

# Application of competing risks model to the analysis of the first goal time

Petr Volf

UTIA AV CR, *volf@utia.cas.cz*

## **O U T L I N E:**

1. Motivation: football match score models with dependent components,
2. Competing risks scheme, observed data, problem of model identifiability,
3. Use of copula to express multivariate distribution,
4. Model of exponential distributions with Barnett copula,
5. Application: Time to 1-st goal in a football match, discussion (data from 2014-15 season of the Synot League).

# 1 Motivation

A basic probability model for final score of a football (soccer) match

(Maher, 1982): Two independent Poisson random variables.

More flexible models – generalizations (inflated models),

time development of model parameters, use of covariates

=> conditional independence, counting process model (Volf, 2009), ..

## Another direction of model improvement:

– an explicit form of dependence of both teams scoring distributions.

Karlis and Ntzoufras (2003) – a special case of bivariate Poisson distribution.

McHale and Scarf (2011) – dependence via a copula model,

the copula is used for joint distribution of 2 discrete random variables.

## **Results of these models:**

In 'bivariate Poisson' the correlation  $> 0$  (by definition),

McHale and Scarf conclude that *'the correlation is negative and is absolutely larger in more competitive matches'*.

Remark: The use of copula in discrete distribution models is not easy technically (and then computationally), marginal distribution functions are as a rule expressed by sums of point probabilities, not having a reasonably closed form.

## **In the present contribution:**

the copula model is used, too, but for continuous distribution of the time to the first scored goal in a match,

=> we deal with the scheme of competing risks given by two dependent continuous (exponential) distributions.

## 2 Competing risks scheme in survival analysis

Situation: End of observation can be caused by one (1-st) of  $K$  causes (events),  
– there is  $K$  (possibly dependent) random variables (times)  $T_j, j = 1, \dots, K$ ,  
(plus variable  $C$  of random right censoring)

Denote  $\bar{F}_K(t_1, \dots, t_K) = P(T_1 > t_1, \dots, T_K > t_K)$  joint survival function of  $\{T_j\}$ .

**We observe**  $Z = \min(T_1, \dots, T_K, C)$

and indicator  $\delta = j$  if  $Z = T_j$ ,  $\delta = 0$  if  $Z = C$ .

**Estimable:** distribution of  $Z = \min(T_1, \dots, T_K)$

e.g.  $S(t) = P(Z > t) = \bar{F}_K(t, \dots, t)$  – its survival function (by KM PLE)

and **Cause-specific (“crude”) hazard functions** for events  $j = 1, 2, \dots, K$ :

$$h_j^*(t) = \lim_{d \rightarrow 0} \frac{P(t \leq Z < t + d, \delta = j \mid Z \geq t)}{d},$$

overall hazard rate for  $Z = \min(T_1, \dots, T_K)$ :

$$h^*(t) = \lim_{d \rightarrow 0} \frac{P(t \leq Z < t + d \mid Z \geq t)}{d} = \sum_{j=1}^K h_j^*(t),$$

integrals = cumulated hazard rates  $H_j^*(t)$ ,  $H^*(t)$ , by Nelson-Aalen est.,

and **Cumulated incidence functions** (consistently - e.g. Lin, 1997):

$$F_j^*(t) = P(Z \leq t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s) ds.$$

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

## **Non-identifiability problem:**

In general, from data  $(Z_i, \delta_i)$ ,  $i = 1, \dots, N$  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  $h_j^*(t)$ .

On the other hand, if the form of marginal and joint distributions is assumed, then in many cases the identification of right parameters is possible (e.g. Basu and Ghosh, 1978, and numerous more recent papers).

### 3 Competing risk and copula

In the sequel we shall consider just 2 competing events, i.e. random variables  $S, T$  and data  $Z_i = \min(S_i, T_i, C_i), \delta_i = 1, 2, 0$ .

Copula offers a way how to model joint distribution or survival function:

$$\overline{F}_2(s, t) = C(\overline{F}_S(s), \overline{F}_T(t), \theta), \quad (1)$$

$\overline{F}_S, \overline{F}_T$  are marginal survival functions of  $S, T$ ,  $\theta$  is a copula parameter.

“Knowledge” of copula is still an unrealistic supposition. Nevertheless, we can try to use certain sufficiently flexible class of copulas, for approximation.

