random variables, namely the times.

The outline of the paper is the following: The next section recalls the scheme of competing risks and the problem of identifiability. Then, in Section 3, a copula model is formulated and the way of its application is proposed. A counter-example of a non-identifiable case is presented, too. As an alternative, a model based on Gauss copula is considered, later it is used to support the results of real data study. Section 4 then contains a real example, namely the analysis of data from the 2014-2015 season of Czech "Synot League" (the premier football league in the Czech Republic). The results are discussed, the impact of the first goal to the final match result is examined, too.

2. Competing Risks Model

The competing risks model assumes that certain event (e. g. a failure of a device) can be caused by K different reasons. Such a situation is then modeled by K (possibly dependent) random variables (random times, as a rule) $T_j, j = 1,...,K$, sometimes accompanied by a variable C of random right censoring. C is then independent of all T_j. Let $\overline{F}_K(t_1,...,t_K) = P(T_1 > t_1,...,T_K > t_K)$ be the joint survival function of $\{T_j\}$. However, instead the 'net' times T_j we standardly observe just 'crude' data (sometimes called also 'the identified minimum') $Z = \min(T_1,...,T_K, C)$ and the indicator $\delta = j$ if $Z = T_j$, $\delta = 0$ if $Z = C$.

The data structure described above allows a direct estimation of the distribution of $Z = \min(T_1,...,T_K)$, for instance its survival function $S(t) = P(Z > t) = \overline{F}_K(t,...,t)$. Further, we can estimate so called incidence densities

$$f_j^*(t) = dP(Z = t, \delta = j) = -\frac{\partial \overline{F}_K(t_1,...,t_K)}{\partial t_j} | (t_1 = ... = t_K = t),$$

and also their integrals, cumulative incidence functions

$$F_j^*(t) = \int_0^t f_j^*(s)\, ds = P(Z \le t, \delta = j).$$

Notice that $\lim F_j^*(t) = P(\delta = j) < 1$ if $t \to \infty$, $S(t) = 1 - \sum_{j=1}^K F_j^*(t)$.

Cumulative incidence functions, often called also the crude distribution functions, are estimable consistently by standard survival analysis methods, see for instance Lin (1997).

However, in general, from data $(Z_i, \delta_i), i = 1,\ldots,N$ it is not possible to identify neither marginal nor joint distribution of $\{T_j\}$. A. Tsiatis (1975) has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish the models. Namely, this 'independent' model is given by cause-specific hazard functions $h_j^*(t) = f_j^*(t) / S(t)$.

The situation can be better in a regression case, because the covariates provide an additional information, especially when their structure is rich enough. There are numerous results showing conditions

for full model identifiability, let us mention here Heckman and Honoré (1989) and their proof of identifiability in the Cox or the AFT model cases. Later on, Lee (2006) has studied more general transformation models of regression. However, all these studies rely on an assumption that the dependence structure does not change with covariates. If it is not the case, the problem of identifiability arises anew.

As a consequence of the Tsiatis (1975) result, in competing risks models without regressors it is necessary to make certain functional form assumptions about the type of both marginal and joint distribution in order to identify them. Several such cases are studied in Basu and Ghosh (1978) and in some other papers. More recent results on identifiability can be found for example in Schwarz et al (2013) dealing with non-parametric setting, or in Escarela and Carriere (2003) considering Frank copula and parametric models.

3. Copula Model for Competing Risks Distribution

In the sequel we shall consider just 2 competing events, i.e. random variables S,T and data $Z_i = min(S_i, T_i, C_i), \delta_i = 1,2,0$. Then the notion of copula offers a way how to model bivariate distributions, namely the joint distribution function $F_2(s,t)$ of S,T:

$$F_2(s,t) = C(F_S(s), F_T(t), \theta), \tag{1}$$

where F_S, F_T are marginal distribution functions of S,T, $C(u,v,\theta)$ is a copula, i.e. a two-dimensional distribution function on $[0,1]^2$, with uniformly on $[0,1]$ distributed marginals U,V, θ is a copula parameter. The parameter is, as a rule, uniquely connected with correlation of U,V, hence also with correlation of S,T. We can connect also density functions. Let $c(u,v)$ be the joint density of (U,V), then

$$c(u,v) = \frac{f_2(s,t)}{f_S(s) \cdot f_T(t)}, \tag{2}$$

with $u = F_S(s), v = F_T(t)$.

Sometimes we may prefer to use a copula as a "survival copula" for modeling the joint survival

function $\overline{F_2}(s,t)$ of S,T:

$$\overline{F_2}(s,t) = P(S > s, T > t) = C(\overline{F_S}(s), \overline{F_T}(t), \theta),$$

where $\overline{F_S}, \overline{F_T}$ are marginal survival functions of S,T. It is seen that the use of copula allows to model the dependence structure separately from the analysis of marginal distributions. From this point of view, the identifiability of the copula (and its parameter) and of marginal distributions can be considered as two separate steps.

Zheng and Klein (1995) proved that when the copula is known, the marginal distributions are estimable consistently (and then the joint distribution, too, from (1)), even in a non-parametric (so that quite general) setting. However, in general, the value of θ has to be known. The problem of proper copula choice is analyzed in a set of papers, let us mention here Kaishev et al (2007) comparing performance of several copula types. A common agreement is that the knowledge (or a good estimate) of parameter θ is much more crucial for correct model of joint distribution. As a consequence, because the knowledge of copula type is still an unrealistic supposition, we can try to use certain sufficiently flexible class of copulas, as approximation, and concentrate to reliable estimation of parameters.

3.1. Copula Based on Tsiatis' Example

Let us return to the example of Tsiatis (1975) considering $K = 2$ random variables S,T, both with exponential distribution, and the following marginal and joint survival functions,

$$\overline{F}_S(s) = e^{-\lambda s},\ \overline{F}_T(t) = e^{-\mu t},\ \overline{F}_2(s,t) = e^{-\lambda s - \mu t - \gamma st}.$$

Hence, $S(t) = \overline{F}_2(t,t) = \exp(-\lambda t - \mu t - \gamma t^2)$. Corresponding cause-specific hazard rates and their integrals are

$$h_S^*(t) = (\lambda + \gamma t),\ h_T^*(t) = (\mu + \gamma t),\ H_S^*(t) = (\lambda t + \frac{\gamma}{2} t^2),\ H_T^*(t) = (\mu t + \frac{\gamma}{2} t^2).$$

It follows that $S^*(t) = exp(-H_S^*(t) + H_T^*(t))$ is the same as $S(t)$ above, which means that independent random variables with marginal survival functions

$$\overline{G}_S(s) = e^{-\lambda s - \frac{\gamma}{2} s^2},\ \overline{G}_T(t) = e^{-\mu t - \frac{\gamma}{2} t^2}$$

yield the same competing risk scheme. Notice, however, that 'true' marginal distributions are exponential while derived independent distributions are not. It gives a chance that, when the type of marginals is

assumed, they (and parameter γ, too) can be estimated, uniquely. Tsiatis' example actually uses the following copula

$$C(u,v) = u \cdot v \cdot \exp(-\theta \cdot \ln u \cdot \ln v), \tag{3}$$

as survival copula connecting two exponential survival functions. Its parameter $\theta \geq 0$, corresponding correlation $\rho(U,V) \leq 0$, $\theta = 0$ means independence of U,V. In fact (3) belongs to a set of Archimedean copulas and is known as the Barnett copula. The parameters are connected in the following way: $\gamma = \theta \cdot \lambda \cdot \mu$. Figure 1 shows the dependence of correlation on parameters. The identifiability of model consisting of Barnett copula (3) and exponential marginal distributions has been proved already in Basu and Ghosh (1978). In Appendix we offer an alternative proof showing the uniqueness of maximum likelihood estimates.