It “remains” to estimate its parameter  $\theta$  (the same non-identifiability problem).

## Non-identifiability example of Tsiatis (1975):

Consider random variables  $S, T$  following the distribution

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

Hence,  $S(t) = \bar{F}_2(t, t) = \exp(-\lambda t - \mu t - \gamma t^2)$ .

Corresponding cause-specific hazard rates and their integrals are

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

and  $S^*(t) = \exp(-H_S^*(t) - H_T^*(t))$  is the same as  $S(t)$  above.

It means that independent random variables with marginal survival functions

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

yield the same competing risk scheme.



Notice that 'true' marginals are exponential while 'independent' are not  
=> a chance that, when type of marginals is assumed,  
they can be estimated, and  $\gamma$ , too.

Again, several such cases are analyzed already by Basu and Ghosh (1978),  
and then by others.

Tsiatis' example actually uses the Barnett copula

$$C(u, v) = u \cdot v \cdot \exp(-\theta \cdot \ln u \cdot \ln v)$$

with  $\theta \geq 0$ , then  $\rho(U, V) \leq 0$ ,  $\theta = 0 \Leftrightarrow$  independence.

Tsiatis' parameter  $\gamma = \theta \cdot \lambda \cdot \mu$ .

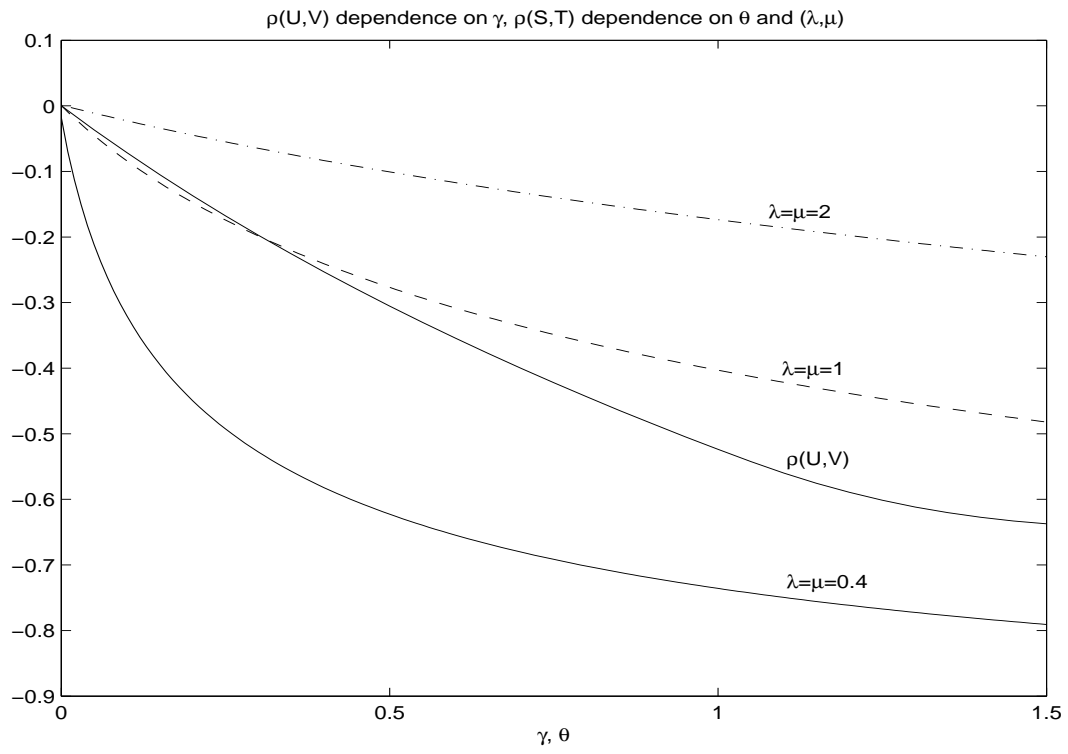


Figure 1: Dependence of  $\rho(U,V)$  on parameter  $\theta$  and  $\rho(S,T)$  or  $\gamma$ , when  $S \sim \text{Exp}(\lambda)$ ,  $T \sim \text{Exp}(\mu)$ .

## 4 APPLICATION

to distribution of times to 1-st goal in football (soccer) match.