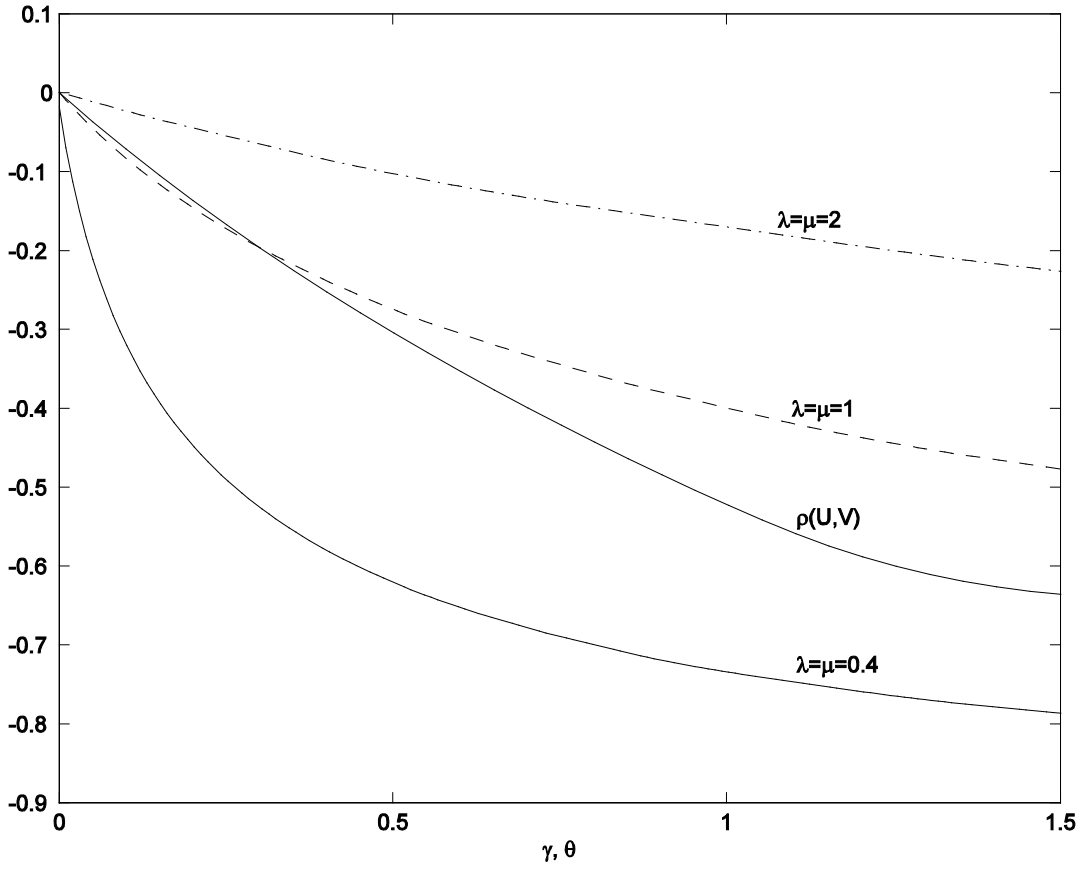

Figure 1. Dependence of $\rho(U,V)$ on parameter θ and $\rho(S,T)$ on γ, when $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$.

3.2. Other Forms of Two-Dimensional Exponential Distribution

We shall now show that the identifiability need not hold for other selection of copula type. For instance, let us consider the Gumbel copula

$$C(u,v) = \exp\{-[(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{1/\theta}\},$$

with $\theta \geq 1$. Here $\rho(U,V) \geq 0$, $\theta = 1$ corresponds to independence.

Let again $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$, then

$$\overline{F_2}(s,t) = \exp\{-[(\lambda s)^{\theta} + (\mu t)^{\theta}]^{1/\theta}\}, \ i.e. \ S(z) = \exp\{-[\lambda^{\theta} + \mu^{\theta}]^{1/\theta} \cdot z\},$$

$$f_S^*(z) = -\frac{\partial F_2}{\partial s}_{|s=t=z} = S(z)\frac{1}{\theta}[(\lambda^{\theta} + \mu^{\theta})^{1/\theta-1} z^{1-\theta}]\theta z^{\theta-1}\lambda^{\theta} = S(z)(\lambda^{\theta} + \mu^{\theta})^{1/\theta-1} \cdot \lambda^{\theta},$$

similarly $f_T^*(z) = S(z)(\lambda^{\theta} + \mu^{\theta})^{1/\theta-1} \cdot \mu^{\theta}$. After re-parametrization

$$\alpha = \lambda^{\theta}, \ \beta = \mu^{\theta}, \ A = (\alpha + \beta)^{1/\theta-1} \cdot \alpha, \ B = (\alpha + \beta)^{1/\theta-1} \cdot \beta,$$

we obtain that

$$S(z) = \exp\{-(A + B) \cdot z\}, \ f_S^*(z) = S(z) \cdot A, \ f_T^*(z) = S(z) \cdot B.$$

Hence, the model is determined fully by just two parameters, initial model is over-parametrized, we cannot estimate parameters λ, μ, θ uniquely. Notice also that A, B are the intensities of independent exponential distributions forming independent model forecasted by Tsiatis.

Another often used model for bivariate exponential distribution is the Marshall–Olkin model: Let X_1, X_2, X_3 be independent exponential random variables with parameters $\lambda_1, \lambda_2, \lambda_3$, respectively, set $S = \min(X_1, X_3)$ and $T = \min(X_2, X_3)$. Then marginal distributions of S, T are also exponential, with parameters $\lambda_1 + \lambda_3$, $\lambda_2 + \lambda_3$, resp., their correlation equals $\lambda_3 / (\lambda_1 + \lambda_2 + \lambda_3)$. However, as $P(S = T) = \lambda_3 / (\lambda_1 + \lambda_2 + \lambda_3)$, too, the joint distribution of S, T is not of continuous type and, therefore, is not convenient for our purposes. Let us note that this distribution is closely connected with bivariate Poisson model used for instance in Karlis and Ntzoufras (2003).

3.3. Gauss Copula

There exists a large number of different copula functions, for instance a set of Archimedean copulas. Let us recall here another rather universal and flexible type, namely the Gauss copula. Let X,Y be standard normal random variables $\sim N(0,1)$ tied with (Pearson) correlation $\rho = \rho(X,Y)$. We denote ϕ,φ univariate standard normal distribution function and density and by $\phi_2(x,y),\varphi_2(x,y)$ corresponding 2-dimensional functions. Then

$$\varphi_2(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2} x' \Sigma^{-1} x\}$$

with $x = (x,y)'$ and the covariance matrix $\Sigma = [1,\rho;\rho,1]$. If we define $U = \phi(X), V = \phi(Y)$, we obtain a 2-dimensional distribution on $(0,1)^2$ with the copula

$$C(u,v) = \phi_2(\phi^{-1}(u),\phi^{-1}(v)). \tag{4}$$

Naturally, $\rho(U,V) \neq \rho(X,Y)$ (though they are rather close, as a rule), while Spearman's correlations coincide, namely $\rho_{\mathrm{SP}}(X,Y) = \rho_{\mathrm{SP}}(U,V) = \rho(U,V)$. As we are primary interested in the model for dependence of competing variables S,T, let us assume that their joint distribution function is given by Gauss copula (4),

$$F_2(s,t) = \phi_2(\phi^{-1}(F_S(s)),\phi^{-1}(F_T(t))), \tag{5}$$

and $S = F_S^{-1}(\phi(X))$, $T = F_T^{-1}(\phi(Y))$. Again $\rho_{\mathrm{SP}}(S,T) = \rho_{\mathrm{SP}}(U,V)$, and "initial" $\rho = \rho(X,Y)$ is the only parameter describing the dependence of S and T. It, naturally, differs from $\rho(S,T)$, however, all values $\rho(S,T)$ can be achieved by convenient choice of $\rho(X,Y)$. Let us remark here that the real dependence among S,T can be much more complicated, nevertheless the use of Gauss copula offers here certain rather simple and sufficiently flexible (as regards the correlation) set of distributions.

The identifiability of the model based on Gauss copula can be proved by the same arguments as used in Escarella and Carriere (2003), Sect. 3, for the case of Frank copula. Namely, as there exists a unique

monotone relationship between $\rho(S,T)$ and ρ of copula, it is not possible to get the same competing risk scheme from two distinct models with marginals of the same parametric type.

3.4. Estimation in Gauss Copula Model

Provided the data are $(Z_i,\delta_i), i=1,\ldots,N$, the likelihood function then has the form

$$L=\prod_{i=1}^{N}\left\{-\frac{\partial}{\partial s}\overline{F}_2(s,t)\right\}^{I[\delta_i=1]}\cdot\left\{-\frac{\partial}{\partial t}\overline{F}_2(s,t)\right\}^{I[\delta_i=2]}\cdot\overline{F}_2(s,t)^{I[\delta_i=0]},$$

evaluated at $s=t=Z_i$, with $\overline{F}_2(s,t)=P(S>s,T>t)=1-F_S(s)-F_T(t)+F_2(s,t)$. From transformation (5) it follows that $F_2(s,t)=\phi_2(x,y)$ with $x=\phi^{-1}(F_S(s))$, $y=\phi^{-1}(F_T(t))$. Hence, when we put $X_i=\phi^{-1}(F_S(Z_i)), Y_i=\phi^{-1}(F_T(Z_i))$, we obtain after some computation – integration of 2-dimensional Gauss density $\varphi_2(x,y)$, that

$$L=\prod_{i=1}^{N}\left\{f_S(Z_i)\left[1-\phi_1(Y_i;\rho X_i,1-\rho^2)\right]\right\}^{I[\delta_i=1]}\cdot$$

$$\cdot\left\{f_T(Z_i)\left[1-\phi_1(X_i;\rho Y_i,1-\rho^2)\right]\right\}^{I[\delta_i=2]}\cdot\left\{1-F_S(Z_i)-F_T(Z_i)+\phi_2(X_i,Y_i)\right\}^{I[\delta_i=0]},$$

where $\phi_1(x;\mu,\sigma^2)$ denotes the distribution function of normal distribution $N(\mu,\sigma^2)$, evaluated at x. Parameter ρ is hidden in ϕ_1 and in ϕ_2. Distributions of S and T are present both explicitly and also implicitly, in transformed X_i,Y_i. It is seen that the problem of maximization is not an easy task and has to be solved by a convenient search procedure. It also means that confidence intervals for estimated parameters cannot be derived directly from the likelihood, we have to search other ways. The bootstrap method offers one of possibilities. Alternatively, Bayes credibility intervals can be obtained in the Bayes approach framework accompanied by the MCMC procedure.