**DATA:** Synot Liga 2014-15, 16 teams, 240 matches,

*<http://www.sport.cz/fotbal/synot-liga/#vysledky>*

**Question:** How dependent are the 'latent' times to 1-st goal of two teams?

**MODEL** - based on the standard model used for modeling score of (football) matches (e.g. Maher, 1982):

Each team ( $i$ ) has an attack parameter  $a_i$  and defense parameter  $b_i$ ,  
additional parameter  $h$  of home team advantage.

Scoring in a match between teams  $i$  (=home) and  $j$  (away) - two Poisson processes with intensities  $a_i \cdot b_j \cdot h$ ,  $a_j \cdot b_i$ , resp.

**The time to 1-st goal** - two competing exponential random variables

$$S_{ij} \sim \text{Exp}(\lambda_{ij} = a_i \cdot b_j \cdot h), \quad T_{ij} \sim \text{Exp}(\mu_{ij} = a_j \cdot b_i).$$

Dependence expressed via 'Tsiatis' model with parameter  $\gamma$ .

The model has 34 parameters:  $a_i, b_i$  of 16 teams, and  $h, \gamma$ .

Barnett copula parameter  $\theta = \gamma / (\lambda_{ij} \mu_{ij})$  is different for each match, hence also the correlation differs, depends on the match.

Final order	Team	Score	Points	1-st goal scored	home obt.	away scored	away 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.

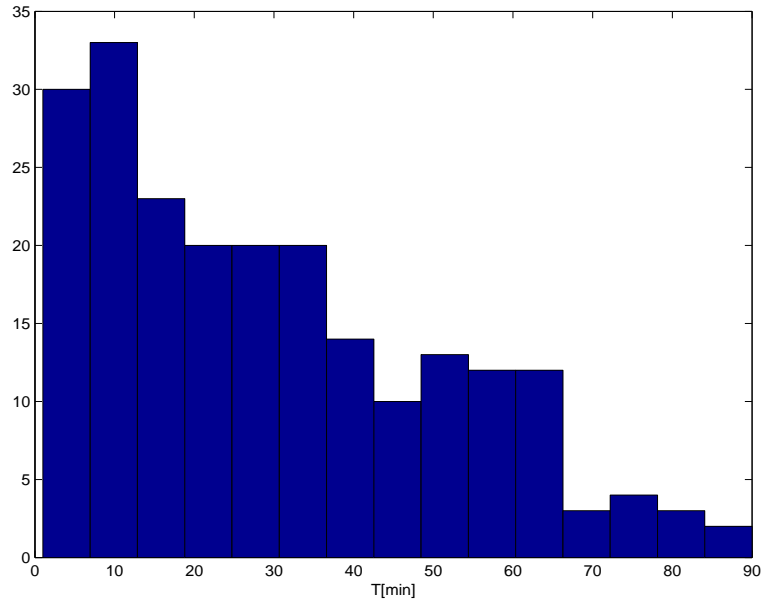


Figure 2: Histogram of times of 219 first goals. 21 matches were without goals, i.e. observations were censored at 90-th minute.

Distribution of times to first goals is not far from exponential distribution (?).

ML estimate of the intensity yielded  $\hat{\lambda} = 0.0261$ ,  $(0.0228, 0.0297)$ ,  
the mean time to first goal  $1/\hat{\lambda} \sim 38$  minutes.

Team	$\alpha$		$\beta$		$a$	$b$
Plzen	0.9742	(0.4664)	-1.8874	(1.7569)	2.6490	0.1515
Sparta	0.3662	(0.6055)	-0.9755	(0.8235)	1.4422	0.3770
Jablonec	0.2115	(0.5584)	-1.4791	(1.1286)	1.2356	0.2278
Ml.Boleslav	0.8080	(0.5539)	-0.2759	(0.6479)	2.2433	0.7589
Pribram	-0.0464	(0.6898)	-0.7362	(0.7491)	0.9547	0.4789
Dukla	-0.2479	(0.8046)	-0.0606	(0.5797)	0.7804	0.9412
Teplice	0.0205	(0.6216)	-1.5465	(1.2794)	1.0207	0.2130
Bohemians	-1.3719	(1.4189)	-0.6103	(0.6467)	0.2536	0.5432
Slovacko	0.2151	(0.6469)	-0.2541	(0.6111)	1.2400	0.7756
Jihlava	-0.2056	(0.7296)	-0.8168	(0.7615)	0.8141	0.4419
Slavia	0.3320	(0.5780)	-0.7249	(0.7983)	1.3938	0.4843
Liberec	-0.5043	(0.8504)	-0.3311	(0.6083)	0.6039	0.7181
Ostrava	-0.3343	(0.7779)	-0.4247	(0.6289)	0.7159	0.6540
Brno	-0.6091	(0.9150)	0.0883	(0.5231)	0.5438	1.0923
Hradec Kr.	-0.3694	(0.8226)	-0.1606	(0.5757)	0.6912	0.8517
C.Budejovice	-0.0435	(0.9000)	0.3128	(0.4921)	0.9574	1.3672