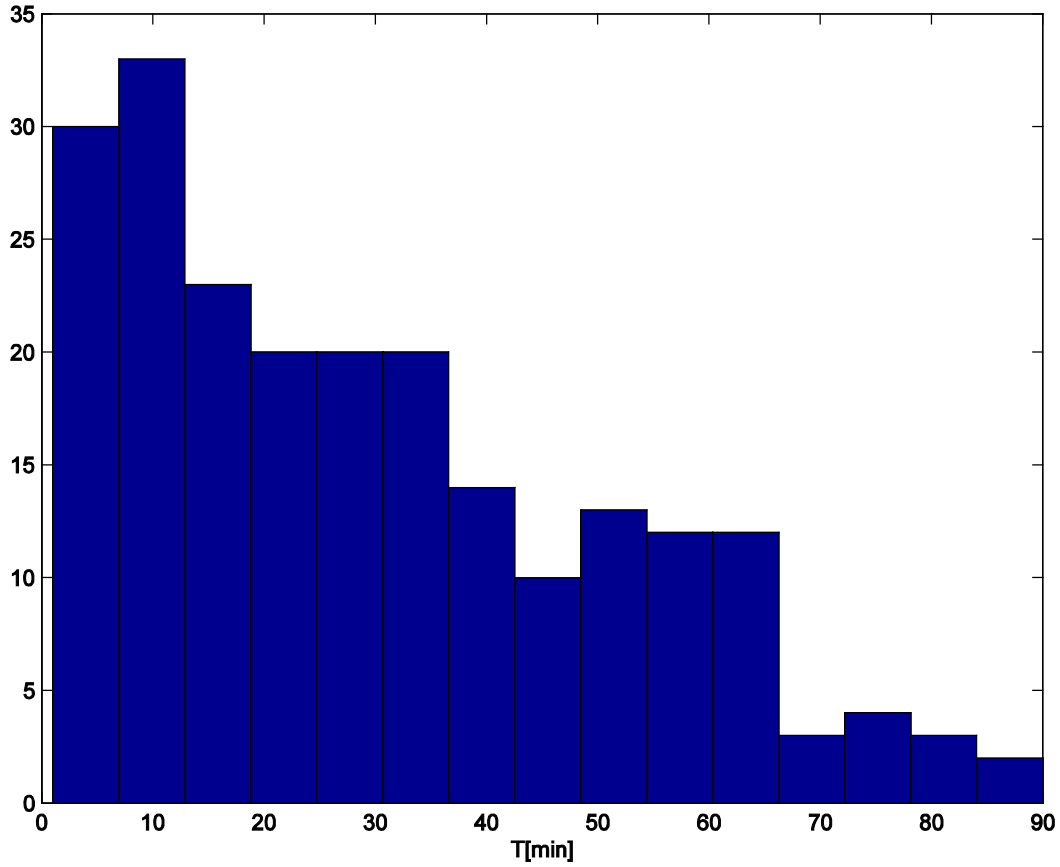

Figure 2. Histogram of times of first goals (without 21 cases censored at 90-th minute).

Final order	Team	Score	Points	1-st goal scored	Home obt.	away scored	obt.
1	Plzen	70:24	72	12	3	11	3
2	Sparta	57:20	67	9	6	9	4
3	Jablonec	58:22	64	12	1	9	6
4	Ml.Boleslav	43:34	46	11	4	5	8
5	Pribram	40:45	43	11	4	5	7
6	Dukla	34:40	41	7	5	3	10
7	Teplice	41:37	38	8	5	7	7
8	Bohemians	35:41	38	6	6	4	11
9	Slovacko	43:46	37	8	7	6	8
10	Jihlava	33:38	36	6	8	7	6

11	Slavia	40:45	34	9	6	7	7
12	Liberec	39:43	33	5	7	6	8
13	Ostrava	23:41	33	7	5	3	10
14	Brno	34:45	33	6	8	2	12
15	Hradec Kr.	26:52	25	6	6	4	11
16	C.Budejovice	29:72	22	6	9	2	11

Table 1. Brief statistics of 2014-15 season of Synot League.

Team	α		β		a	b
Plzen	0.9304	(0.4666)	-1.7413	(1.7616)	2.5354	0.1753
Sparta	0.3217	(0.6060)	-0.8270	(0.8241)	1.3795	0.4374
Jablonec	0.1667	(0.5588)	-1.3309	(1.1298)	1.1814	0.2642
Ml.Boleslav	0.7641	(0.5541)	-0.1269	(0.6481)	2.1470	0.8808
Pribram	-0.0914	(0.6905)	-0.5875	(0.7496)	0.9127	0.5557
Dukla	-0.2924	(0.8051)	0.0885	(0.5798)	0.7465	1.0926
Teplice	-0.0239	(0.6220)	-1.4007	(1.2831)	0.9764	0.2464
Bohemians	-1.4190	(1.4221)	-0.4620	(0.6473)	0.2419	0.6300
Slovacko	0.1708	(0.6473)	-0.1054	(0.6114)	1.1862	0.9000
Jihlava	-0.2503	(0.7302)	-0.6688	(0.7623)	0.7785	0.5123
Slavia	0.2879	(0.5783)	-0.5770	(0.7992)	1.3336	0.5616
Liberec	-0.5494	(0.8513)	-0.1822	(0.6086)	0.5773	0.8335
Ostrava	-0.3784	(0.7783)	-0.2759	(0.6293)	0.6850	0.7589
Brno	-0.6535	(0.9155)	0.2373	(0.5233)	0.5202	1.2678
Hradec Kr.	-0.4139	(0.8232)	-0.0117	(0.5760)	0.6611	0.9884
C.Budejovice	-0.0877	(0.9006)	0.4620	(0.4923)	0.9160	1.5873

Table 2. Results: Estimated parameters $\alpha_i = \ln a_i, \beta_i = \ln b_i$ (with half-widths of approximate 95% conf. intervals in brackets), then a_i, b_i.

4. Application to the Time of First Goal

We shall now use the competing risk model derived in Part 3.1 to modeling the time to first scored goal during a football match. Marginal variables are the latent times of 1-st goal of each team, however only the incidence of one (the first) of them is observed. Or, in the case of draw 0:0, we have censoring by a fixed value 90 minutes. Except statistical estimation of model parameters, we are interested in the following question: How dependent are these latent times to 1-st goal?

In our study we shall use the data from the Czech Synot League, season 2014-15. 16 participating teams played together 240 matches (i.e. twice with each other, home and away). All observed times of first goals (219 cases) are displayed in Figure 2, 21 matches ended without goals. Figure 2 suggests that the times to first goals can be modeled via exponential distribution. The maximum likelihood estimate of its intensity parameter yields $\hat{\lambda} = 0.0261,$ with 95% confidence interval $(0.0228, 0.0297)$. It also means that the mean time to first goal was $1/\hat{\lambda} \sim 38$ minutes. Table 1 shows the final order after the season, with some additional statistics, also concerning first goals. Namely, its right part contains number of matches in which the team scored or obtained the first goal (home or away). More information on the Czech football Synot league can be found on *http://www.sport.cz/fotbal/synot-liga/#vysledky*.

As regards marginal models, the source was the standard model of Maher (1982). More specifically, each team (i) was characterized by its attack parameter a_i and defense parameter b_i. Another parameter, h, denotes the advantage of home field. The sequence of scoring in a match between home team i and away team j is then described by two Poisson processes with intensities $a_i \cdot b_j \cdot h$, $a_j \cdot b_i$, respectively. Consequently, the time to the 1-st goal arises from two competing exponential random variables

$$S_{ij} \sim Exp(a_i \cdot b_j \cdot h),\ T_{ij} \sim Exp(a_j \cdot b_i).$$

Further, it was assumed that their mutual dependence can be expressed via Tsiatis' model described in Part 3.1. Notice also that in the present setting the teams parameters values are not given uniquely, that the parameters $a_i \cdot c,\ b_j / c$ yield the same model, for any $c > 0$. The parameters are related to "time unit" 100 minutes, one reason for it was also a better stability of numerical procedures.

4.1. Numerical Results

We were solving the problem of the maximum likelihood estimation (MLE) of 34 parameters: a_i, b_i of 16 teams, home advantage parameter h, and γ characterizing the dependence. It was assumed that both h and γ were the same for all couples of teams, i.e. in all matches. The results of the MLE of teams parameters are displayed in Table 2. For computational convenience, we estimated $\alpha_i = \ln a_i, \beta_i = \ln b_i$, also $\delta = \ln h$. The ML estimates of two common parameters (with half-widths of 95% confidence intervals) were

$$\hat{\delta} = 0.6423\,(0.2048),\ \hat{h} = \exp(\delta) = 1.9008,\ \ \hat{\gamma} = 1.1700\,(0.1192).$$

The correlation in each particular match depends on teams parameters and on two common parameters h and γ. Its value can be traced roughly from Figure 1, or computed from corresponding two-dimensional exponential model. It is possible to say that the first goal matters. For instance in the match of two leaders, Plzen and Sparta, numerical computation yielded $\rho(S,T) = -0.569$. Then, in a match of teams with rather poor attack and yet fair defence, as for instance Bohemians and Jihlava, obtained $\rho(S,T) = -0.800$ could be interpreted that the first goal was even more important. Further, the value of parameter $h = 1.9$ indicated that the chance of home team to score first was about $1.9 / 2.9 = 0.66$, while in reality from 219 first goals 129 were scored by home teams, 129/219 = 0.59.

4.2. Discussion of Results

Though, theoretically, the identification of underlying distributions, and therefore also the consistency of parameter estimates, is guaranteed (see Part 3.1), simulated experiments show rather slow convergence of estimates to 'true' values. It is also seen that the confidence intervals for parameters (approximate, i.e. based on asymptotic normality of estimates) are rather wide, which is a natural consequence of rather large proportion of the number of parameters to the number of data (matches). Simultaneously, however, this follows from the fact that the log-likelihood function is rather flat.

Though the following problem is not in the center of our attention, we still can provide a brief statistics concerning the impact of the first goal to final result. From 240 matches of the season 21 ended without goals. From remaining 219 matches there were 37 draws (other then 0:0), 116 home wins, 66 away wins. In these 219 matches with goals in 156 cases the team scoring first was also the winner, in 26

cases the opposite had occurred (and 37 ended by draw). The proportion $\hat{p} = 156/219 = 0.7123$ is significantly larger than 0.5, approximate (asymptotic) 95% confidence interval for 'true' proportion equals $(0.6524, 0.7723)$. On the other hand, the chance to turn over the score after obtaining the first goal can be described by estimated proportion $\hat{q} = 26/219 = 0.1187$, which is significantly larger than zero, yielding 95% confidence interval $(0.0759, 0.1616)$.

It is worth to mention here some contributions devoted to the analysis of the first goal impact directly, as for example Nevo and Ritov (2013). Paper of Volf (2009) also contains a small simulation study analyzing probabilities of final results conditionally on the first goal author and time.

4.3. Use of Gauss Copula

Just in order to check and support the results of the preceding part, we repeated the analysis using the Gauss copula, in a way described in Parts 3.3. and 3.4, connecting two exponential distributions characterizing each match. Now, we estimated directly parameters a_i, b_i, h and ρ, the last two were common for all matches. Results, ML estimates of teams parameters, are displayed in Table 3, estimated parameter of home advantage was $h = 1.7229$. It can be said that values are comparable to those of the preceding analysis. Further, optimal correlation estimate was $\rho = -0.520$. As the values maximizing the log-likelihood were found by a random search, we were not able to obtain reliable confidence intervals. While Barnett copula allows only for non-positive correlation, Gauss copula is more universal, it was another reason for using it for supporting the solution.