Table 2: **Results:** MLE of parameters  $\alpha_i = \ln a_i$ ,  $\beta_i = \ln b_i$  (with half-widths of approximate 95% conf. intervals in brackets), then  $a_i$ ,  $b_i$ . Parameters are related to 90 minutes, in order to have reasonable scales (and avoid numerical problems, too).

Further,  $\hat{\delta} = 0.6417$  (0.2046),  $\hat{h} = \exp(\hat{\delta}) = 1.8997$ ,  $\hat{\gamma} = 0.945$  (0.078).

Parameter values are "relative",  $a_i \cdot c, b_i/c$  yield the same, for any  $c > 0$ .

## 5 Discussion of results

Correlation in each match depends on teams parameters (and on  $h$  and  $\gamma$ ).

Can be computed numerically from corresponding two-dimensional exponential model.

For instance, in the match Plzen and Sparta

$$\rho(S, T) = -0.563,$$

Bohemians vers. Jihlava (teams with rather poor attack and yet fair defence)

$$\rho(S, T) = -0.799.$$

**Interpretation(?)**:

The smaller correlation, the more is the first goal important.

$h = 1.9$  indicated that the chance of home team to score first was about  $1.9/2.9 = 0.66$ ,

– in reality from 219 first goals 129 were scored by home teams,  
 $129/219 = 0.59$ .



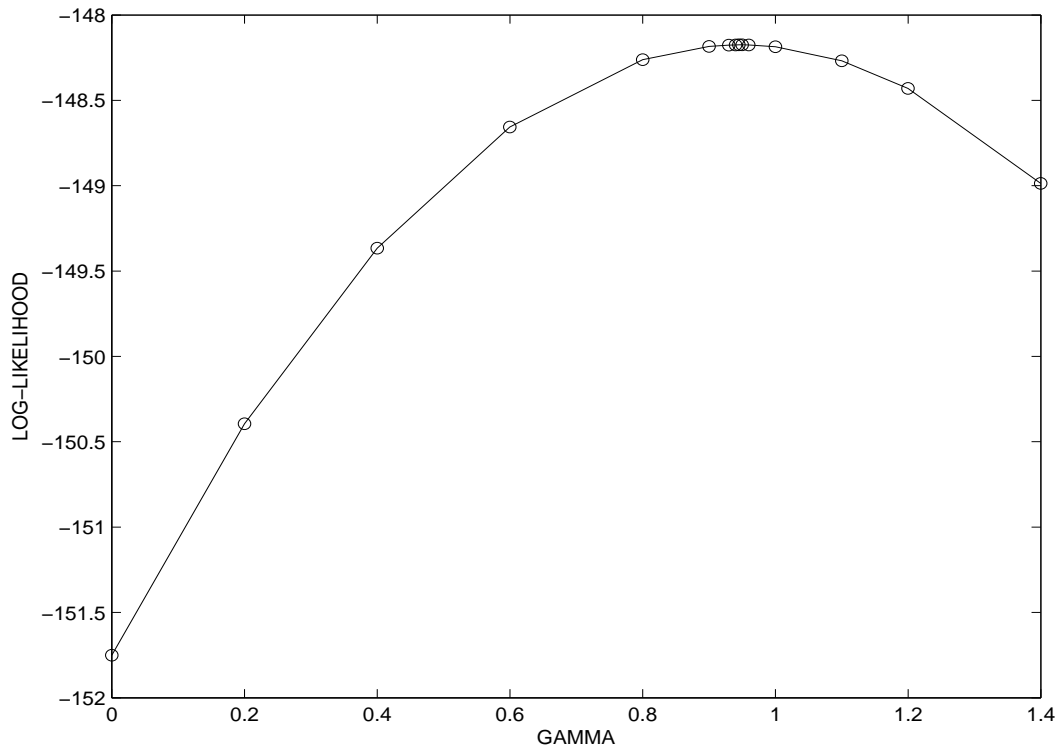


Figure 3: Dependence of max log-likelihood on  $\gamma$ , i.e. maximized over all other parameters when  $\gamma$  is fixed. It shows that the log-likelihood is "flat".