Team	a	b	Team	a	b
Plzen	2.5572	0.2838	Slovacko	1.1704	0.9205
Sparta	1.5007	0.4857	Jihlava	0.9253	0.6620
Jablonec	1.4321	0.3362	Slavia	1.3612	0.6721
Ml.Boleslav	1.8877	0.7706	Liberec	0.6933	0.8550
Pribram	1.0976	0.6171	Ostrava	0.6679	0.7775
Dukla	0.7127	1.0347	Brno	0.5318	1.2938
Teplice	1.1318	0.4710	Hradec Kr.	0.6817	1.0116
Bohemians	0.5240	0.8343	C.Budejovice	0.8395	1.4094

Table 3. Estimated parameters a_i and b_i in the model using Gauss copula.

In fact, in joint exponential distribution given through a copula, the resulting correlation of involved exponentially distributed variables does not depend on their parameters, it depends just on parameter of the copula. It concerns also the Gauss copula model. Thus, in our case, the value $\rho(X,Y) = -0.520$ leads to (computed numerically) $\rho(U,V) = -0.5027$ and $\rho(S,T) = -0.3775$ in each match. The explanation is seen easily from the formula for correlation. Namely, for two exponentially distributed random variables S,T with parameters μ,λ, respectively, we shall obtain that $\rho(S,T) = E(S\cdot T)\cdot\mu\cdot\lambda - 1$. Further

$$E(S\cdot T) = \int_0^\infty \int_0^\infty s\,t\,f_{ST}\,ds\,dt = \int_0^1 \int_0^1 F_S^{-1}(u)\,F_T^{-1}(v)\,c(u,v)\,du\,dv =$$

$$= \frac{1}{\mu\lambda}\int_0^1 \int_0^1 ln(1-u)\,ln(1-v)\,c(u,v)\,du\,dv,$$

after substitution $u = F_S(s), v = F_T(t)$, taking into account relation (2). It is seen that μ,λ vanish from the expression for $\rho(S,T)$.

This property concerns, naturally, also to Barnett copula, however notice that in our approach using this copula the parameters μ,λ were actually a part of copula parameter θ, because we concentrated to estimation of parameter $\gamma = \theta\cdot\mu\cdot\lambda$. Therefore, $\rho(S,T)$ depended on both. In such a way, we were actually using a set of Barnett copulas and it can be said that such a model is more flexible, than, for instance, the Gauss copula model presented above.

5. Concluding Remarks

We have studied the dependence of random variables – latent times of scoring the first goal in a football match, with the aid of the competing risks model. Achieved results lead to conclusion that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches, i.e. the matches of teams with good defence and comparable attack abilities. This conclusion is thus in accord with results of McHale and Scarf (2011). It has to be pointed out that the teams parameters evaluated above concern just to stage of match up to the first goal. It can be expected that the team performance changes during the match and is related to actual score, elapsing time, and to other factors characterizing the match state. This

aspect is also reflected by more advanced models of score development, see again for instance Dixon and Robinson (1998), Volf (2009) and an overview of models provided there. Hence, the approach proposed in the present study can be extended to the analysis of times to next goals. Another generalization can consider different copula parameters for certain groups of matches or teams.

Acknowledgement

The research has been supported by the project No 13-14445S of the Czech Scientific Foundation.

References

[1]. Basu, A.P., Ghosh, J.K. (1978). Identifiability of the Multinormal and Other Distributions under Competing Risks Model. Journal of Multivariate Analysis, 8, 413-429.

[2]. Dixon, M.J., Robinson, M.E. (1998). A birth process model for association football matches. The Statistician, 47, 523-538.

[3]. Escarela, G., Carriere, J.F. (2003). Fitting competing risks with an assumed copula. Statistical Methods in Medical Research, 12(4), 333-349.

[4]. Heckman, J.J., Honoré, B.E. (1989). The identifiability of the competing risks model. Biometrika, 76, 325-330.

[5]. 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

[6]. Karlis, D., Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. J. R. Stat. Soc. Ser. D, 52, 381-394.

[7]. Lee, S. (2006). Identifcation of a competing risks model with unknown transformations of latent failure times. Biometrika, 93, 996-1002.

[8]. Lin, D.Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. Statistics in Medicine, 16, 901-910.

[9]. Maher, M.J. (1982). Modelling association football scores. Stat. Neerl., 36, 109-118.

[10]. McHale I., Scarf P.A. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. Statistical Modelling, 11, 219-236.

[11]. Nevo, D., Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. Journal of Quantitative Analysis in Sports, 9, 165-177.

[12]. Schwarz, M., Jongbloed, G., and Van Keilegom, I. (2013). On the identifiability of copulas in bivariate competing risks models. Canadian Journal of Statistics, 41, 291-303

[13]. Tsiatis, A. (1975). A nonidentifiability aspects of the problem of competing risks. Proc. Nat. Acad. Sci. USA, 72, 20-22.

[14]. Volf, P. (2009). A random point process model for the score in sport matches. IMA Journal of Management Mathematics, 20, 121-131.

[15]. Zheng, M., Klein, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika, 82, 127-138.

Appendix

We shall now prove that the case of competing risks with two exponential marginal distributions tied together by Barnett copula (2) is identifiable. Let us again consider random variables $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$, and their joint survival function $\bar{F}_2(s,t) = e^{-\lambda s - \mu t - \gamma st}$, $\lambda, \mu > 0, \gamma \geq 0$. Hence, in the competing risks setting, $S(z) = \bar{F}_2(z,z) = exp(-\lambda z - \mu z - \gamma z^2)$, incidence densities are $f_S^*(z) = (\lambda + \gamma z) \cdot S(z)$, $f_T^*(z) = (\mu + \gamma z) \cdot S(z)$, and the likelihood, given data $\{z_i, \delta_i, i = 1,...,N\}$ is

$$L = \prod_{i=1}^{N} (\lambda + \gamma z_i)^{[\delta_i = 1]} \cdot (\mu + \gamma z_i)^{[\delta_i = 2]} \cdot S(z_i).$$

It will be shown that the model fulfils the regularity conditions. In particular, the 2-nd derivative of the log-likelihood is a negative definite matrix, therefore there exists unique maximum likelihood estimate of parameters λ, μ, γ. The log-likelihood and its first derivatives are:

$$l = \ln L = \ln(\lambda + \gamma z) \cdot [\delta = 1] + \ln(\mu + \gamma z) \cdot [\delta = 2] - (\lambda z + \mu z + \gamma z^2).$$

$$\frac{\partial l}{\partial \lambda} = \frac{[\delta = 1]}{(\lambda + \gamma z)} - z, \quad \frac{\partial l}{\partial \mu} = \frac{[\delta = 2]}{(\mu + \gamma z)} - z, \quad \frac{\partial l}{\partial \gamma} = \frac{[\delta = 1]}{(\lambda + \gamma z)} \cdot z + \frac{[\delta = 2]}{(\mu + \gamma z)} \cdot z - z^2.$$

It is easy to show that expectations of the first derivatives equal zero, provided we take Z as random variable having survival function $S(z)$ with 'true' λ, μ, γ:

$$E\frac{\partial l}{\partial \lambda} = \int_0^\infty \frac{f_S^*(z)dz}{(\lambda + \gamma z)} - \int_0^\infty z(-dS(z)) = \int_0^\infty S(z)dz - \int_0^\infty S(z)dz = 0,$$

the second term was obtained by the *per-partes* integration. Similarly, $E\frac{\partial l}{\partial \mu}=0$, too, and

$$E\frac{\partial l}{\partial \gamma}=\int_0^\infty z\cdot S(z)dz+\int_0^\infty z\cdot S(z)dz-\int_0^\infty z^2\cdot(-dS(z))=0,$$

again after *per-partes* integration of the last term. As regards the second derivatives,

$$\frac{\partial^2 l}{\partial \lambda^2}=-\frac{[\delta=1]}{(\lambda+\gamma z)^2},\quad \frac{\partial^2 l}{\partial \mu^2}=-\frac{[\delta=2]}{(\mu+\gamma z)^2},\quad \frac{\partial^2 l}{\partial \mu\partial \lambda}=0,$$

$$\frac{\partial^2 l}{\partial \lambda\partial \gamma}=-\frac{[\delta=1]}{(\lambda+\gamma z)^2}\cdot z,\quad \frac{\partial^2 l}{\partial \mu\partial \gamma}=-\frac{[\delta=2]}{(\mu+\gamma z)^2}\cdot z,$$

$$\frac{\partial^2 l}{\partial \gamma^2}=-\frac{[\delta=1]}{(\lambda+\gamma z)^2}\cdot z^2-\frac{[\delta=2]}{(\mu+\gamma z)^2}\cdot z^2.$$

Again, it is easy to show that the matrix of 2-nd derivatives (let us denote it D) is negative definite for all finite z,λ,μ,γ: Let $c=(c_1,c_2,c_3)'$ be arbitrary non-zero vector, then

$$c'Dc=-c_1^2\frac{[\delta=1]}{(\lambda+\gamma z)^2}-c_2^2\frac{[\delta=2]}{(\mu+\gamma z)^2}-2c_1c_3\frac{[\delta=1]}{(\lambda+\gamma z)^2}z-$$

$$-2c_2c_3\frac{[\delta=2]}{(\mu+\gamma z)^2}z-c_3^2\{\frac{[\delta=1]}{(\lambda+\gamma z)^2}z^2+\frac{[\delta=2]}{(\mu+\gamma z)^2}z^2\}=$$

$$-[\delta=1]\{\frac{c_1}{(\lambda+\gamma z)}+\frac{c_3 z}{(\lambda+\gamma z)}\}^2-[\delta=2]\{\frac{c_2}{(\mu+\gamma z)}+\frac{c_3 z}{(\mu+\gamma z)}\}^2<0.$$

Published: Volume 2016, Issue 10 / October 25, 2016