## Simulation

of times to first goal for each match, from it we can estimate:

- probabilities of first scoring team
- the first goal time (rather roughly, the model using exponential distribution has rather large variance).

**Illustration:** 1000x generated matches Plzen–Sparta, Sparta–Plzen, Jihlava–Bohemians and Bohemians–Jihlava,

$p_1$ ,  $p_2$ ,  $p_0$  are relative frequencies of scoring the first goal by the home, away team, or of match without goal.

Match	$\lambda$	$\mu$	$p_1$	$p_2$	$p_0$	Mean(T)	Std(T)
P–S	1.8975	0.2185	0.787	0.162	0.051	32.2196	27.9710
S–P	0.4151	0.9987	0.337	0.561	0.102	43.5117	33.7654
J–B	0.8402	0.1122	0.632	0.231	0.137	51.1139	36.0780
B–J	0.2129	0.4422	0.327	0.483	0.190	59.3336	41.5438

Table 3: Characteristics of randomly generated results of selected matches. Again, intensities  $\lambda$  and  $\mu$  are related to 90 minutes.

In real matches, the first goals were scored by Plzen at min. 12, by Plzen at min. 52, by Bohemians at min. 32, the last match ended 0:0.

It is seen that the prediction of the 1-st goal is rather unreliable.

Some global statistics:

1-st goal	home	away	0:0
reality:	129	90	21
$\sum p$	136	87	17
predicted	201	39	0
agreement	104	15	0

Table 4: Comparison of real and predicted results.

The second row contains sum of probabilities  $p_1, p_2, p_0$ , resp., over all matches.

Prediction in row 3 is based on  $\max(p_1, p_2, p_0)$ . The last row contains the number of cases where the prediction agreed with the reality.

The low prediction reliability is, unfortunately, evident.

## Use of Gauss copula

– in order to check and support the results, we repeated the analysis using the Gauss copula.

Barnett copula yields only non-positive correlation, Gauss copula is universal.

Team	$a$	$b$	Team	$a$	$b$
Plzen	2.4235	0.2695	Slovacko	1.1111	0.8742
Sparta	1.4244	0.4612	Jihlava	0.8771	0.6280
Jablonec	1.3582	0.3193	Slavia	1.2899	0.6387
Ml.Boleslav	1.7917	0.7328	Liberec	0.6587	0.8117
Pribram	1.0402	0.5860	Ostrava	0.6330	0.7384
Dukla	0.6755	0.9844	Brno	0.5048	1.2286
Teplice	1.0727	0.4473	Hradec Kr.	0.6461	0.9620
Bohemians	0.4984	0.7923	C.Budejovice	0.7980	1.3384

Table 5: Estimated parameters  $a_i$  and  $b_i$  in model using Gauss copula

Further, estimated  $h = 1.7229$ ,  $\rho = -0.520 = \rho(X, Y)$   
of involved standard normal  $X, Y$ .

It leads to  $\rho(U, V) = -0.5027$  and  $\rho(S, T) = -0.3775$ ,  
the same for each match.

All estimates were obtained by numerical procedures, hence I do not give CI-s

Quite generally, for two exponentially distributed random variables  $S, T$  with parameters  $\mu, \lambda$ , connected by a copula, their  $\rho(S, T)$  does not depend on  $\mu, \lambda$ :

$\rho(S, T) = E(S \cdot T) \cdot \mu \cdot \lambda - 1$ . Further

$$\begin{aligned} E(S \cdot T) &= \int_0^\infty \int_0^\infty s t f_{ST} ds dt = \int_0^1 \int_0^1 F_S^{-1}(u) F_T^{-1}(v) c(u, v) du dv = \\ &= \frac{1}{\mu \lambda} \int_0^1 \int_0^1 \ln(1 - u) \ln(1 - v) c(u, v) du dv, \end{aligned}$$

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

It is seen that  $\mu, \lambda$  vanish from the expression for  $\rho(S, T)$ .

This concerns also to Barnett copula, however notice that in our approach  $\mu, \lambda$  were a part of copula parameter  $\theta$ , because we estimated parameter  $\gamma = \theta \cdot \mu \cdot \lambda$ . Therefore,  $\rho(S, T)$  depended on both, we were actually using a set of Barnett copulas.

	Order 2014-15	a(14)	b(14)	a(15)	b(15)	order 2015-16
1	Plzen	2.6490	0.1515	1.8218	0.2592	1
2	Sparta	1.4422	0.3770	1.0447	0.6478	2
3	Jablonec	1.2356	0.2278	0.7470	0.5481	7
4	Ml.Boleslav	2.2433	0.7589	1.5209	0.6408	4
5	Pribram	0.9547	0.4789	0.6197	0.9066	14
6	Dukla	0.7804	0.9412	1.2129	0.5980	10
7	Teplice	1.0207	0.2130	0.8845	0.8982	12
8	Bohemians	0.2536	0.5432	0.8931	0.5389	9
9	Slovacko	1.2400	0.7756	0.3463	1.0957	8
10	Jihlava	0.8141	0.4419	0.7611	0.5699	11
11	Slavia	1.3938	0.4843	1.4646	0.3974	5
12	Liberec	0.6039	0.7181	1.0913	0.5251	3
13	Ostrava	0.7159	0.6540	1.3502	1.1322	16
14	Brno	0.5438	1.0923	0.6325	0.5334	6
15	H.Kr./Olomouc	0.6912	0.8517	0.4542	0.4836	15
16	C.Budej./Zlin	0.9574	1.3672	1.1406	0.9276	13

Table 6: Comparison of results, i.e. estimated parameters and final order of teams in seasons 2014-15 and 2015-16.

**COMPARISON** – continued:

2014-15:  $\hat{\delta} = 0.6417$  (0.2046),  $\hat{h} = \exp(\hat{\delta}) = 1.8997$ ,  $\hat{\gamma} = 0.945$  (0.078).

2015-16:  $\hat{\delta} = 0.4837$  (0.1966),  $\hat{h} = \exp(\hat{\delta}) = 1.6221$ ,  $\hat{\gamma} = 1.450$  (0.117).

Our main interest is the correlation,

– influenced by teams parameters and  $h$ ,  $\gamma$ .

For instance, now

for Sparta vers. Plzen  $\rho = -0.602$  (-0.569 in 2014/14),

for Bohemians vers. Jihlava  $\rho = -0.676$  (-0.800 in 2014/15),

– the first-goal intensity of the Bohemians has increased considerably.



## 6 Conclusion

- We have studied the dependence of random variables – latent times of scoring the first goal in football matches, with the aid of the competing risk model.
- Achieved results lead to conclusion that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches (compare with McHale and Scarf, 2011).
- The approach can be extended to the analysis of times to next goals, further generalization can consider different copula parameters for certain groups of matches and/or teams.
- Further, in a more general models the intensities can also depend on other factors and on match development (see also Volf, 2009 for an overview of models).

## 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] Chen, Y.H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J.R.Statist. Soc. B*, 72, Part 2, 235-251.
- [3] Dixon, M.J., Robinson, M.E. (1998). A birth process model for association football matches. *The Statistician*, 47, 523-538.
- [4] Escarela, G., Carriere, J.F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4), 333-349.
- [5] Heckman, J.J., Honoré, B.E. (1989). The identifiability of the competing risks model. *Biometrika*, 76, 325-330.
- [6] 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
- [7] Karlis, D., Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. Ser. D*, 52, 381-394.
- [8] Lee, S. (2006). Identification of a competing risks model with unknown transformations of latent failure times. *Biometrika*, 93, 996-1002.
- [9] Lin, D.Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16, 901-910.
- [10] Maher, M.J. (1982). Modelling association football scores. *Stat. Neerl.*, 36, 109-118.
- [11] McHale I., Scarf P.A. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 11, 219-236.
- [12] 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.
- [13] 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
- [14] Tsiatis, A. (1975). A nonidentifiability aspects of the problem of competing risks. *Proc. Nat. Acad. Sci. USA*, 72, 20-22.
- [15] Volf, P. (2009). A random point process model for the score in sport matches. *IMA Journal of Management Mathematics*, 20, 121-131.
- [16] Zheng, M., Klein, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82, 127-138.